FEUP EDIÇÕES

DSIE'11

Doctoral Symposium in Informatics Engineering

Proceedings
of the
6th Doctoral Symposium in Informatics Engineering

27-28 JAN 2011
Porto - Portugal

Editors:
A. Augusto Sousa
Eugénio Oliveira

www.fe.up.pt/dsie11

Sponsored by:

DEI Departamento de Engenharia Informática
FEUP
U.PORTO
Tintas 2000
claranet
SISMA
unicer
aefeup
montaco
labs.sapo.pt/up
Delta
REBAU

FEUP EDIÇÕES

DSIE'11

Doctoral Symposium in Informatics Engineering

Proceedings
of the
6th Doctoral Symposium in Informatics Engineering

27-28 JAN 2011
Porto - Portugal

Editors:
A. Augusto Sousa
Eugénio Oliveira

www.fe.up.pt/dsie11

COPYRIGHT

1ª Edição/ 1st Edition 2011

ISBN: 978-972-752-129-6

Editors: A. Augusto Sousa and Eugénio Oliveira

Faculdade de Engenharia da Universidade do Porto

Rua Dr. Roberto Frias, 4200-465 Porto

DSIE'11 SECRETARIAT:

Faculdade de Engenharia da Universidade do Porto

Rua Dr. Roberto Frias, s/n

4200-465 Porto, Portugal

Telephone: +351 22 508 21 34

Fax: +351 22 508 14 43

E-mail: dsie11@fe.up.pt

Symposium Website: http://www.fe.up.pt/dsie11

FOREWORD

2011 Doctoral Symposium in Informatics Engineering - DSIE'11 is an event representing the 6th edition of a scientific meeting organized by PhD students of the FEUP Doctoral Program in Informatics Engineering (ProDEI). These kind of meetings have been held since the school year 2005/06 and their main objective has always been to provide a forum for discussion on, and demonstration of, the practical application of scientific research issues, particularly in the context of information technology, computer engineering and computer science. DSIE symposium comes out as a natural conclusion of a mandatory ProDEI course called Methodologies for Scientific Research (MSR) leading to a formal evaluation of the student acquired competencies.

The aim of that specific course (MSR) is to teach students the processes, methodologies and best practices related to scientific research, particularly in the mentioned areas, as well as to improve their capability to produce adequate scientific texts. With a mixed format based on multidisciplinary seminars and tutorials, the course culminates with the realization of DSIE meeting, seen as a kind of laboratory test of the concepts learned by students. In the scope of DSIE, students are expected to play various roles, such as paper authors, both scientific and organization committee member as well as reviewers, duly accompanied by more senior lectures and professors.

DSIE is then seen as a “leitmotif” for the students to write scientific correct and adequate papers following the methods and good practices currently associated to outstanding research activities in the area. Last two DSIE editions only admitted submissions from ProDEI students and, in many cases, papers already point to interesting research themes PhD students are willing to pursue. Although, still at an embryonic stage, and despite some of the papers are not yet enough mature or only report a state of the art, we already can find some interesting research work or an interesting perspective about future work. At this time, it was not essential, nor even possible, for the students in their PhD first year, to produce strong and deep research results. However, we hope that the basic requirements for presenting an acceptable scientific paper have been fulfilled.

DSIE'11 Proceedings include 26 articles accepted in the context previously defined. They have been put together according to the seven technical symposium sessions. These sessions group some different, although adjacent topics, once, as it was expected, the paper themes are very much heterogeneous. The sessions go from more theoretical to more applied topics, from more focused to broader areas, but always including the algorithmic and scientific research methods flavor. The sessions are "on Information Systems/Data Processing (4 papers), Artificial Intelligence (7 papers), Computer Graphics (5 papers), Mobile Computing and Networks (6 papers) and Other Topics (3 papers)

The complete DSIE'11 meeting has a two days program that includes also an invited talk by an outstanding researcher in Computer Animation.

MSR course and ProDEI responsible professors, were proud to participate in DSIE'11 event and would like to acknowledge all the students who were deeply involved in the success of this event that, we hope, has contributed for a better understanding of the themes that have been addressed during the course, the best scientific research methods and the good practices for writing scientific papers.

Eugénio Oliveira and A. Augusto Sousa
(In charge of the MSR course – Methodologies for Scientific Research) January 2011

PREFÁCIO

O Simpósio Doutoral em Engenharia Informática "2011 *Doctoral Symposium in Informatics Engineering*- DSIE'11" consubstancia a 6ª edição dos Encontros Científicos promovidos pelos Estudantes de Doutoramento do Programa de Doutoramento em Engenharia Informática (ProDEI) da FEUP que se realiza sem interrupção desde o ano lectivo 2005/06. O principal objectivo do DSIE é o de promover um fórum de discussão e de aplicação de práticas de investigação científica, nomeadamente no âmbito da informática, da engenharia informática e da Ciência da Computação. Organiza-se este simpósio como conclusão natural de uma unidade curricular, obrigatória no curso, designada por Metodologias de Investigação Científica (MIC) permitindo uma avaliação completa das várias competências adquiridas pelos estudantes.

Esta disciplina, pertencente ao primeiro semestre da componente curricular do ProDEI, pretende transmitir aos estudantes os processos, metodologias e boas práticas associados à investigação científica, sobretudo no âmbito das áreas referidas, assim como melhorar a sua capacidade de produção adequada de textos científicos. Com um formato misto baseado em tutoriais e seminários multidisciplinares, a unidade curricular culmina com a realização deste encontro que se destina a funcionar, figurativamente, como um laboratório de prova dos conceitos apreendidos pelos estudantes. Estes desempenham vários papéis, como sejam o de autores dos artigos e membros das comissões de organização e científica, devidamente acompanhados pelos docentes da unidade curricular e de outros docentes que aceitam o desafio de colaborar na revisão assim como, em alguns casos, na produção dos artigos submetidos.

O DSIE'11 surge assim como o mote para a produção de artigos com formato científico correcto, onde os autores colocam em prática os conhecimentos adquiridos ao longo da unidade curricular, limitando- se, nas últimas edições, a participação aos estudantes da edição corrente do curso. Em muitos casos, as contribuições apresentadas apontam já para os temas que os estudantes pretendem seguir na componente de investigação do programa doutoral. Numa fase embrionária da investigação, é natural que alguns dos trabalhos se apresentem ainda algo incipientes, ficando-se por uma pesquisa de estado-da-

arte numa área, ou por uma descrição pouco profunda de um trabalho apenas perspectivado e ainda a realizar no futuro. Não é fundamental, nem tal seria possível, na totalidade dos casos, que os trabalhos sejam muito aprofundados, apresentando-se, no entanto, com o mínimo de requisitos que são os normalmente exigidos a um artigo científico.

O presente volume inclui os 26 artigos publicados nesse contexto, reunidos, de acordo com os assuntos versados, em 7 sessões técnicas do simpósio. Estas sessões agrupam e classificam os temas, expectavelmente heterogéneos, face à diversidade das áreas cobertas pelo ProDEI, dos artigos em publicação. Estas sessões incluem tópicos que vão da teoria às aplicações, de tópicos focados a áreas mais abrangentes, mas sempre atendendo à importância devida aos algoritmos e métodos de investigação em estudo:

Sistemas de Informação/Processamento de Dados (4 artigos), Inteligência Artificial (7 artigos), Computação Gráfica (5 artigos), Computação Móvel e Redes (6 artigos), Outros temas (3 artigos).

O programa completo que se estendeu por dois dias, o DSIE'11 incluiu ainda uma sessão com um orador convidado que transmitiu a sua experiência, simultaneamente, científica e aplicacional na área da Animação por Computador.

Os docentes de MIC do ProDEI, agradecem o empenho de todos quantos participaram nesta realização que, esperam, tenha contribuído para uma melhor apreensão dos temas tratados ao longo da unidade curricular, nos domínios das metodologias de investigação científica e das boas práticas de escrita de documentos relacionados com essa investigação.

Eugénio Oliveira e A. Augusto de Sousa,
(Docentes de MIC: Metodologias de Investigação Científica, Programa Doutoral em Engenharia Informática) Janeiro 2011

CONFERENCE COMMITTEES

STEERING COMMITTEE

A. Augusto Sousa
Eugénio Oliveira

ORGANIZING COMMITTEE CO-CHAIRS

Jorge F. P. G. Ribeiro Teixeira
José Augusto de A. Monteiro
José Carlos Lobinho Gomes

ORGANIZING COMMITTEE

Adela Antónia Ortiz Castillo
Alex Fernando de Araújo
Ali Azarian
Carlos Daniel S. Constantino
Carlos Jose Ribeiro Campos
Carlos Manuel Dias Viegas
Helder Martins Fontes
João Emilio S. C. Almeida
João Tiago Pinheiro Neto Jacob
Jorge F. P. G. Ribeiro Teixeira
José Augusto de A. Monteiro
José Carlos Lobinho Gomes
Luis Filipe Guimarães Teofilo
Mario Salvador G. Rodriguez
Nima Shafii
Nuno Jorge Pereira Saleiro
Nuno Miguel Monteiro Barbosa
Paulo Manuel Ferreira Neto
Pedro Amorim Brandão da Silva
Pedro dos Santos Saleiro da Cruz
Pedro Filipe de Monteiro Rocha
Pedro Miguel Moreira da Silva
Rui Alberto Ferreira de Castro
Rui Jorge Martins da Silva Chilro
Tiago Pinto Fernandes
Vítor Manuel Rodrigues da Cunha
Zafeiris Kokkinogenis

SCIENTIFIC COMMITTEE CO-CHAIRS

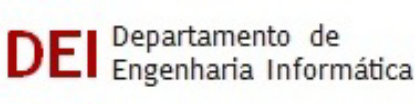
Carlos Manuel Dias Viegas
Mario Salvador G. Rodriguez
Rui Jorge Martins da Silva Chilro
Tiago Pinto Fernandes

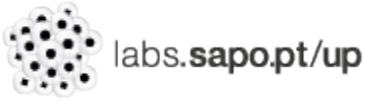
SCIENTIFIC COMMITTEE

Adela Antónia Ortiz Castillo
Alex Fernando de Araújo
Ali Azarian
Carlos Daniel S. Constantino
Carlos Jose Ribeiro Campos
Carlos Manuel Dias Viegas
Helder Martins Fontes
João Emilio S. C. Almeida
João Tiago Pinheiro Neto Jacob
Jorge F. P. G. Ribeiro Teixeira
José Augusto de A. Monteiro
José Carlos Lobinho Gomes
Luis Filipe Guimarães Teofilo
Mario Salvador G. Rodriguez
Nima Shafii
Nuno Jorge Pereira Saleiro
Nuno Miguel Monteiro Barbosa
Paulo Manuel Ferreira Neto
Pedro Amorim Brandão da Silva
Pedro dos Santos Saleiro da Cruz
Pedro Filipe de Monteiro Rocha
Pedro Miguel Moreira da Silva
Rui Alberto Ferreira de Castro
Rui Jorge Martins da Silva Chilro
Tiago Pinto Fernandes
Vítor Manuel Rodrigues da Cunha
Zafeiris Kokkinogenis
Ana Cristina R. Paiva Pimenta
Ana Paula Cunha da Rocha
A. Fernando V. C. Castro Coelho
António Jesus Monteiro de Castro
António Manuel Correia Pereira
A. Miguel P. Pimenta Monteiro
Claudia Melania Chituc
Gabriel de Sousa Torcato David
Henrique Daniel de A. L. Cardoso
João Carlos Pascoal de Faria
João Manuel Paiva Cardoso
João Pedro C. Leal Mendes Moreira
Jorge Manuel Gomes Barbosa
Luís A. Diniz F. de Morais Sarmento
Luis Paulo Gonçalves dos Reis
Maria Eduarda S. Mendes Rodrigues
Pedro A. G. L. Ferreira do Souto
Rosaldo José Fernandes Rossetti
Rui Filipe L. Maranhão de Abreu
Sérgio Sobral Nunes
Teresa Cristina de Sousa Azevedo

SPONSORS

DSIE'11 – Doctoral Symposium in Informatics Engineering is sponsored by:





WELCOME MESSAGE

Dear Dean, dear Students and Participants, dear Guests,

In the name of ProDEI Scientific Committee, it is my pleasure to open this Doctoral Symposium on Informatics Engineering, to welcome you all, and to have the opportunity to address just a few words at this opening session.

I, firstly, would like to thank the presence of the FEUP's Dean and Informatics Department director, as well as the invited speaker, professors and, what is most important, all the participants who are crucial to this event, you the PhD students.

We are here for the 6th Workshop of this kind, organized by PhD ProDEI (Doctoral Program in Informatics Engineering) students at FEUP, in the scope of the "Scientific Research Methods" course.

Main goals of this event simultaneously include the enhancement of students scientific expertise in several Informatics related subjects, the use of adequate formats for both conveying ideas and to write scientific papers, as well as the capability to organize and deal with the many different aspects that are implied by a scientific meeting organization.

ProDEI program, as an all, already includes about 90 active PhD students. However, what really measures the impact of a program like this is the number of the students that really finish their PhD.

We are now seeing students PhD thesis being concluded and at least 14 have been successfully defended. Some of these Dissertation defenses had the presence of well-known international experts on the field, assuring us that ours are international quality standards. ProDEI is proud about this success.

This Symposium includes 7 sessions concerning different and somewhat broad subjects like: Information Systems, Data Processing, Artificial Intelligence, Computer Graphics, Mobile Computing, Configurable Computing and Semantic Web.

We are not here expecting outstanding scientific results although we do expect the rise of a genuine interest in scientific topics to be investigated further. I urge all of you to contribute to a fruitful discussion here, at this meeting, about your topics on the Informatics Engineering area, which encompasses both Computer Science and Computers Engineering. Be curious and not shy. Please make the symposium LIVE.

I would like to finish by strongly acknowledge all the ProDEI PhD students and, if you allow me, to thank in particular those who took in charge the Chairs of both the Scientific and Organizing Committees. All of you have done an excellent job so far.

Thank you and enjoy the opportunity.

Eugénio Oliveira

ProDEI Director

CONTENTS

TECHNICAL PROGRAMME

SESSION 3 - ARTIFICIAL INTELLIGENCE

SESSION 4 - COMPUTER GRAPHICS

SESSION 5 - COMPUTER GRAPHICS / MOBILE COMPUTING

SESSION 1

INFORMATION SYSTEMS / DATA PROCESSING

Chairman: Nuno Jorge Pereira Saleiro

Rui Alberto Castro
School Performance Evaluation in Portugal: A Data Warehouse Implementation to Automate Information Analysis

José Augusto Monteiro
The Guardian of the Republic: A conceptual system to detect outliers on Public Contracts

Jorge Teixeira
Automatic Generation of a Training Set for NER on Portuguese journalistic text

Pedro Saleiro
Web sessions clustering for behavioral targeting

School Performance Evaluation in Portugal: A Data Warehouse Implementation to Automate Information Analysis

Rui Alberto Castro

Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, s/n, 4200-465, Porto, Portugal
pro10022@fe.up.pt

Abstract. School performance evaluation is nowadays a hot subject both in political and pedagogical terms. The ability to perform comparative analysis between different schools in different regions using the national exams results is becoming an important issue for school management boards. Although this inter-school comparison (usually named school rankings) is very important, we believe that also intra-school comparison is of major importance as the most significant decisions that affect student performance are made by teachers and school directors inside a particular school. This paper presents the implementation of a Data Warehouse system using the Dimensional Model suited for inter-school and intra-school performance analysis. The used Data is real data that comes from the Ministério da Educação and from a specific school. We run a set of analysis that show both the usability of the built system and the influence of internal factors (teachers, courses) in final results. We present also some comparisons between SQL queries built over traditional relational model and built over our Data Warehouse system.

Keywords: Data Warehouse, Dimensional Model, SQL, School Ranking.

1 Introduction

Present days are characterized by the importance of accountability in all aspects of life. The need to measure, compare and rank are spread out over knowledge areas that just a few year ago were not used to it. In Portugal, pre-university school exams results has been published by the Ministério da Educação (ME) [1] since 2002 and since then, the so called school rankings are built and published every year by different sources mainly the media [2]. Although this inter-school comparison, usually named school rankings, is very important we believe that also intra-school comparison is very significant. This inter-school analysis of student results using internal factors such as the student teacher, the student course, the student background in terms of social and economic factors and the student history in terms of past

schools is of major importance as the most significant decisions that could affect student performance are made by their own teachers and by the school directors inside a particular school. Also is at school level that most important improvement measures can be taken in order to present better results.

To perform the inter and intra-school analysis some difficulties arise due to the very different data sources and incompatible data systems. In order to overcome this problem and also to provide a simple query system that even, in future work, could be transformed in an automatic query system tool, we present the implementation of a Data Warehouse system using the Dimensional Model [3] suited for inter and intra school performance analysis.

To test and evaluate the proposed model, we loaded it using real data from exams provided by the Ministério da Educação [1] and with data from a specific school (Colégio Paulo VI de Gondomar [4]). We present some of the major steps of the ETL process (Extraction, Transformation and Loading) and address some of the important issues of data validation and consolidation.

The main goal of this paper is to show a system suited for inter and intra school performance analysis and to demonstrate its feasibility with a particular example. Some performance analysis using the system are provided in order to compare its usability and performance.

This document has been structured in following order: In Section 2 we will present other school performance studies and techniques, in Section 3 a general view of the Data Warehouse system and the Dimensional Model applied to our particular case, in Section 4 we describe how the system has been implemented and how the real data has been loaded, in Section 5 we will present some results and query comparisons and in Section 6 we will present some conclusions.

2 Related Work

The school performance analysis in Portugal is a recent topic as only since 2002 the Ministério da Educação de Portugal has disclosed the final exams results of the Portuguese students. Since then the so called school rankings are built every year based on the final marks in 12º year and in university access exams.

In foreign countries is very usual to make school rankings but, as in Portugal, they make an inter-school analysis based on exams or standard tests. In UK, for example, is usual to publish school rankings based on GCSE (General Certificate of Secondary Education) [6].

A lot of different institutions, mainly the press, make these rankings in the past years but there is no known publication of an analysis of intra-school performance and its relation with inter-school rankings of a Portuguese school.

Some reference press publish every year a very deep inter-school analysis [5] and compare the results over the years decomposing in some key factors such as regional distribution, school size and student course. We will extend this work to the intra-school factors that we think are determinant to student performance, mainly the teacher.

3 Data Warehouse System Description

The main objective of an implementation of a Data Warehouse System is to make the information available and increase its quality, reliability and usability. The final goal of the system is to provide information that can drive the school managers to better and more fundament decisions.

Our system is made around two stars (fact tables) with three and five dimensions respectively. Star one, shown in figure one, stores all the exams results of all students of all schools since 2002.

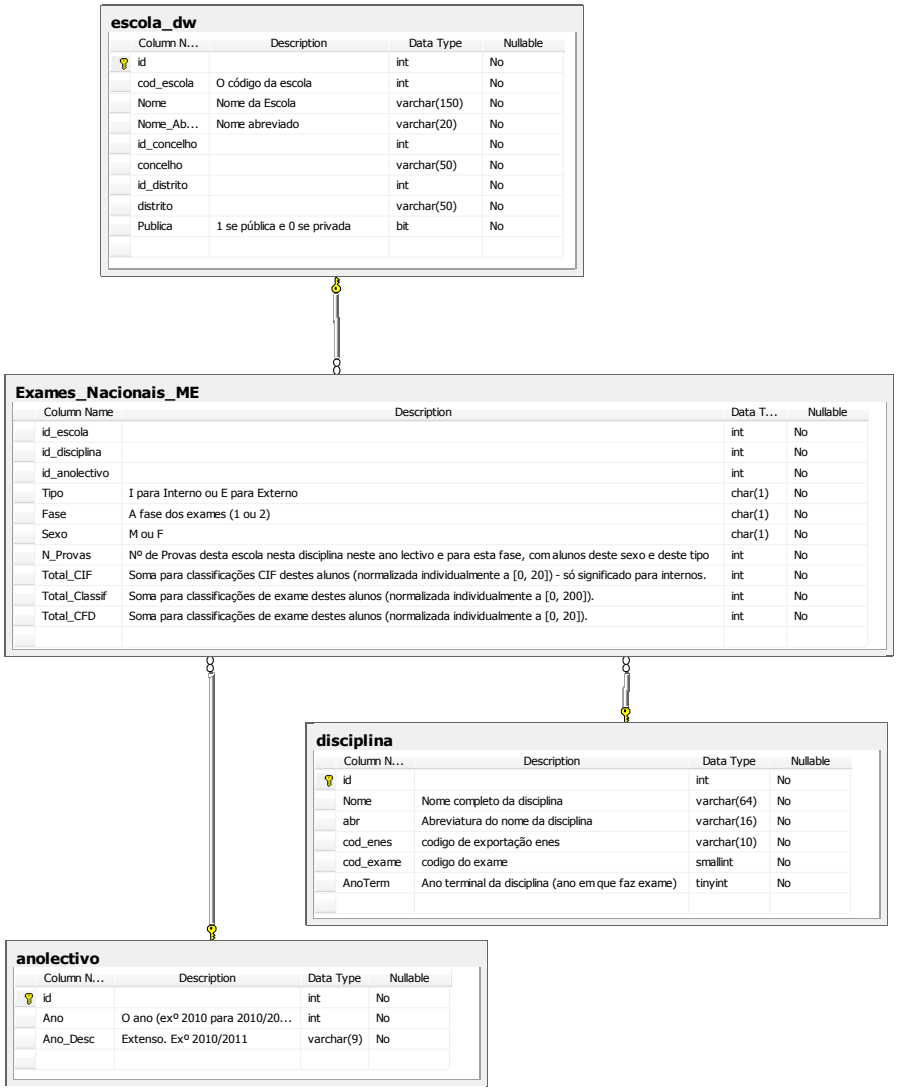


Fig. 1. Star One. Results from national exams.

Star two, shown in figure two, stores all the grades and exams results of all students of a specific school since 2002.

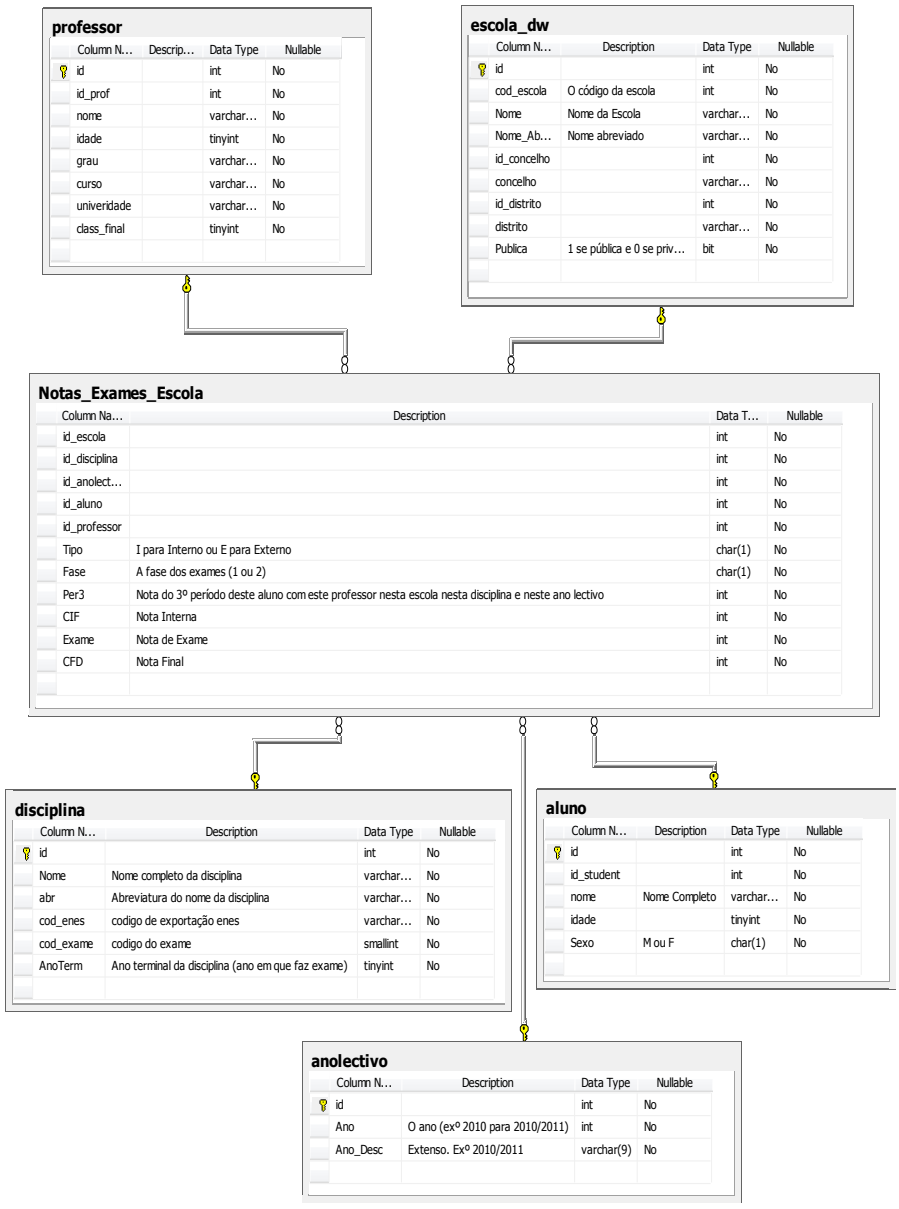


Fig. 2. Star Two: Results of a specific school

The fact tables store all grades and exams:

- Star One stores the results from the national exams.
- Star Two stores the grades and exams of the specific school.

Both stars share three common dimensions:

- anolectivo: Stores the school year to which the grades refers.
- escola_dw: Stores data about a specific school namely its name, code, geographical location and type (public or private).
- disciplina: Stores all information about all disciplines, namely their name, cod_enes (a specific code administered by the Ministério da Educação (ME) and used for electronic export), cod_exame (a specific code administered by ME that specifies the exam) and the year of discipline termination.

Star Two has two more dimensions:

- aluno: Stores information about students.
- professor: Stores information about the student teachers.

The star one (national exams) has only aggregate results as ME doesn't disclose specific student information, that is there is no way to relate the results from one student from one year to the other. Having this in mind we aggregate all the information across the specific dimensions (school, discipline and year) and also across sex, phase (phase 1 or phase 2) and type (Internal or External student) as these are the only information available.

The star two (specific school) stores all the results with no aggregation as we have enough information to make studies along the years. Each fact is a grade and exam of a particular student.

With the proposed system a lot of interrogation about national exams are possible across years, school, regional location, type of school, discipline, phase, type of student and student sex. In addition to the above interrogations, we are able also to get analysis about the specific school data across student, teacher and teacher specific data as age, degree and university.

4 Data Warehouse Implementation

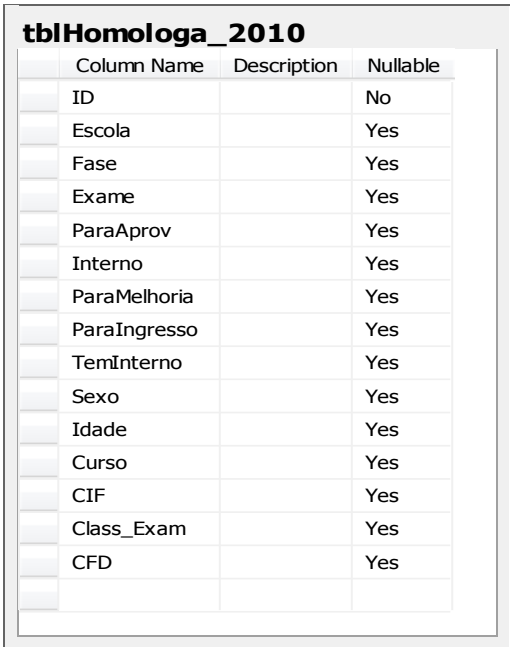
To implement the proposed system we have to make two main phases:

- The ETL process: Data Extraction, Transformation and Loading.
- Data Verification and Validation.

4.1 The ETL Process - Data Extraction

Data extraction is, in this case, a straightforward process. Both the national exams data and the specific school data are already in electronic format and are directly importable to the SQL Server System. The ME data is available online [1] in Access

tables format that are directly importable to SQL Server. There are nine tables (tblCodsConcelho, tblCodsDistrito, tblCodsPubPriv, tblCursos, tblCursosSubTipos, tblCursosTipos, tblEscolas, tblExames e tblHomologa2010 (for the 2009/2010 lecture year). The Colégio Paulo VI (CPVI) school has already a relational Database implemented and has a set of 47 different tables from where de data should be extracted. Figure 3 show as an example the table design directly extracted from ME exams data.



tblHomologa_2010

Column Name	Description	Nullable
ID		No
Escola		Yes
Fase		Yes
Exame		Yes
ParaAprov		Yes
Interno		Yes
ParaMelhoria		Yes
ParaIngresso		Yes
TemInterno		Yes
Sexo		Yes
Idade		Yes
Curso		Yes
CIF		Yes
Class_Exam		Yes
CFD		Yes

Fig. 3. Design of the ME exams data table directly extracted

4.2 The ETL Process - Data Transformation

The Data Transformation phase was more complex due to ambiguity of data, specially on courses definition, and lack of information on ME tables. In fact, ME tables define course by a name and an exam code and CPVI school define course by a name not always equivalent to the ME one and a ENES code which has no automatic conversion in an exam code.

Due to these ambiguities a semi-automated process has been used with some manual aid. Here is an example of a SQL code for an automated transformation of data and its storage in the transformation area using a temporary table named `_temp_tab_course`:

```
select distinct course.name, abr, cod_enes, termina,
    class.year as AnoD,cast('' as varchar(200))as
    nome_tbl,cast(0 as integer)as CodExame, cast('' as
    varchar(25))as AnoTerm, 0 as OkFlag
into _temp_tab_course
    from course, year_area, year, class_course, class
where course.id_year_area=year_area.id and
    id_year=year.id and year.year>=2001 and exame=1
    and id_course=course.id and id_class=class.id
    and (termina=class.year or termina=0)
```

Here is the SQL code for the automatic math for data from the school and from ME:

```
update _temp_tab_course
    setnome_tbl=descr,codexame=exame,
    anoterm=[anoterminal], okflag=1
from tblexames t
where name=descr
```

As an example, with this code we are able to get 28% of automatic finalization and the remain 72% needed manual intervention and validation.

4.3 The ETL Process - Data Loading

Having done the complex data transformation process for all the ME exams and CPVI data, the task of loading the final tables data is a simpler process. Here is some examples of SQL code to load some of the dimensions:

```
--Insert the anolectivo Dimension
insert anolectivo
    select year,cast(year as varchar(4)) + '/' +
          cast(year+1 as varchar(4))
    from year
    where year>=2002
    order by year

-- Insert the disciplina Dimension
insert disciplina
    select name,abr,cod_enes,codexame,termina
    from _temp_tab_course
    where okflag=1
```

Here is as an example the code to load the fact table Star One (national exams results). In this example we are loading the data for the year 2009/2010. A similar SQL expression is used for other years:

```
insert exames_nacionais_me
    select escola_dw.id, disciplina.id, anolectivo.id,
          'I', fase, sexo, count(*), sum(cif),
          sum(class_exam), sum(cfd)
    from tblhomologa_2009, escola_dw, disciplina,
          anolectivo
    where ano=2009 and escola=cod_escola and
          cod_exame=exame and interno='s'
    group by escola_dw.id, disciplina.id,
          anolectivo.id, fase, sexo
    union
    select escola_dw.id, disciplina.id, anolectivo.id,
          'E', fase, sexo, count(*), sum(cif),
          sum(class_exam), sum(cfd)
    from tblhomologa_2009, escola_dw, disciplina,
          anolectivo
    where ano=2009 and escola=cod_escola and
          cod_exame=exame and interno='n'
    group by escola_dw.id, disciplina.id,
          anolectivo.id, fase, sexo
```

4.4 Data Verification and Validation

After all data loading process is completed, a new phase of verification and validation is done. This is an automatic process done by some SQL processes in order to verify that all valid data is loaded in the Data Warehouse and that the loaded data is consistent with the data sources.

Here is a sample of the SQL process code to perform this task:

```
select escola,nome,count(*) as Ex_Number,
    (select count(*) from exames_nacionais_me
    where id_escola=escola_dw.id)
from tblhomologa_2010, escola_dw
where escola=cod_escola
group by escola, nome, escola_dw.id
having count(*)<>
    (select count(*) from exames_nacionais_me
    where id_escola=escola_dw.id)
```

5 Results and Query Comparisons

In this section we present some examples of queries to our implemented system, its results and a comparison between that query and the equivalent in the CPVI system (traditional relational database). Here are some examples:

- Grades and exams of all students of CPVI in MAT and in 2009/2010 (Data Warehouse Version):

```
select ano_desc,abr,anoterm,professor.nome,per3,cif,
    exame,cfd
from notas_exames_escola,professor,disciplina,
    escola_dw, anolectivo
where id_disciplina=disciplina.id and id_anolectivo =
    anolectivo.id and id_professor=professor.id and abr
    = 'mat-a' and ano=2009
```

- Grades and exams of all students of CPVI in MAT and in 2009/2010 (Relational Database Version):

```
select year.year,abr,class.year,class.name,code,nf,cif,
    exame_f1,exame_f2,cfd
from student_bio,class_course,course,year_area,class,
    year,prof_bio,prof
where year.year=2009 and abr='mat-a' and class.year=12
    and cif>=10 and course.id_year_area=year_area.id
    and id_course=course.id and id_class=class.id and
    student_bio.id_class_course=class_course.id
    and year_area.id_year=year.id
    and prof_bio.id_class_course=class_course.id and
    id_prof=prof.id
order by class.year,class.name
```

Table 1. Grades and exams of all students of CPVI in MAT and in 2009/2010 (sample).

year	*abr*	*code*	*nf*	*cif*	*exame_f1*	*exame_f2*	*cfd*
2009	MAT-A	JC	10	11		121	11
2009	MAT-A	JC	15	15	171		16
2009	MAT-A	JC	17	16	147	135	16
2009	MAT-A	JC	14	14	147	156	15
2009	MAT-A	JC	20	20	196		20
2009	MAT-A	JC	9	11	96		11
2009	MAT-A	JC	19	19	178		19
2009	MAT-A	JC	18	19	185	177	19
2009	MAT-A	JC	15	15	160		15
2009	MAT-A	JC	17	18	196		19

- Top 12 schools with more exams in 2009/2010:

```
select top 30 nome,count(*) Total_Exames_Nacionais
from exames_nacionais_me,anolectivo,escola_dw
where id_anolectivo=anolectivo.id and
    id_escola=escola_dw.id
    and ano_desc='2009/2010'
group by nome
order by count(*) desc
```

Table 2. Top 12 schools with more exams in 2009/2010.

Escola	*Nº de Exames em 2009/2010*
Escola Secundária Camões	591
Escola Secundária Alberto Sampaio	522
Escola Secundária Jaime Moniz	482
Escola Secundária Santa Maria de Sintra	482
Escola Secundária da Amadora	466
Escola Secundária Alexandre Herculano	459
Escola Secundária de Odivelas	454
Escola Secundária Maria Amália Vaz de Carvalho	443
Escola Secundária Leal da Câmara	429
Escola Secundária Avelar Brotero	428
Escola Secundária de Cascais	428
Escola Secundária Alves Martins	419

- Year by year analysis of the last three years performance of CPVI Math teachers:

```
select ano_desc,disciplina.nome,id_prof,avg(exame)
    Média_Exame_Prof
from notas_exames_escola,anolectivo,disciplina,
    professor
where id_anolectivo=anolectivo.id and
    id_disciplina=disciplina.id and
    id_professor=professor.id
    and disciplina.abr='mat-a'
group by ano_desc,disciplina.nome,id_prof
order by ano_desc,avg(exame) desc
```

Table 3. Three year analysis of CPVI Math teachers performance.

Ano Lectivo	*Disciplina*	*Professor*	*Média Exame*
2006/2007	Matemática A	6	124
2006/2007	Matemática A	21	109
2006/2007	Matemática A	157	82
2007/2008	Matemática A	6	163
2007/2008	Matemática A	21	115
2007/2008	Matemática A	157	102
2008/2009	Matemática A	41	158
2008/2009	Matemática A	21	151
2008/2009	Matemática A	157	122
2009/2010	Matemática A	6	143
2009/2010	Matemática A	41	140
2009/2010	Matemática A	21	115

6 Conclusions

Nowadays all the human activities are expected to be evaluated and compared. In this paper we present an implementation of a Data Warehouse system suited for school performance evaluation and we present some results obtained from the proposed system. With the presented system we are able to obtain an enormous different figures and tables suited for a managerial and pedagogical analysis.

The advantages of building a system as the one proposed are mainly three:

- Increase Information reliability, confidence and completeness
- Protect the information
- Make available in an easier way a big amount of information

With the proposed system a set of different results can be obtained in order to analyze the school performance from different points of view. In section 5 we present some examples and show the simplicity of obtaining a lot of results using the proposed model.

As future work and due to the Dimensional Model low complexity we will be able to provide an implementation of automatic analysis and interrogations suited for online site results presentation.

References

1. Ministério da Educação de Portugal. Direcção Geral de Inovação e Desenvolvimento Curricular - DGIDC. Exams Data: http://www.dgidc.min-edu.pt/JNE/Paginas/estatistica.aspx.
2. Sociedade Independente de Comunicação - SIC. Ranking das Escolas: http://sic.sapo.pt/online/noticias/pais/especiais/ranking-escolas-2010/.
3. Ralph Kimball, Margy Ross: The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, ISBN: 0471200247. John Wiley & Sons, 2nd edition, 2002 pp 37-89.
4. Colégio Paulo VI de Gondomar: http://www.colegiopaulovi.com.
5. Jornal Público - Dossier with School Ranking: http://static.publico.clix.pt/docs/educacao/especiaranking2010.pdf.
6. The Telegraph Journal - GCSE league tables 2010 school-by-school: http://www.telegraph.co.uk/education/leaguetables/8254332/GCSE-league-tables-2010-school-by-school.html.

The Guardian of the Republic: A conceptual system to detect outliers on Public Contracts

José Augusto Monteiro

Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, n/n, 4200-465, Porto, Portugal
jose.monteiro@fe.up.pt

Abstract. The paper focuses on the processing of the information available in government portals. Nowadays, in an e-Government, common citizens can have access to various contents, such as government decisions, government acts, court decisions, legislation and other information displayed on the official portals. However, the high volume of contents and the way the information is presented raises issues about their utility and validity to the community. The problem is how to process the available information and make it useful for citizens. In our work, we suggest an improvement of the present solutions by proposing a conceptual system to automate the processing of the information, analyze patterns, detect outliers, and send alerts about the deviations found. A preliminary analysis of the problem reveals that the type and the volume of the dataset are important factors to select the more appropriated algorithm in outlier detection.

Keywords: Outlier detection, Support vector machines, Learning machine, Training data, Natural language processing.

1 Introduction

In this paper we will focus on the detection of outliers on dataset of the public contracts.

Nowadays, one of the most common problems that citizens claim for resolution is the control of the huge public expenditure. In one hand, some expenses do not seem necessary. On the other hand, some expenses are necessary, but are much more expensive than expected. It is precisely to control the last mentioned that we propose the building of a system.

A preliminary investigation of the problem raised the following issues: i) how do we know what are the acceptable values for a given category of expenses? ii) how to define a pattern of expenses? iii) How to detect expenses that are out of the pattern and that require attention?

On our dataset we have two relevant data fields: one with a description of the subject of the contract and, another field with a price value. These are the elements that help us judging if the transaction is correct or not. To a human, even if he does not know anything about some kind of good, with some research he will be capable of judge a specific situation. To a computer, it is more complicated. It is necessary to provide an example so that the machine can compare and make a correct judgment.
For example: A contract where the object is a vehicle acquisition. Knowing the characteristics of the vehicle, how do we know the mentioned price on the contract is the normal price for that type of vehicle or if it is too expensive?
On the dataset chosen for this research work the object of the contract describes goods and services. Some examples of goods can be: computers, vehicles, office material among others. Some examples of services can be: vehicle maintenance, rehabilitation of buildings, consultancy, among others. The description of goods or services that are the object of the contract can be defined by a small sentence that can have a few words like "*Aquisição de viatura*" or a long sentence like "*Fornecimento de uma viatura, do tipo ligeiro de passageiros, com retoma de uma viatura, aquirida em 1995 para a Segurança Social em Angra do Heroísmo*". Also, misspelled words can be found in the description. There are no specific fields where the type or the class of the object of the contract is defined within atomic data. There is not either a field where the quantity of some good acquisition is specifically defined. Continuing with the example of the vehicles, in an acquisition of multiple vehicles, the object of the contract can be described by an expression like: "*2 viaturas Opel Vivaro L1 H1 Combi 9L 2.0 CDTI 114 cv*". Another different description for a multiple vehicle acquisition is: "*Aquisição de viaturas Toyota Procº C - 55/09*". When we try to compare the unitary values of both contracts, we cannot do an objective comparison. In the first example it is clear that we can divide the price value by two. In the second example we know that we have multiple vehicles, but we do not know how many. The absence of a normalized description of the object of the contract can result in ambiguous analysis.
The main goal of this research work is to conceptualize a system that helps to detect outliers on the dataset of the public contracts. This problem is known as an outlier detection problem.
Starting from the point that we do not know what we are looking for, we have to deal with unlabeled data [7]. Also we need an algorithm that helps to identify what are the patterns and which are the points out of the patterns (outliers).
Outlier detection is a technique that helps to discover fraud detection or "strange" behaviors on areas such as computer intrusion, financial analysis, among several others. On appliance, it differs between users or datasets.
This document has been structured in the following order: In the Section 2 there will be presented other studies and techniques used in outlier detection, in the Section 3 a conceptualization of the system, in the section 4, we describe how the system will be evaluated, in the Section 5 we describe how the system will be implemented, and in the final section we will present some preliminary conclusions.

2 Related Work

To try to solve outlier detection problems, different approaches have already been tested according to pre-defined concepts. Some of those, such as distance-based referred by Tang et al. (2007), and density-based referred by Zhu et al. (2010), are more recent. However, there are others less recent, such as support vector machines (SVM) [4], [2], [6], that still remain valid. In fact, this binary classification algorithm is a standard tool for machine learning and, according to Platt (1998), Hearst et al. (2002), and more recently Zhu et al. (2010), it is known for having a very good performance. Another method is transductive confidence machines proposed by Barbara et al. (2006), that uses an "extra" dataset as the outlier reference [1].
Comparing outlier detection algorithms, density-based and the connectivity-based, Tang et al. (2007) consider that the density-based schemes are effective in an environment where patterns possess sufficiently higher densities than outliers, while the connectivity-based scheme works better for isolated outliers that possess comparable densities with the patterns [5].
As referred in the introduction section, we do not know yet what will be the behavior of our dataset in the outlier detection process. However, considering our case, the outlier detection represents a well defined problem. Regarding the cases that may be a fraud, we expect that our outlier had a low density. Also, it will be expected a pattern with a low density too. Starting from this assumption, these two methods are not the most appropriate to use in a prototype of our system.
We also have compared another two methods that have in common the particularity of having what we call "assisted" outlier detection: i) transductive confidence machines; ii) outliers by example.
In one hand, the transductive confidence machines method processes every point in a separate dataset and decides individually which point is an outlier according to an existing clustering model. To detect outliers in a dataset, it processes both datasets, the training and the test [1]. On the other hand, the method used by Zhu et al. (2010), outliers by example (OBE) which combine the SVM algorithm with the user relevance feedback, has proven to be adequate to deal with a low dimension datasets [8]. The first is "assisted" by a model and the second is "assisted" by the user for example. Despite the similarity on the main idea of these two outlier detection methods, the OBE approach appears to be more flexible and more adequate to implement in our system. Regarding the characteristics of our dataset (low density, ambiguous characterization of object of contract), we believe that the models based on user feedback could be more adequate to use in our system in the prototype stage to help training data. In the prototype stage we will use the OBE algorithm.

3 System Description

In this section we start by describing our system as a system that is capable to detect outliers from unlabeled data.

In a broad sense the system follows the conventional patterns (input, processing, and output):

- Data input: Dataset obtained from the site of the public contracts[1]. The information domains are composed by information of public entities, provider entities and the contract that involves both;
- Processing data: Querying contract description field from dataset, using a single word or word combination, the system should return a subset of registers described by contract description (CD) field and price value (PV) field. Starting from the returned subset, system should training data to create a pattern which will be the reference to detect the outliers.
- Output data: The processed data should return results that confirm or not the existence of outliers on a given category of goods.

The initial results that we expect to obtain with the system will depend on the quality of the dataset. In the next paragraphs we provide an overview of how the information that the system has to deal, is organized on the chosen dataset.
Describing the two fields (CD, PV) that will be used in outlier detection analysis, PV data are already in an atomic format. However, as we referred on introduction section, CD data is based on a set of words (string format) that seems not follow a well defined pattern, lexical or grammatical rules, as can be observed in figure 1. To a better understanding of how the system should work, we support our explanation on vehicle acquisitions. This will be the chosen subject for testing a prototype of our system.

Label
Aquisição de viatura ligeira de passageiros
Aquisição de viatura Renault Kangoo Confort 1.5 DCI
Aquisição de viatura ligeira.
Aquisição de viaturas marca Toyota
Aquisição de viaturas marca Opel
Aquisição de viatura Renault Grand Espace
Aquisição de viatura ligeira - Volkswagen Passat Limousine 2.0 TDI 140 cv Confortline
Aquisição de viatura ao abrigo dos contratos públicos de aprovisionamento (contrato n.º 412027)
Aquisição de viatura de 9 lugares
Aquisição de viatura de serviço
Aquisição de Viatura Mitsubishi L200 4*2 –Contrato DGP 412045 n.º ordem 420

Fig. 1. An example of a subset related with vehicle acquisition.

Starting by giving an expression to obtain information about the vehicle acquisition (*aquisição viatura*), the system should return a subset of registers where terms "*aquisição*" and "*viatura*" are included in the CD field. From the returned subset are expected that the system creates a pattern of prices that will be the reference value for the chosen expression. If some register deviates from the created reference, it could be an outlier.

[1] http://www.base.gov.pt

As referred before, we may experience some problems caused by the meaning of terms combination. Imagine that we want to detect outliers in a category of vehicles that have an engine with 1.500 cm^3. Querying dataset with expressions like "*1.5 cm3*", "*1.500 cm3*" or "*15 cm3*", may not return the desired results. In the example of the figure 1, we can observe that some of the vehicle acquisitions do not mention vehicle characteristics. Also some registers refer more than one vehicle, but is not clear about how many vehicles has been buy in that specific contract.
To have a more precise idea how many registers we have on dataset related with vehicles, we query dataset with a single term and have obtained the presented results:

- viatura: 2125 registers;
- carro: 74 registers;
- veículo: 252 registers;
- ligeiro: 59 registers;
- pesado: 23 registers;
- cilindrada: 6 registers;
- cm3: 3 registers;
- cv: 33 registers

The returned registers for the term "*viatura*" are all related with vehicles. However, they relate to different classes of situations connected with the vehicles: acquisitions, renting, maintenance, etc. results. The returned results from other terms are less precise and revealed that a part of the returned registers is not related with vehicles.
The PV of the contracts for a certain type of query can have a wide range. For the term "*viatura*" we can find a register that returns a description of a utility vehicle acquisition that costs about 15.000€, or for the same term, the returned register may describe an acquisition of a fleet of vehicles that costs 300.000€. Also, there can be found registers that the PV is zero or "*NULL*". With these examples the system may experience more difficulties to detect the outliers, because in some registers there are no defined contexts, categories, or quantities. To obtain more efficient results it seems more adequate that the system should use the term combination.
To evaluate the quality of the returned results with term combination, it has been done the following queries that returned the presented results:

- aquisição viatura: 159 registers;
- ligeiro passageiros: 12 registers;
- veículo ligeiro mercadorias: 7 registers;
- veículo mercadorias: 12 registers;
- veículo ligeiro: 33 registers.

In these given examples, all the returned registers are related with vehicles. Nevertheless, only the first query has returned results that match the desired context. We also observe that the query "*veículo ligeiro*" combines results from the query "*veículo ligeiro passageiros*" and "*veículo ligeiro mercadorias*". As mentioned before, we still cannot differentiate the characteristics of vehicles, quantity, or what type of commercial transaction it refers to (acquisition, renting, maintenance, etc.).

Beyond the absence of a clear detail in the object of the contract description, we also identify a problem that is related with the semantics concepts.
In order to increase the overall precision of the system, we need to give positive and/or negative examples to the system, thus building a training set that will be used in the learning of a OBE algorithm.

4 Evaluation

On this particular problem, the evaluation of the system is indexed to human utility. In fact, when a human evaluates a system results he may judge in a qualitative perspective (ex.: how satisfied he is with the results of the system), or in a quantitative perspective (ex.: which are the percentage of the correct results). In the prototype phase, the evaluation of our system will use a manual method. The chosen option is justified by the prediction of a certain ambiguity that may occur in certain cases. For example: In a vehicle acquisition, given the power of the engine and the price value, if the system returns an outlier, in fact, the returned register may be not an outlier. The vehicle may have some special characteristics that justify the high price value and not make it an outlier. For these particular situations, human judgmental appears to be the accurate option.
The chosen dataset has about 200.000 registers and tends to increase . Nevertheless, only a small part of the registers are related to a subject. For example, to the subject "vehicles" it is suitable to know with a certain precision how many registers are in dataset. For this subject we estimate that the dataset does not contain more than 2500 registers. The acquisition of the vehicles will be adopted as the subject to the first experimentation tests.
To evaluate the adequacy of the dataset regarding the queries' that need to be made, the adopted metrics are the precision and recall [3]. In the prototype test stage the registers returned as outliers will be evaluated by human judgment as true or false.
To build a fair system evaluation the human judgment should be complemented with other mechanisms that allow cross validation and make the judgment more consistent:

- Human opinion: Composed by 5 people to allow odd results.
- Specialized magazines: To provide a reference in price value for a given situation.
- Newspapers: News with no more than 1 or 2 years maximum about the cases that are fraud suspicious.
- Investigation cases: Known cases under authority investigation.

Those "assisted" mechanisms are expected to refine the system to avoid the situations that we have mentioned on the system description section.
In this starting stage, it was decided to considerer an outlier all objects that the PV exceeds 50% of the medium value of a given object category.

5 Experimental Set-up

To conceptualize our system the problem has been divided in four main tasks: i) Obtain data from website "base.gov.pt" and pre-process the text; ii) Insert clean data on database; iii) Provide an interface to users for querying system; iv) Implement the outlier detection mechanism.

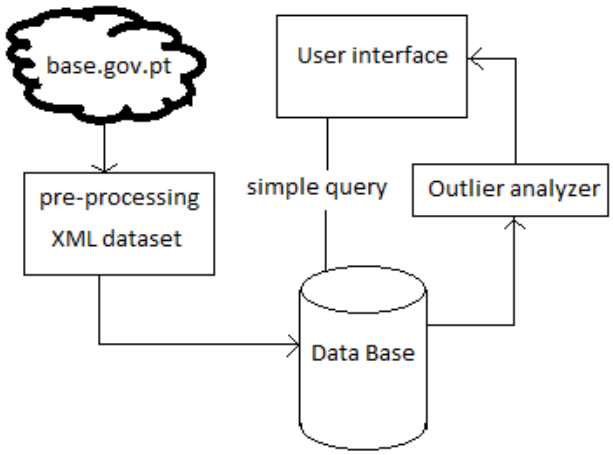


Fig. 2. Overview of conceptual system

In figure 2, the four modules that we indented to implement are represented. On pre-processing module data will be obtained from website and after cleaning HTML tags are converted to a valid XML format. This process allows preserving a local dataset in file format. An example of a register is presented bellow. The second module loads dataset on database for a easier management and analysis. The module outlier analyzer is responsible training data and outlier detection. The user interface module is responsible from the interaction between users and the system.
At this stage it has been completed the first task and the second is under development.
Providing an overview of the dataset, it is composed by two main information domains: i) organizations; ii) contracts.
Organizations contain registers of two types of entities: i) public organizations, composed mainly by government participated institutions, and ii) provider organizations, composed by all type organizations that interact in business market with the public organizations.
The contract is the element that joins both, public and provider organizations, in one business.
Organizations are described by institutional name and tax identification number. The contract is a little more complete. Refer the elements that describe the business and the involved entities.

Example of a register in XML format:

```
<registo id="30590">
  <dataregisto>27-02-2009</dataregisto>
  <numeroprocedimento>30581</numeroprocedimento>
  <adjudicantes>
    <nif>600000117</nif>
    <nome>DGSP-Estabelecimento Prisional de
    Izeda</nome>
  </adjudicantes>
  <adjudicados>
    <nif>502544180</nif>
    <nome>Vodafone</nome>
  </adjudicados>
  <objectocontrato>Aquisição serviços-comunicações
  carrinhas celulares</objectocontrato>
  <datacelebracao></datacelebracao>
  <precocontrato>186,10</precocontrato>
  <prazoexecucao>30 dia(s)</prazoexecucao>
  <localexecucao>Bragança</localexecucao>
  <criteriomaterial>Al c) do n°1 do art. 24 do Dec-Lei
  n.° 18/2008 de 29 de Janeiro - Ajuste Directo em
  função de critérios materiais</criteriomaterial>
</registo>
```

.

As can be observed in the XML example above, each register has an unique "id" and the contracts are described by the following:

- Contact number (numeroprocedimento) - CN: represents the official number of the contract on dataset;
- Registration date (dataregisto) - RD: represents the date when the contract has been registered on online;
- Contract description (objectocontrato) - CD: represents the main information about the object of the contract. This is the field where our system gets the information about the object to analyze;
- Contract date (datacelebracao) - DCel: represents the date when the contract has been celebrated;
- Price value (precocontrato) - PV: represents the value of the business. This is the field where our system gets the values that will be or not the outliers.
- Execution time (prazoexecucao) - PE: represents the delay to the contract execution;
- Place (localexecucao) - Pl: represents the place where the contract has been celebrated;
- Legal criteria (criteriomaterial) - LC: represents the legal justification to certain contact celebration conditions. In the major part of the contracts this field could be empty.

- Organizations (adjudicantes), (adjudicatarios): describes each entity and its position on the contract. Each organization is described by tax identification number (nif) and name (nome).

To the outlier analyzer, the data of major importance are unique ID, CD and PV. These are the sources of the data. The information about the contract and the organizations is used to identify the cases that are marked as outliers.

6 Conclusions

Nowadays most systems which provide analysis about government data are private. The information available can be obtained with a simple query to the database that returns a list of registers. This may become limited for who doesn't know where to start looking for in a large volume of registers. Our proposal is to adopt a system supported by a mechanism of outlier detection that helps users to detect more easily cases that may be relevant to analyze.

The data stored on government portals increases almost every day, which means that it is more difficult for users to deal with a large volume of registers when they try to do research on a subject. This emerging problem suggests the necessity of a mechanism to process the stored data that, given a few arguments, will return valid information. On this research it was found relevant work on outlier detection algorithms. Based on literature, the binary classification algorithm based on SVM's appears to be the more efficient and with given proof. However, in our research, given the characteristics of our dataset, we considerer that OBE algorithm may be more capable to deal with a small dataset.

A preliminary analysis of dataset reveals that the data that describes a contract has not enough elements to analyze cases regarding particular characteristics like per example, vehicle acquisitions given the power of the engine. However, in more generalized situations like simple vehicle acquisitions regardless the characteristics, it may be appropriate. Another difficulty found on dataset is the absence of a price value in some registers or even the detail about the quantity.

As a consequence, in some cases, the system may return more false outliers than expected. This will force us to start with a more generic approach.

As future work we intend to: i) implement a prototype; ii) test the prototype; iii) redefine the mechanism of outlier detection according to the suggestions given by the analysis of the results of the tests iv) present the first version of the system.

References

1. Barbara, D., Domeniconi, C., Rogers, J. P.: Detecting outliers using transduction and statistical testing. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 55--64. ACM (2006).
2. Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., Scholkopf, B.: Support vector machines. Intelligent Systems and their Applications, IEEE, 13(4):pp. 18--28 (2002)
3. Manning, C. D., Raghavan, P., Schutze, H.: Evaluation in information retrieval. In An introduction to information retrieval, Online edition, pp. 151--175. Cambridge UP (2009)
4. Platt, J.: Sequential minimal optimization: A fast algorithm for training support vector machines. (1998)
5. Tang, J., Chen, Z., Fu, A., Cheung, D.: Capabilities of outlier detection schemes in large datasets, framework and methodologies. Knowledge and Information Systems, 11(1):pp. 45--84, (2007)
6. Thorsten, J. "SVM-Light Support Vector Machine," SVM-Light Support Vector Machine. [Online]. Available: http://svmlight.joachims.org/. [Accessed: 12-Jan-2011]
7. Yamanishi, K., Takeuchi, J.: Discovering outlier filtering rules from unlabeled data: combining a supervised learner with an unsupervised learner. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 389--394. ACM (2001)
8. Zhu, C., Kitagawa, H., Papadimitriou, S., Faloutsos, C.: Outlier detection by example. Journal of Intelligent Information Systems, pp. 1--31, (2010)

Automatic Generation of a Training Set for NER on Portuguese journalistic text

Jorge Teixeira

Labs Sapo UP and LIACC - FEUP
Faculdade de Engenharia da Universidade do Porto
Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal
jft@fe.up.pt

Keywords: Named Entity Recognition, Conditional Random Fields, Natural Language Processing, Portuguese, Journalistic text

Abstract. *Tracking* names of public personalities from news is nowadays impossible to be performed without semi-automatic techniques, and usually require human intervention for annotation and validation of corpora. The main goal of this paper is to automatically generate a training set of news for Named Entity Recognition on journalist text. This allow us to build the entire pipe-line of a NER with no human intervention. A news corpus, containing 20,000 news crawled from online newspapers, is automatically annotated with a list, extracted from Voxx, of approximately one thousand names of well-known and frequently mentioned people on news. Additionally, we describe examples from this corpus with a set of feature vectors that include syntactic, semantic and structural information of its words. We intend to create a NER system, based on Conditional Random Fields, which is specialized for names of people. We use HAREM (an annotated corpus of Named Entities for Portuguese) as our gold-standard corpus and results obtained for the annotation of the training set have precision value of 95% and recall value of 74%. Regarding the NER system, we obtained values for precision of 78% and for recall of 23%.

1 Introduction

Nowadays the number of news published everyday on the web easily exceed thousands. This enormous amount of data means that all sort of human tasks that concern with the organization of information became hard to solve. *Media Clipping* and *entity tracking* are two examples of typical tasks that are usually performed by experts and implying semi-manual techniques.

One may want to receive (by email, feed RSS, etc.) all the news articles being published where José Sócrates, the portuguese Prime Minister, is mentioned, or even study the different perspectives and evolutions of Obama quotations for the last years. Even though it may seem trivial how to recognize these names, we have to be aware that frequently names appears in *non-typical* structural contexts

(e.g.: "*Ronaldo*: This is the perfect time to go to Camp Nou" versus "*WikiLeaks*: Hillary Clinton continues contacts with foreign leaders"). In many cases, simple hand-crafted rules as typical context that occurs around these names or identify capitalized letters for proper nouns should be enough for the recognition of these specific Named Entities. However, real systems with real text, like news, tend to be much more complex and these techniques are insufficient to deal with this problem. To address this problem, we must use automatic techniques that are able to identify and extract names of public personalities from news.

We thus propose the use of machine learning techniques - Conditional Random Fields (CRF) - to automatically recognize and extract names (proper names) of public personalities on news published on the web. This process has three main steps. In the first one we automatically extract names of people from Voxx[1] - and build a gazetteer of names of people. This gazetteer will be used as our *Initial Set*. Second, we automatically annotated a news corpus C^{news} of twenty thousand news. This annotation is described by a set of rules to perform matches between the Initial Set of names on the news corpus. The annotated corpus will be used on the CRF training process as the set of positive examples. We also describe C^{news} with a set of feature vectors including syntactic information (category, person, number, etc.), semantic information (job, nationality, type of verb, etc.) and structural information (position in the sentence, previous and following words and a marker of begin/middle/end of the Named Entity). Finally we use CRFs to train a model. This model will be evaluated with a test corpus, HAREM (an advanced NER evaluation contest for Portuguese) (Santos et al. [10]).

The remaining of the paper is organized as follows. In Section 2 we discuss some related work. In Section 3 we describe our Method and in Section 4 the Experimental Set-up. The results obtained are described in Section 5, its Analysis and Discussion in presented in Section 6 and the Conclusions and Future Work are presented in Section 7.

2 Related Work

Minkov et al. [7] investigate Named Entity Recognition on a very specific type of text, the email. The authors propose a set of *specialized structural* features for identifying personal names on emails. They used four copora with 573 manually annotated documents. The experiments were performed based on Conditional Random Fields and on four different models that differ on the set of features used. Results obtained for F-measure vary from 68,1 to 91,9. Interestingly, they found Part Of Speech-tagging too noisy to be useful as features. Authors also presented two additional methods for improving the overall performance: one based on the repetition of named entities in emails and other based on a dictionary of names and its variations. Results achieved show that the performance increased significantly.

[1] Voxx is a system that automatically extracts and classifies quotations from online news and is available at http://voxx.sapo.pt

Feature induction and web-enhanced lexicons are two different methods presented by McCallum et al. [4] for Named Entity Recognition with Conditional Random Fields. Feature induction enables the use of richer and higher Markov models, thus offering more freedom to choose which features may be more relevant for the task. Authors claimed that automated feature induction offers both improved accuracy and significant reduction in features count. Web-enhanced lexicons is described by the authors as a method that obtains seeds for the lexicons from the labeled data, and then uses the web to significantly augment those lexicons. Authors presented results on the CoNLL (Conference on Computational Natural Language Learning) 2003 dataset, a corpus of english newspaper articles annotated with four entity categories (PER, LOC, ORG and MISC) and containing 964 documents on the training set. Authors obtained an F-measure of 84,04 on the test set by using CRFs, feature induction and Web-augmented lexicons.

Talukdar et al. [14] also used complex feature generation methods and patterns for the recognition of Named Entities and based their work on the English CoNLL dataset. Nadeau et al. [9] used and unsupervised named-entity recognition approach for generating gazetteers and resolving ambiguity using the Enamex corpus from MUC-7 (Message Understanding Conference). In both cases, results obtained slightly improved the baseline systems.

Jun'ichi et al. [1] considered that these kind of methods required complicated induction of patterns or statistical methods to extract high-quality gazetteers. The authors explored the use of the Wikipedia as external knowledge to improve Named Entity Recognition. The method proposed extracts category labels from Wikipedia based on simple word sequences as "*Jimi Hendrix* (...) was an *American guitarist*". These labels where then used as features in a CRF-based NER model. The authors used the CoNLL 2003 dataset, a corpus of english newspaper containing 964 documents on the training set. This dataset also provides gazetteers files for PER, LOC, ORG and MISC categories. They compared features using the CoNLL gazetteers with those using wikipedia. Results obtained show that the Wikipedia model improved F-measure by 1.58 points from the baseline.

Although Wikipedia is an extremely vast resource of Named Entities, it is a community-edited resource, making its structure not standardize and not as easy to analyze as one may expect. Mikheev et al. [5] mentioned that the compilation of extensive gazetteers is sometimes the bottleneck in the design of NER systems. The authors have shown that, for the MUC-7 test corpus used, it was sufficient to use relatively small gazetteers of well-known names rather than large gazetteers of low-frequency names.

Despite the good quality and diversity of scientific work that has been done for Named Entity Recognition, mainly for the english language, it is hard to adapt these resources and tools for other languages rather than its original ones, and foreign names are rarely included on gazetteers, making their recognition even more difficult for NER systems. Also, these systems usually involve manual annotation of large corpora, thus very human and time consuming.

Regarding the Portuguese language, Sarmento L. [11] developed SIEMES, a named-entity recognition system for Portuguese that relies on a set of similarity rules to base the classification procedure. These rules try to obtain soft matches between candidate entities found in text and instances contained in a wide-scope gazetteer, and avoid the need for coding large sets of rules by exploiting lexical similarities.

Milidi et al. [6] describe several machine learning algorithms (Hidden Markov Models, Support Vector Machines and Transformation Based Learning) for Portuguese NER. The authors used a corpus with 2100 sentences already annotated with POS-tagging and a set of gazetteers for the three main categories (location, person and organization). The first two gazetteers were extracted from the web and the third from a magazine. The training corpus was manually annotated with NER-tags, totaling 3,325 named entities. Results obtained shown that SVM (Support Vector Machines) approach achieved the best results, with F-measure of 88.11%, even though HMM (Hidden Markov Models) alternative gives good results for precision and recall without the support of any specific linguistic intelligence.

Our approach focus on two main issues: (i) automatically annotate a corpus of news with a small and high frequency gazetteer of names of people extracted from news; (ii) develop a NER system specialized on names of people and based on a well stablished machine learning technique - Conditional Random Fields - which will learn a model from the previously annotated training set.

3 Method

In this section we describe both our proposed methods, automatically annotate a corpus of news, our training corpus, and build a NER system for names of people based on CRF.

3.1 Initial Set

The first step to automatically annotate the news corpus with names of people is to build a good *Initial Set* of names. The Initial Set of names will be build with the help of Voxx. Voxx is a system that automatically extracts quotations from online news sources and classifies them according to its topic/theme. The identification and extraction of quotations is performed in such a way that we are able to easily collect the names of public personalities mentioned on such quotations.

Let us consider a name n_i and a set of names $\mathcal{N} = \{n_1, n_2, ...n_i\}$. Based on the information we collected from Voxx, we have a set of names $\mathcal{N}^{Voxx}$, where $|\mathcal{N}^{Voxx}| = 1045$. Our Initial Set $\mathcal{N}^{initial}$ is thus defined defined as $\mathcal{N}^{initial} = \mathcal{N}^{Voxx}$. This set of names is composed by names of people extracted from quotations on news that have at least three or more occurrences. This restriction allow us to build a small but well-known list of names of people, following the idea of Mikheev et al. [5].

3.2 Annotation Process

Let us consider a News Corpus $\mathcal{C}^{news}$, composed by a set of 20,000 news items n_i (totaling approximately 110,000 sentences) obtained from 16 generic content Portuguese online newspapers. Each news item $n_i = (title, body)$ is composed by its title and body, such that $\mathcal{C}^{news} = \{n_1, n_2, ..., n_{1999}\}$.

We intend to automatically annotate this corpus based on the $\mathcal{N}^{initial}$, our list of names. In order to perform this annotation, we use a set of rules, executed by the order described bellow:

1. Exact matches starting by the longest name towards the shortest;
2. Exact matches between $n_i \in \mathcal{N}^{initial}$ and the candidate of the news corpus;
3. Soft matches between $n_i \in \mathcal{N}^{initial}$ and the candidate of the news corpus, will allow us to include parts of names in common to both the $n_i \in \mathcal{N}^{initial}$ and the news corpus;
4. Avoid names with only one word because introduce noise to the system and are not frequent on news;
5. Avoid names with more than 4 words for the same reasons as for the previous rule.

By executing these rules on the news corpus, we mark the identified names with the tags "$< PN >$" and "$< /PN >$". We were able to automatically annotate the news corpus with 6,600 instances of names of well-known people, which corresponds to 562 different names.

3.3 Features Generation

One of the main principles of Named Entity Recognition is the choice of features [8]. Let us consider a set of features $\mathcal{F} = \{f_1, f_2, ..., f_n\}$, where f_i is a specific feature that will be applied to our news corpus $\mathcal{C}^{news}$. Our training corpus $\mathcal{C}^{train}$ can be described as:

$$\mathcal{C}^{train} = S(\mathcal{C}^{news}, f_i), f_i \in \mathcal{F} \quad (1)$$

where S is a function that maps the features with its corresponding words on the $\mathcal{C}^{news}$. We will use *word-level* features and a window of 3 tokens to the left and to the right of the focus word. Table 1 lists the features used to annotat $\mathcal{C}^{news}$.

For the first group of features from Table 1, "Capitalized Word", "Acronym" and "Word Length", we develop simple methods that fits these cases. Regarding the "End of sentence" features, we used a tokenizer specially trained for this type of texts [2]. This tokenizer, based on a text classification approach, tries to separate all the tokens from the input text and identifies the beginning and the end of the sentence. For the "Syntactic Category" and "Semantic Category" features, we used LSP (Léxico Semântico do Português), a lexicon developed for the Portuguese Language which is able to perform a syntactic (and for some words a semantic) analysis of the words of a given text. The last set of features, $\mathcal{F}_{names}$ is a list of names extracted from a Portuguese gazetteer, REPENTINO. REPENTINO is a gazetteer for the Portuguese language that stores nearly 100

Table 1. Set of features used for the annotation of $\mathcal{C}^{news}$

	Features	Examples
$\mathcal{F}_{cap}$	Capitalized word	*Pedro* or *Miguel*
$\mathcal{F}_{acr}$	Acronym	*NATO* or *USA*
$\mathcal{F}_{lng}$	Word Length	"musician" - *8*
$\mathcal{F}_{end}$	End of sentence	
$\mathcal{F}_{syn}$	Syntactic Cat.	"said" - *verb*
$\mathcal{F}_{sem}$	Semantic Cat.	"journalist" - *job*
$\mathcal{F}_{names}$	Names of people	*Paulo Portas*

categories and subcategories. For this work, we are only interested in names of people, which are identified by the category *HUM*, subcategory *EN_SER*. The task of extracting names from REPENTINO is thus straight forward and consists simply on building a list of all entities tagged on REPENTINO with the previous described category and subcategory.

3.4 CRF Model

Conditional Random Fields are undirected statistical graphic models, and McCallum et al. [4] have shown that are well suited to sequence analysis, particularly on named entity recognition for newswire data.

According to Lafferty et al. [3] and McCallum et al. [4] , let $o = \{o_1, o_2, ..., o_n\}$ be a sequence of words from a text with length s. Let $\mathcal{S}$ be a set of states in a finite state machine, each of which is associated with a label $l \in \mathcal{L}$ (e.g.: name, job, etc.). Let $s = \{s_1, s_2, ..., s_n\}$ be a sequence of states that corresponds to the labels assigned to words in the input sequence o. Linear chain CRFs define the conditional probability of a state sequence given an input sequence to be:

$$P(s|o) = \frac{1}{Z_o} exp\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_j f_j(s_{i-1}, s_i, o, i)\right) \quad (2)$$

where Z_o is a normalization factor of all state sequences, $f_j(s_{i-1}, s_i, o, i)$ is one of the m functions that describes a feature, and λ_j is a learned weight for each such feature function. For this work we only use binary feature functions, a first order Markov independence assumption. A feature function may be defined, for example, to have value 0 in most cases, and have value 1 if and only if state s_{i-1} is state #1 (this state may have, for example, label *verb*) and state s_i is state #2 (for example a state that have label *article*). Intuitively, the learned feature weight λ_j for each feature f_j should be positive for features that are correlated with the target label, negative for features that are anti-correlated with the label, and near zero for relatively uninformative features, as described by [13]. CRFs are described in more detail by [3].

We used CRF++ (version .054)[2], a customizable implementation of CRFs for segmentation/labeling of sequential data, and we set to 100 the maximum number of iterations of the algorithm. On one hand, the convergence becomes extremely slow for large sets of data, as the one we are using, and on the other hand 100 iterations is certainly enough for the algorithm to converge in our scenario. We also specify a template that will be used by the CRF++ algorithm to learn the model. This template allows to make different combinations of each token, its position and its features along with the other tokens from the sliding window. However, after several tests we conclude that the gains achieved by changing the template description were very low. Thus, we opt by using simple and straightforward templates that only describe each of the tokens, position and feature of the sliding window with no combinations with the other tokens. The CRF implementation used allow us to build a model based on our training corpus $\mathcal{C}^{train}$.

4 Experimental Set-Up

We are mainly interested in evaluating: (i) the quality of the annotation of the news corpus, our training set, (ii) and the quality of the CRF annotator obtained, which is specialized for identifying names of people. Concerning the annotation of the news corpus, a random subset of the annotated corpus will be manually evaluated. This subset will be composed by approximately 1% (200 news items) of our News Corpus. The evaluation of the CRF annotator will be performed using HAREM[3] [10], our *gold-standard corpus*, $\mathcal{C}^{gold}$. To better understand the contribution of the set of features $\mathcal{F}$, we will use different models $\mathcal{E}$, which are listed bellow.

- $\mathcal{E}^{baseline}$: $\mathcal{F}_{names} = \mathcal{N}^{initial}$
- $\mathcal{E}^{simple}$: $\mathcal{F}_{simple} = \mathcal{F}_{cap} \cup \mathcal{F}_{acr} \cup \mathcal{F}_{lng} \cup \mathcal{F}_{end}$
- $\mathcal{E}^{lsp}$: $\mathcal{F}_{lsp} = \mathcal{F}_{syn} \cup \mathcal{F}_{sem}$
- $\mathcal{E}^{comb\_1}$: $\mathcal{F}_{comb\_1} = \mathcal{F}_{names} \cup \mathcal{F}_{simple}$
- $\mathcal{E}^{comb\_2}$: $\mathcal{F}_{comb\_2} = \mathcal{F}_{names} \cup \mathcal{F}_{lsp}$
- $\mathcal{E}^{all}$: $\mathcal{F}_{all} = \mathcal{F}_{names} \cup \mathcal{F}_{simple} \cup \mathcal{F}_{lsp}$

For our baseline model, $\mathcal{E}^{baseline}$, the examples on the corpus will only be described by the names of people obtained from REPENTINO gazetteer. The simple model $\mathcal{E}^{baseline}$ will be used together with other models, since using structural like acronyms and capitalized words is not meaningless alone, when considering the NER task we propose. $\mathcal{E}^{baseline}$ model will describe examples from our corpus with syntactic and semantic information, which is important and will be tested alone. $\mathcal{E}^{comb\_1}$ is a model that combine both our baseline model and the simple model $\mathcal{E}^{simple}$, thus allowing us to understand the contribution of contextual information for the NER task. $\mathcal{E}^{comb\_2}$ is a combination of our

[2] Available at: http://crfpp.sourceforge.net/
[3] Available at http://www.linguateca.pt/HAREM/

baseline model and the $\mathcal{E}^{lsp}$, so we are able to describe examples from our news corpus with both names of people extracted from REPENTINO and syntactic and semantic information from LSP. Finally, $\mathcal{E}^{all}$ combines the baseline model with two other models, $\mathcal{E}^{simple}$ and $\mathcal{E}^{lsp}$.

To evaluate the quality of the annotator with the models previously described, we will use two different measures, precision $\mathcal{P}$ and recall $\mathcal{R}$. The precision $\mathcal{P}$ is given by:

$$\mathcal{P}_{annotator} = \frac{\#crt}{\#crt + \#inc} \tag{3}$$

where $\#crt$ represents the number of correct names identified and $\#inc$ the number of incorrect names. Regarding the recall $\mathcal{R}_{annotator}$, we have:

$$\mathcal{R}_{annotator} = \frac{\#crt}{\#crt + \#inc + \#nil} \tag{4}$$

where $\#nil$ represents the number misses (the annotator did not annotate a specific name).

5 Results

The evaluation of the automatic annotation of the training set, preformed manually on a random subset of 200 news, uses two different measures: (i) to describe the number of correct tags added to the training set, considering all the tags our proposed method was capable of identifying, we used $\mathcal{P}$, the precision, and achieved a result of 95%; (ii) and to represent the number of correct tags for all the labels the method was supposed to identify, including those that were missed, we used the recall, $\mathcal{R}$, and obtained a value of 74%. Also, the number of cases were the annotator misses names with a single word (e.g.: Obama or Jardim) represents 27% off all the cases (missing names).

Results regarding the performance obtained by the CRF annotator on identifying names of people are presented in Table 2

Table 2. Evaluation results for the CRF annotator

Model $\mathcal{E}$	Precision $\mathcal{P}$	Recall $\mathcal{R}$	F-measure $\mathcal{F}$
$\mathcal{E}^{baseline}$	55.3%	8.25%	14.3%
$\mathcal{E}^{lsp}$	82.4%	7.74%	14.2%
$\mathcal{E}^{comb\_1}$	80.2%	19.5%	31.4%
$\mathcal{E}^{comb\_2}$	83.1%	18.6%	30.4%
$\mathcal{E}^{all}$	78.7%	23.3%	**36.0%**

6 Analysis and Discussion

The evaluation results regarding our proposed method for automatically annotate, and thus build, a training set corpus, show that the precision achieved is very high (95%) for a relatively high value of recall, 74%. A high value for the precision indicates that the proposed method, whenever it founds a name, it almost always correctly identify its boundaries. For instance, the correct boundaries for *José Eduardo Ferreira Neto* are <PN>*José Eduardo Ferreira Neto*</PN>, and not *José* <*PN*>*Eduardo Ferreira*</*PN*>. Considering the recall, one can see that its value is significantly lower when compared with the precision. This indicates that the method under evaluation was not able to identify all the names. An error analysis was performed over the missed names, and we conclude that approximately 27% of all the missed names were names with only one word. This case fails in our training set annotation method since our initial assumption was defined to consider only names of people with at least two words, which seemed to be the most typical cases of names on journalistic text. However, one have to notice that considering names with only one word may introduce considerable *noisy* information to the system and degrade its performance.

In what concerns to the annotator performance (see Table 2), several considerations should be taken into account. The baseline system, $\mathcal{E}^{baseline}$, with features that are only based on names extracted from REPENTINO gazetteer, clearly achieved poor results, with a low precision (55%) and a low value for recall (8,3%). From these results one can see that the training corpus and the REPENTINO features are not enough to generate a NER system based on CRF and specialized on names of people. When considering $\mathcal{E}^{lsp}$, we discard information from the gazetteer, but instead use syntactic information as the syntactic category and the semantic category. Results show that the precision achieved (82%) clearly outperforms the baseline model, even though the recall value (7,7%) has dropped slightly. The considerable increase on the precision value shows that syntactic information is more relevant for NER tasks than the use of large gazetteers, which follows the conclusions of Mikheev et al. [5]. In $\mathcal{E}^{comb\_1}$ we combine features using the gazetteer with features obtained from simple structural analysis, and the results achieved show a significative improvement of the recall value (20%), with a F-measure of approximately twice the one obtained for the previous model, $\mathcal{E}^{lsp}$. These results show that combining features with names from a gazetteer and structural features are a good, yet simple approach for our specific task on NER. On $\mathcal{E}^{comb\_2}$ we combine features of names extracted from the gazetteer with features with syntactic information, and results show a slight improvement on the precision value (83%), even though the recall value drops approximately 1%. One can see that results achieved for $\mathcal{E}^{comb\_1}$ and $\mathcal{E}^{comb\_2}$ are similar and there are no relevant information that allow us to conclude with is the best option to choose. Finally, on $\mathcal{E}^{all}$ we combine features from names, structural features and syntactic features, which led to the best F-measure value (36%) of all methods.

Comparing results obtained using our automatic approach with the literature, Milidi et al. [6] achieved fairly better results with HMM (F-measure of

88%), however the manual annotation of the corpus, together with the small size of the corpus (2100 sentences) may be a decisive factor for the differences of F-measure obtained.

An additional error analysis was performed in order to figure possible causes and justifications for the errors obtained for the $\mathcal{E}^{all}$ model. Table 3 presents the distribution of the three major types of errors identified. The error type I is the

Table 3. Error analysis for the $\mathcal{E}^{all}$ model

Error Type	Cause	%
I	Incorrectly identified	36%
II	Name used in a different context	**33%**
III	Missed	31%

most frequent and include mainly names with wrong boundaries, but also some tokenization problems (originality from HAREM corpus) as for example "Paulo Pinto Mascarenhastem" instead of "Paulo Pinto Mascarenhas tem (has)". The error type II is quiet unexpected. It represent cases where our system is able to correctly identify names of people, however these names are used in a different context as a street name or a public institution as for example "Maternity Alfredo da Costa", where "Alfredo da Costa" is also a name of a person. The error type III represent the cases where our system misses the identification of names of people.

As a final consideration, it is important to notice that, even though the proposed methods, both automatic annotation of the training corpus and the NER system for names of people, are tested on Portuguese text, they could be applied to other languages with minor changes, namely using a POS-tagger adapted for the new language.

7 Conclusions and Future Work

We proposed two different methods to automatically identifiy names of people on journalistic text. First we automatically annotate a corpus with 20,000 news with a set of names extracted from Voxx. This corpus was used as our training corpus. Second, we describe examples from our news corpus with a set of specialized features. We used the annotated corpus, our training set, together with the examples described by set of features, to learn a Conditional Random Field model. This model allowed us to create a specialized system capable of identifying names of people on news.

The evaluation was performed on both proposed methods. The automatic annotation of the news corpus - the training set - achieved precision of 95% and recall of 74%. In what concerns to the NER system for names of people,

we obtained precision of 79% and recall of 23% for the model that combines structural and syntactic information and the gazetteer of names extracted from REPENTINO. When comparing the results obtained with similar work from other authors, as Milidi et al. [6], one can see that our results are not as good as Milidi ones. However, one must notice the main purpose of this work was to build a NER system for names of people with no human intervention, where all processes from the creation of the training corpus to the definition of features is performed automatically. For this reason, direct comparison with other similar work but performed (partially) manually may not be meaningful.

As future work, and regarding the method for automatically create a training set for CRF, and considering the case of using names with only one word, additional work has to be performed in order to evaluate the trade-off between precision and recall in order to achieve the best results possible. For the NER system itselft, we could expand the proposed method to include other Named Entity categories, as companies names, locations and jobs. In order to achieve these objective, wikipedia is a interesting resource to explore since most articles are about Named Entities (Jun'ichi et al. [1]). Another interesting topic for future work is to perform active learning. The system could improved its model by using recently identified Named Entities, thus enriching its overall performance.

Acknowledgments

This work was partially supported by Labs Sapo UP from Portugal Telecom.

References

1. Jun'ichi, K., Torisawa, K.: Exploiting Wikipedia as external knowledge for named entity recognition. Proc. EMNLP-CoNLL (March), 698–707 (2007)
2. Laboreiro, G., Sarmento, L., Teixeira, J., Oliveira, E.: Tokenizing Micro-Blogging Messages using a Text Classification Approach. AND'2010 - ACM pp. 81–87 (2010)
3. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Machine Learning - International Workshop. pp. 282–289. Citeseer (2001)
4. McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. pp. 188–191. Association for Computational Linguistics (2003)
5. Mikheev, A., Moens, M., Grover, C.: Named Entity Recognition without Gazetteers. In: Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics. pp. 1–8. Association for Computational Linguistics (1999)
6. Milidiú, R.L., Duarte, J.C., Cavalcante, R.: Machine Learning Algorithms for Portuguese Named Entity Recognition. Inteligencia Artificial 11(36), 67–75 (Dec 2007)
7. Minkov, E., Wang, R., Cohen, W.: Extracting personal names from email: Applying named entity recognition to informal text. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language

Processing. pp. 443–450. No. October, Association for Computational Linguistics (2005)
8. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes (1991), 1–20 (2007)
9. Nadeau, D., Turney, P., Matwin, S.: Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. Advances in Artificial Intelligence pp. 266–277 (2006)
10. Santos, D., Seco, N., Cardoso, N.: HAREM: An advanced NER evaluation contest for portuguese. Resources and Evaluation, (2006)
11. Sarmento, L.: SIEMÊS - A Named-Entity Recognizer for Portuguese Relying on Similarity Rules (2006)
12. Sarmento, L., Pinto, A., Cabral, L.: REPENTINO A Wide-Scope Gazetteer for Entity Recognition in Portuguese. Computational Processing of the Portuguese Language pp. 31–40 (2006)
13. Settles, B.: Biomedical named entity recognition using conditional random fields and rich feature sets. Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications - JNLPBA '04 p. 104 (2004)
14. Talukdar, P.P., Brants, T., Liberman, M., Pereira, F.: A context pattern induction method for named entity extraction. In: Proceedings of the Tenth Conference on Computational Natural Language Learning. pp. 141–148. CoNLL-X '06, Association for Computational Linguistics, Morristown, NJ, USA (2006)

Web sessions clustering for behavioral targeting

Pedro Saleiro

Faculdade de Engenharia da Universidade do Porto

Abstract. In this paper we present our on going effort to compare web sessions clusters based on different web sessions representations. Sessions within the same cluster represent common navigation patterns. We assume that users with the same navigation patterns have common interests and motivations at a point in time. Therefore, we represent sessions based on descriptions extracted from the URLs as well as, using temporal references. The session clusters were obtained using the K-Means algorithm. The results show that is possible to find groups of sessions containing similar pages visited. It is also possible to extract intelligible descriptions from the sessions clusters. However, in the future, it is expected to evaluate the results obtained using an internal criteria.

Keywords: Behavioral targeting, web usage mining, clustering

1 Introduction

In the $11.2 billion [1] online display advertising market, the effectiveness of web ads is a major requirement. Ad targeting aims to raise ad's effectiveness by attempting to identify the right ad to deliver to the right consumer at the right time.

There are several approaches to ad targeting [5,23] such as geographical, demographic, contextual or behavioral. These approaches differ in the way how the consumer's information is collected. The first two, geographical and demographic collect consumer's information in explicit way. They are user centric and require access to a profile explicitly created by each consumer. However, just large web sites containing data sets of registered users such as Facebook, tend to be the best to apply these approaches.

On the other hand, both contextual and behavioral targeting approaches extract consumer's information in an implicit way, which means, without explicitly ask that. Contextual targeting is based on the web page content such as in Google AdSense. The downside of this approach is that it is only based on the momentary interest of the user. In this type of ad targeting the challenge relies on having the best algorithm to determine what the page is about.

Behavioral targeting uses the user navigational patterns in order to understand user preferences and interests which mean that it is both user centric and content centric. This results in more personalized advertisement which in turn can fetch more consumer interest [23].

During this work it is not possible to explicitly access user's information towards raising the effectiveness of ads offered by an ad server. Therefore we

propose to use a behavioral targeting approach, i.e, segment users based on their behavior. The main goal of this work is to verify the existence of user's groups (segments) with similar behaviors and additionally try to obtain intelligible descriptions of these groups.

In order to create the user's groups, i.e., user profiles, we only have access to the user's browsing history. We assume that users with similar browsing patterns upon a point in time should have similar interests and motivations. User profiles can be created through offline web usage mining, which consists in discovering web usage patterns to better understand the users' behavior.

For the task of creating user groups based on the web server access logs, web usage mining uses clustering techniques. During a visit to a web site, the users' requests are registered in a web server access log format stored in the web server. Therefore, the web server access logs provide the means to create a data set prepared for the application of clustering algorithms.

User's interests are affected by the temporal context, thus in some research work [14] instead of creating user clusters it is presented the concept of session clustering. A session comprises the browsing history of small time windows, usually 30 minutes. Therefore by clustering sessions it is easier to comprehend the contextual motivations of each user and provide ads suitable for current user's interests.

We propose to create user's profiles in a two stage process. First, one creates session clusters, then creates users clusters based on common sessions between users. This work is going to focus on the first part, create session clusters. We compare the resulting session clusters by using different attributes to describe a session. Our approach for representing a session consists in combining descriptions extracted from the URLs of the pages visited with temporal frames based on date, such as "Monday morning".

2 Related Work

Clustering web users or user sessions has been subject of research for the last decade and the most common approaches for representation of web usage are in a vector model [10,14,15] such as usage-based for access/not access web page or frequency-based, containing the number of accesses to a page.

We may refer the work of Yan et al [24] which employed the First Leader clustering algorithm with frequency-based representation. This algorithm is computationally fast due to reading only one time each object to be clustered however, the order of the objects influences the composition of the groups. Furthermore, it is necessary to introduce the number of expected clusters.

Fu et al. [8] applied the BIRCH [25] algorithm in a space of generalized sessions. Despite their method scaling well, it needs the specification of a similarity threshold and it is also dependent on the order of the objects.

The work of Wang et al. [20] presents a new way of measuring similarity between web sessions. The authors consider each session as a sequence and in order to measure the similarity between two pages they use the idea of sequence

alignment from bioinformatics. Each URL is divided in tokens and then a sequence is determined. The best matching between two sessions is found through dynamic programming. This method is targeted to clustering sessions that contain visits to different pages inside the same web site or domain. However in our case, the sequence of pages visited within a session is not relevant and different pages served by the ad server can represent a common topic of interest to the user.

Time aware web users or session clustering has been taken into consideration by mainly considering the time spent in a given page and the succession of visits, the "clickstreams" [21,3]. Nevertheless, the time spent in a web page is prone to noise, as many external reasons may lead to spend more or less time in a web page. Another drawback of these kind of approaches is considering the succession of visits to a web page in an overall time span instead of in simultaneous intervals. A simplified approach from Lingras et al. [11] uses boolean values to represent day or night visit but it introduces reduced accuracy of users preferences over time.

3 Method Description

In this section, we present our approach to create clusters of sessions which represent groups of users with similar browsing patterns. Furthermore, we present how we expect to extract intelligible descriptions of these groups. In order to obtain such descriptions we suppose that users with similar browsing patterns should have similar interests.

First we describe how to identify a session from the web server access logs followed by an explanation of our approach for representing a session. At last we present the clustering techniques suitable for grouping similar sessions.

3.1 Session Identification

A session is a list of web pages accesses from a given user during a period of time. Each access is registered in a line of the web server access log. For the task of identifying the list of web pages visited during a user's session it is necessary to clean all the information contained in the web server access logs that is meaningless or not relevant. Though, browser and proxy caching represent a major drawback to the creation of a reliable user session data set [7].

The web server access log is a text file that contains all the requests made to the web server, and usually they are in a Common Log Format [12], which means that it contains the following fields:

- IP address or domain name
- User ID
- Date and time of the request
- HTTP request (including method and page requested)
- Status code response to the request

- File size
- Referrer (web page that contain the hyperlink that originated the request)
- Web agent (user's browser)

The web server access logs used during this work contain accesses to web pages from several web sites and in this case, the URL of the web pages is in the referrer. There is also extra information about the request such as a session cookie and a long duration cookie. The session cookie identifies a 30 minutes session and the long duration cookie identifies a user. Therefore, only web server access log entries containing the session cookie were considered. From these entries the web page URL (referrer), date and session cookie are the meaningful data for the purpose of this work. Thereafter, these parameters were grouped by common session cookie in order to create each session representation vector.

3.2 Session Representation

As stated in the Introduction section, it is our goal to compare the impact of using different attributes to describe a session in the resulting session clusters. From the web server access log entries selected, different attributes can be defined based on date and web page URL.

Pallis [17] only extracted one attribute of each page visited, which was the URL of the web page as a whole. In our case, different URLs could represent pages from the same subject because the web server access logs used contain entries of web pages from several different web sites.

Thus, we propose to extract description tags $D = \{d_1, d_2, ..., d_n\}$ from the URLs. Two different verbosity levels, α and β are defined. In the level α, it is assigned the tag d_1^α to the URL domain and the tag d_2^α to the second token of the URL. In the level β of tag verbosity, it is assigned a tag d_i^β for each "word" in the URL. Thus we can define the description tags:

- URL verbosity α: http://d_1^α/d_2^α/...
- URL verbosity β: http://d_1^β.d_2^β.d_3^β/d_4^β/.../d_n^β

Thus, for the URL http://praias.sapo.pt/alentejo/sesimbra/meco we can extract the following description tags:

- URL verbosity α: *praias.sapo.pt*, *alentejo*
- URL verbosity β: *praias*, *sapo*, *pt*, *alentejo*, *sesimbra*, *meco*

For both levels of verbosity we create a unique list of tags based on the URLs extracted from the web server access logs. The total number of tags is based on the number of unique tags extracted from the group of URLs.

We consider that user's interests may be temporal. Therefore, we define temporal tags $T = \{t_1, t_2, ..., t_n\}$ in order to group web pages accessed in simultaneous temporal frames. Each attribute t is defined by a description tag $d_i^\alpha$and

a time frame. We defined two levels of temporal tags granularity, π and μ. In the π level, the temporal tags t_i^π correspond to the 7 week days. While in the μ level a division of days in "M-morning" (8h00-15h59), "AE-afternoon/evening" (16h-23h59) and "N-night" (00h00-7h59) is also considered in the temporal tags t_i^μ.

Thus, for the URL http://praias.sapo.pt/alentejo/sesimbra/meco visited in October 15 at 12:32:22, we can extract the following temporal tags:

- Temporal granularity π: $praias.sapo.pt - friday$, $alentejo - friday$
- Temporal granularity μ : $praias.sapo.pt - fridayM$, $alentejo - fridayM$

Let A be a set of attributes describing the sessions extracted from the web server access logs, $A = \{a_1, a_2, ..., a_n\}$, each of which is a tag d_i^α,d_i^β,t_i^π or t_i^μ.

In order to facilitate the clustering operation, each session is represented as an n-dimensional vector over the space of attributes. Let S be a set of user's sessions. Thus, $S = \{w(a_1, s), w(a_2, s), ..., w(a_n, s)\}$, where each $w(a_i, s)$ is a weight assigned to the ith attribute extracted in a session. The weight is boolean, i.e, if an attribute is extracted from a URL of a page visited within a session then it is assigned the value 1 in the session sector. If it is not extracted then it has the value 0 in the session vector. We organize this vectors in a matrix, where the rows are sessions and attributes are columns and the elements are the frequency weights $w(a_i, s)$. During this work, different types of session representation are going to be compared:

- $S = \{w(d_1^\alpha, s), w(d_2^\alpha, s), ..., w(d_n^\alpha, s)\}$
- $S = \{w(d_1^\beta, s), w(d_2^\beta, s), ..., w(d_n^\beta, s)\}$
- $S = \{w(t_1^\pi, s), w(t_2^\pi, s), ..., w(t_n^\pi, s)\}$
- $S = \{w(t_1^\mu, s), w(t_2^\mu, s), ..., w(t_n^\mu, s)\}$

Each representation depends of the type of tags (d_i^α,d_i^β,t_i^π or t_i^μ) representing the attributes. Hence, it is possible to evaluate the impact on the resulting clusters of using different description tags granularity, as well as, using temporal tags for sessions representation.

3.3 Session Clustering

Thereafter the transformation of user sessions into a multi-dimensional space as vectors of extracted attributes, clustering algorithms can partition this space into group of sessions. Each session within a group has a close distance between the others in the group, based on a distance measure.

Regarding the clustering algorithms both model-based and similarity-based are used to group users or sessions, as well as, hierarchical and partitional [4,9] techniques. The most common model-based algorithm is the Expectation-maximization (EM) algorithm which has been used to identify associations among users and pages [6,16] as well to provide user profiles [22]. K-means algorithm

[19] is the most common similarity-based algorithm. It has been used with different distance measures such as the squared Euclidean, the cosine and Manhattan distances [4,26,18].

The K-means [13] algorithm is a non-hierarchical clustering algorithm that assumes instances as real-valued vectors. Each cluster is based on a centroid, or mean of points in a cluster c. The objective of the algorithm is to minimize the squared distance $d(x, y) = \frac{1}{2}\sum(x_i - y_i)^2$ of every point to its associated cluster centroid. To start it is necessary to select a random number k of instances as seeds. K-means is then an iterative two-step algorithm where on assignment step, each data point n is assigned to the cluster which distance is minimal. Thereafter, in the update step the seeds are adjusted to the centroid of each cluster.

4 Method Evaluation

The session clusters should capture the overlapping interests and motivations of different users up to a point in time. However, evaluating clusters is a complex task, even more when there are no gold-standards clusters available, as in this case.

The quality of clusters can be determined based on internal criteria. For instance, measure if the inter-cluster similarity is low and if the intra-cluster similarity is high. However, the internal criteria might not imply clusters representing common interests between sessions within that group.

Thus, evaluation comprises both validation and interpretation of the resulting clusters, such as in Pallis et al. [17]. Validating clusters consists in assessing the quality of the clusters while interpretation is about determining if it is possible to extract intelligible descriptions of the clusters.

On the other hand cluster results interpretation is not trivial as it depends on the nature and orientation of the underlying application. Baldi[2] has focused on interpretation using visualization approaches. For instance, Cadez [6] uses a simple visualization scheme of the web users patterns and make some empirical observations.

5 Experimental Set-up

The method described has been applied on a real data set containing about 30 minutes of web server access log entries from a portuguese ad server on October 15, 2010. As we only had access to a reduce time period of entries, the temporal frame of every session is the same. Therefore, it makes no sense extract temporal tags from the web server access logs. The statistics of the experimental data set are depicted in Table 1. The data set contains 43273 entries, which corresponds to 30.6 MB web server access log file. Therefore, the experiments were conducted using the K-means algorithm as it is not necessary to use a stream clustering algorithm. The web server access log entries contained 11958 unique visited pages

and 26919 hits contained the web session cookie value, which is mandatory for an accurate session identification.

Thereafter the data pre-processing, were identified 11008 unique sessions, 5 unique description tags using verbosity α. In order to facilitate the clustering operation we have filtered attributes which were present in just one session or in more than 50% of the sessions. These attributes are not relevant as it is not possible to cluster sessions if there is no common visited pages. The same happens with a page that was visited in the majority of the sessions as it does not representative to a particular group of sessions. Thereafter removing the filtered attributes it was necessary to remove the sessions that didn't contain any attribute of the new attributes space.The dimensionality reduction operations resulted in 171 description tags and 6123 sessions.

Table 1. Statistics of experimental data set.

Attributes	Testing set
Total access entries	43273
Accessed web pages	11958
Identified sessions	11008
Description tags	553
Description tags (<1 session)	379
Description tags (>50% sessions)	3
Filtered description tags	171
Filtered sessions	6123

6 Experimental Results and Analysis

Following the methodology the sessions were clustered using the K-means algorithm on the Weka toolkit. Different numbers of clusters were tested. As described in the section 3.3, the K-means require the introduction of the number of expected clusters. We have tested with random values between K=10 and K=100, however with K>20, the majority of the clusters contained less than 20 sessions (0.32% of testing set). We consider that the navigation patterns represented by these clusters are not representative, therefore we present the results with K=20.

For the 20 clusters we extracted the description tags with highest mean value of each centroid, as a representation of the interest topics of each cluster. Table 2 gives a specific view of each of the 20 clusters obtained, including the number of sessions assigned to each cluster and the most relevant description tags.

Table 2. Clusters description (K=20).

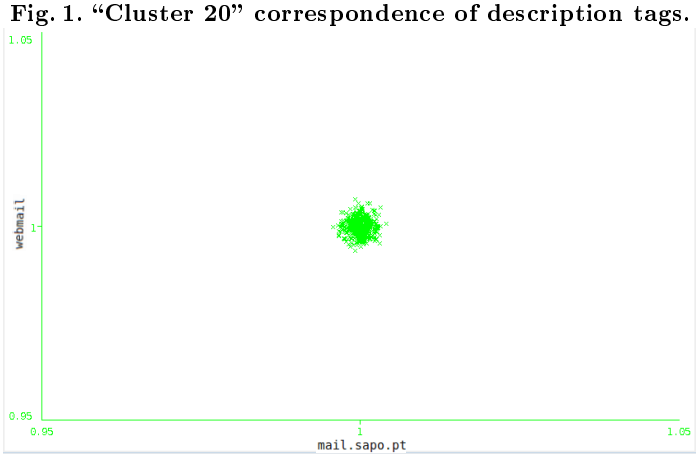
Cluster label	Proportion of sessions set (%)	Number of description tags	Description tags
1	46%	112	astral.sapo.pt, previsoes, tempo.sapo.pt
2	4%	29	noticias, economico.sapo.pt, tek.sapo.pt
3	1%	2	miniclip.sapo.pt, games
4	2%	16	mail.sapo.pt, webmail, services
5	3%	12	noticias.sapo.pt, desporto, economia
6	3%	17	jn.sapo.pt, paginainicial, Dossies
7	3%	21	sabores.sapo.pt, receita, pesquisa
8	2%	12	desporto.sapo.pt, mais_modalidades, sporting
9	3%	15	web.mail.sapo.pt, services, home
10	9%	19	dn.sapo.pt, inicio, sol.sapo.pt
11	1%	6	casa.sapo.pt, Apartamentos, Scripts
12	1%	14	mulher.sapo.pt, actualiade, lazer
13	1%	13	emprego.sapo.pt, resources, h.s.sl.pt
14	2%	12	casa.sapo.pt, Apartamentos, Scripts
15	1%	10	tsf.sapo.pt, common, PaginaInicial
16	1%	9	casa.sapo.pt, WebControls, Arrendamentos
17	1%	14	noticias.sapo.pt, banca
18	3%	23	desporto.sapo.pt, futebol
19	2%	11	auto.sapo.pt, Iframe, carros
20	10%	8	mail.sapo.pt, webmail

It is seen in the table 2 that the "cluster 1" accounts for the largest proportion of the sessions data set, 46%, and contains the most of description tags, 112, among all 20 clusters. These indicate that "cluster 1" stands for the most frequent navigation pattern of the sessions data set. However, the mean values of each description tag are quite low and similar between the 112 present on "cluster 1". This represent that there is not an uniform navigational pattern between the majority of the sessions in the cluster. Thus, the "cluster 1" is not a representative cluster of a group with similar navigational patterns.

As it can be observed on Figure 1, the cluster points are uniform for "mail.sapo.pt" and "webmail" descriptions tags. Every sessions contain a visit to the e-mail page which represents a typical e-mail consulting session. The same happens with "cluster 3" with online arcade gaming. The "cluster 18" with the description tags "desporto.sapo.pt" and "futebol" represents a session about football news.

Despite, we obtained some representative clusters with K=20, it would be valuable to conduct an analysis with increasing numbers of K and the proportion of each cluster as well as with its mean values of intra-cluster and inter-cluster distance.

Fig. 1. "Cluster 20" correspondence of description tags.



7 Conclusion and Future Work

In this paper, we presented our on going work to use a behavioral targeting approach in order to raise the effectiveness of ads offered by an ad server. The web server access logs are the only implicit information about users trends and behaviors that we have access. Therefore, we presented our ongoing effort to create web sessions clusters using web server access logs containing users' browsing history. We assume that sessions with similar browsing patterns should represent users with similar interests.

We proposed to create sessions clusters using different sessions representations. Our approach for representing a session consists in combining different levels of verbosity when extracting features from URLs and temporal descriptions from the time frame of each visit to a web page.

We had access to a 30 minutes period of web server access log entries. Web session clusters were created using K-means algorithm with K=20. It was possible to extract intelligible descriptions from the majority of the clusters, however due to the large number of features different partitional and streaming clusters algorithms shall be tested.

In the future, it is necessary to evaluate the resulting clusters using an internal criteria. Experiments will be conducted using 1 month of web server access log entries in order to apply the both temporal tags representation described in section 3.2. By using temporal tags, the number of features will increase significantly, thus further dimensionality reduction techniques must be applied.

Acknowledgments. This work was partially supported by Luis Sarmento.

References

1. Goldfarb A. and C. Tucker. Online display advertising: Targeting and obtrusiveness. *Marketing Science*.
2. Pierre Baldi, Paolo Frasconi, and Padhraic Smyth. Modeling the internet and the web. 2003.
3. Arindam Banerjee and Joydeep Ghosh. Clickstream clustering using weighted longest common subsequences. 2001.
4. A. Bianco, G. Mardente, M. Mellia, M. Munafo, and L. Muscariello. Web user session characterization via clustering techniques. In *Global Telecommunications Conference, 2005. GLOBECOM '05. IEEE*, 2005.
5. Andrei Broder, Marcus Fontoura, Vanja Josifovski, and Lance Riedel. A semantic approach to contextual advertising. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 559–566, New York, NY, USA, 2007. ACM.
6. Igor Cadez, David Heckerman, Christopher Meek, Padhraic Smyth, and Steven White. Model-based clustering and visualization of navigation patterns on a web site. *Data Mining and Knowledge Discovery*, 7:399–424, 2003.
7. Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Data Preparation for Mining World Wide Web Browsing Patterns. *KNOWLEDGE AND INFORMATION SYSTEMS*, 1:5–32, 1999.
8. K. Sandhu Fu and M.-Y. Shih. Clustering of web users based on access patterns. *WEBKDD workshop*, 1999.
9. David J. Hand, Heikki Mannila, and Padhraic Smyth. *Principles of Data Mining (Adaptive Computation and Machine Learning)*. The MIT Press, August 2001.
10. R. J. Kuo, J. L. Liao, and C. Tu. Integration of art2 neural network and genetic k-means algorithm for analyzing web browsing paths in electronic commerce. *Decis. Support Syst.*, 40:355–374, August 2005.
11. Pawan Lingras and Chad West. Interval set clustering of web users with rough k-means. *Journal of Intelligent Information Systems*, 23:5–16, 2004.
12. Berners-Lee T. Luotonen, A. Cern httpd users guide reference manual. 1994.
13. D.J.C. Mackay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, October 2003.
14. Sanjay Madria, Sourav Bhowmick, W. Ng, and E. Lim. Research issues in web data mining. In Mukesh Mohania and A Tjoa, editors, *DataWarehousing and Knowledge Discovery*, volume 1676 of *Lecture Notes in Computer Science*, pages 805–805. Springer Berlin / Heidelberg, 1999.
15. José D. Martín-Guerrero, Alberto Palomares, Emili Balaguer-Ballester, Emilio Soria-Olivas, Juan Gómez-Sanchis, and Antonio Soriano-Asensi. Studying the feasibility of a recommender in a citizen web portal based on user modeling and clustering algorithms. *Expert Syst. Appl.*, 30:299–312, February 2006.
16. George Pallis, Lefteris Angelis, and Athena Vakali. Model-based cluster analysis for web users sessions. In Mohand-Said Hacid, Neil V. Murray, Zbigniew W. Ras, and Shusaku Tsumoto, editors, *Foundations of Intelligent Systems*, volume 3488 of *Lecture Notes in Computer Science*, pages 219–227. Springer Berlin / Heidelberg, 2005.
17. George Pallis, Lefteris Angelis, and Athena Vakali. Validation and interpretation of web users' sessions clusters. *Inf. Process. Manage.*, 43:1348–1367, September 2007.

18. Sophia Petridou, Vassiliki Koutsonikola, Athena Vakali, and Georgios Papadimitriou. A divergence-oriented approach for web users clustering. In Marina Gavrilova, Osvaldo Gervasi, Vipin Kumar, C. Tan, David Taniar, Antonio LaganÃ , Youngsong Mun, and Hyunseung Choo, editors, *Computational Science and Its Applications - ICCSA 2006*, volume 3981 of *Lecture Notes in Computer Science*, pages 1229–1238. Springer Berlin / Heidelberg, 2006.
19. R. Tibshirani T. Hastie and J. Friedman. Elements of statistical learning: Data mining, inference, and prediction. *The Mathematical Intelligencer*, 27(2).
20. Weinan Wang and O.R. Zaiane. Clustering web sessions by sequence alignment. In *Database and Expert Systems Applications, 2002. Proceedings. 13th International Workshop on*, pages 394 – 398, 2002.
21. Jitian Xiao and Yanchun Zhang. Clustering of web users using session-based similarity measures. In *Computer Networks and Mobile Computing, 2001. Proceedings. 2001 International Conference on*, 2001.
22. Guandong Xu, Yanchun Zhang, Jiangang Ma, and Xiaofang Zhou. Discovering user access pattern based on probabilistic latent factor model. In *Proceedings of the 16th Australasian database conference - Volume 39*, ADC '05, pages 27–35, Darlinghurst, Australia, Australia, 2005. Australian Computer Society, Inc.
23. Jun Yan, Ning Liu, Gang Wang, Wen Zhang, Yun Jiang, and Zheng Chen. How much can behavioral targeting help online advertising? In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 261–270, New York, NY, USA, 2009. ACM.
24. Tak Woon Yan, Matthew Jacobsen, Hector Garcia-Molina, and Umeshwar Dayal. From user access patterns to dynamic hypertext linking. *Computer Networks and ISDN Systems*, 28(7-11):1007 – 1014, 1996. Proceedings of the Fifth International World Wide Web Conference 6-10 May 1996.
25. Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch: an efficient data clustering method for very large databases. In *Proceedings of the 1996 ACM SIGMOD international conference on Management of data*, SIGMOD '96, pages 103–114, New York, NY, USA, 1996. ACM.
26. Ying Zhao and George Karypis. Criterion functions for document clustering: Experiments and analysis. 2001.

SESSION 2

ARTIFICIAL INTELLIGENCE

Chairman: Rui Jorge Martins da Silva Chilro

Lobinho Gomes
Control of machining cutting force Using Artificial neural networks

Mario Gonzalez, David Dominguez and Angel Sanchez
Learning Vehicle Traffic Videos using Small-World Attractor Neural Networks

Zafeiris Kokkinogenis, Lúcio Sanchez Passos, Rosaldo Rossetti and Joaquim Gabriel
Towards the next-generation traffic simulation tools: a first evaluation

Control of machining cutting force Using Artificial neural networks

Lobinho Gomes

Faculty of Engineering - University of Porto (FEUP)
Rua Dr. Roberto Frias, 4200-465 - Porto - Portugal
Lobinho.gomes@gmail.com

Abstract. The constant search of industry for productivity raises and market shares, pushes the development of new products capable of giving an answer to these concerns. Specifically the machine tools makers have tried to solve these problems incrementing the capability of the machines they produce, essentially in speed and precision. The recent study of some problems associated to the machining process, has revealed the possibility of incrementing the productivity of some of vertical milling machine, only through the force control, keeping it constant and equal to the optimum value defined for the tool. The cutting force control, due to the system characteristics, can only be implemented by making use of adaptive control. In order to implement adaptive controllers we have at our disposal two technologies that have been showing good results. These technologies are Neural Networks and Fuzzy Logic.
We thought that it would be of interest to research the use of Artificial Neural Networks in implementation of a controller. This has been the objective of the development of the work described in this paper.
The results obtained have been encouraging, showing the possibility of Implementing those controllers in real systems.

Keywords: Artificial Neural Networks, Cutting force, *Feed-Forward, Recurrent, Backpropagation,* Time Delay Neural Network, Dynamic Recurrent Neural Networks.

1 Introduction

The Industry linked to cutting processes fight currently with two limitations that decrease their production rate. The feed rate, and spindle speed are always programmed in offline and in a conservative way by programmers.

A system allowing the adjustment of cutting force to real conditions of operation of the machine, will certainly allow a significant improvement of the cutting process. For each material, there is a specific cutting force [1], also called cutting pressure by some authors [2].

A specific force of cutting for each material is defined as the force required to cut 1 mm^2 cross-section of the material [1]. In reality, the cutting process of a vertical milling machine has several different characteristics, depending on the material, tool and tool wear, and mainly to piece geometry being machined. The geometry of the work piece to be machined is the main responsible for control difficulties, because when a milling machine is manufactured it must be able to work a very high range of pieces, with geometries completely unpredictable.

Changes of cutting force, suffered during the machining process, remove the tool of the optimal functioning point, implying increased costs of production.

The increased costs of production is mainly due to rapid wear of the cutting tool, to bad finishing surfaces, by poor use of the tool due to be working below their potential, and high machining time.

From the point of view of theory of control, the cutting process can be considered complex, with a reduced knowledge about cutting process and the possibilities of its control. Up to now has not yet found an acceptable solution to deal with such processes [10]. Our problem is to develop an Adaptive controller capable of controlling the cutting force of a machining operation.

An Adaptive controller is one who can modify their behavior in response to the changes of the dynamics of the process and the nature of the disturbance [6]. We can describe the difficulties by dividing them into three important classes, cutting technology used in the machining process, the problem of control, and the technology used to implement the controller.

The first challenge was to realize what is the problem of control, what kind or kinds of control would be liable to be used, and what is the best solution to use here, as the controller of the plant (machining process) is completely unknown and puts great difficulties to thorough knowledge. Another problem related to the control, is the technology that should be used to implement the chosen control. Artificial Neural networks and Fuzzy Logic are currently the most promising opportunities [8], to implement an adaptive control of cutting force. Scientific works [5] demonstrate how these techniques can be applied to a control so successfully. We decided that we would investigate the use of Artificial Neural Networks in the implementation of a cutting force controller. In particular the Recurrent Neural Networks [12], a network topology where exists "feed-back" [17].

Noted that whatever the type of control to be implemented, would be an adaptive control for nonlinear systems. The division between linear and nonlinear systems, suggests a classification for nonlinear systems that do not exist because the differential equations of a nonlinear system are virtually devoid of a general method of resolution, and an exact solution can only exist for a particular group that share some properties [4]. The main objectives of this work, is to demonstrate the importance of control of cutting force, throw some ideas that can be used in future in controller's development to control the cutting force.

So, we will aim to define the overall structure of the controller and its substructure specifications, creating models and simulating their own functioning in programs such as MATLAB. Through this simulation by using actual data whenever possible, define

the structure and training of different Artificial Neural Networks that implement the control system.

This paper initially describes adaptive control operation and its application to control of cutting force. In the following section it is presented and discussed the results, ending with a conclusion.

2 Adaptive Controller

The adaptive control means that the controller, during its operation, will be modified to adapt to new control situations.

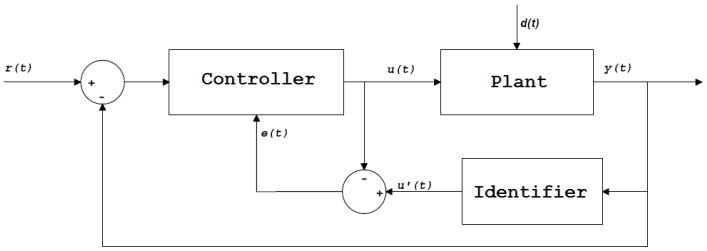


Fig. 1. Generic Adaptive Controller

Fig. 1 represents an adaptive control system with ability to self learning. A reference signal is applied to the controller, which excites the plant, applying its entry a control signal. The plant reacts to input signal, producing an output, which is applied to process model (plant model) that we want control.

The process model generates an error that identifies how the controller output moves away from the desired value. This error signal is used to modify the controller's parameters to adapt to the new situation. So the process model must be able to, for a certain value of the plant's output, generate its input value, i.e. the model of the process must identify, for the plant, the value present at entry.

The process model, also called Identifier, can be analytically implemented if we know the transfer function of the plant. If we can designate f as transfer function of the plant, and if f is known and is possible to invert, then the function f^{-1} describes the Identifier. For complex processes, the f function is not known, nor possible of invert, so it is necessary to use other techniques, such as Artificial Neural Networks, to implement the Identifier.

2.1 Control of cutting force

To define this kind of control we use the process models, the control models and their adaptation by learning.

The problem of controlling the cutting force it is put at two distinct hierarchical levels **Fig. 2**. At a higher level, which will designate of *Decider*, we have the problem

of defining what the cutting force, which best optimizes the cutting process. At a lower level, which we call *Actuator*, we need a controller, which keep constant the cutting force.

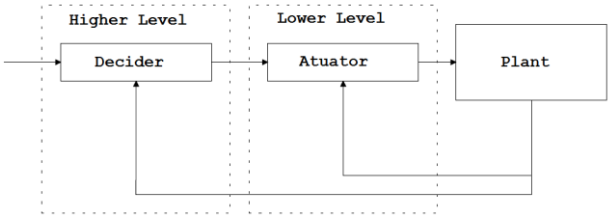


Fig. 2. Controller Model of cutting force

2.1.1 Decider

This block was designated Decider, because its main function is to decide what is the optimum cutting force, the cutting force that minimizes machining costs. However, there are features of controller, to the extent that if there is change on some variables that it depends, it can adjust the cutting force according to this variation.

However this block will be left for future study.

2.1.2 Actuator

The Actuator is constituted by the controller; the identifier and a mid-block between the controller and the plant that we call Adapter, and its function is transform the signal received from the controller, on two other signs, used by the machine.

The machine uses two parameters that together define the machining conditions, i.e. forward speed (*F*) and spindle speed (*S*) of the tool. In reality the controller generates the signal *F/S*, and the Adapter should unfold this sign in signs *F* and *S* respectively. The structure of the Actuator can be observed in **Fig. 3**.

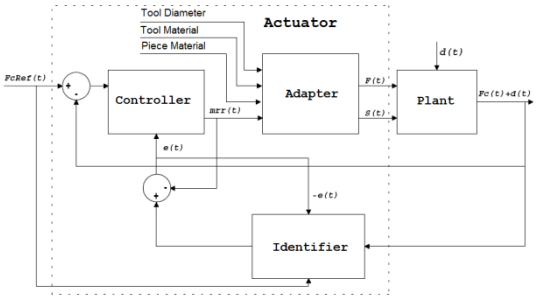


Fig. 3. Structure of low-level controller (Actuator)

The adapter is required because for the same reason *F/S*, *F* and *S* signals may be different, depending on the material and the tool used. When the reason *F/S* switches, even without modification of the material, the increase or decrease of *F* and *S*, may not be linear. The value *F/S* is also called *mrr*-material removes rate, and defines the cutting force. The Adapter to be able to generate signals *F* and *S* needs too information about the tool and the material to be machined.
The operating principle of the Actuator can be described as follows: the reference signal *(FcRef)* is applied to the controller, the controller computes a signal *mrr (t)* so that the cutting force exerted by the machine, is equal to the force of reference. The Adapter receives the signal *mrr (t)*, and from the knowledge of the process, determines the signs *F* and *S*, to be used by the machine, the signal applied to the machine causes a particular cutting force, which may be equal, or not, the reference cutting force. The cutting force is measured by a sensor and the measured signal is used by the Identifier, who determines the sign which, according to the transfer function of the plant, at that time, would have given rise to measured cutting force, this sign represents in terms of controller's output, the error that this made, between the reference cutting force and reference force that the controller originated when defining the control signal.

The error that was obtained is used to adjust the weights of the Controller by using a training based algorithm that uses the delta rule. The identifier that tries modulating the reverse transfer function of the plant, it is also in the presence of a set of information that enables it to determine the error and continue to adapt the network weights.

2.1.2.1 Adapter

Due to difficulties of convergence of a single network, which implements the adapter, it was necessary to split the neuronal network into two, so that the intermediate layer of each could have a different number of neurons.

Networks are of type feed-forward multilayer, trained with the backpropagation algorithm. In **Fig. 4**, we can see how was finally formed the adapter. The entry is common to both Artificial Neural Networks, and each of these estimates the value for which was trained.

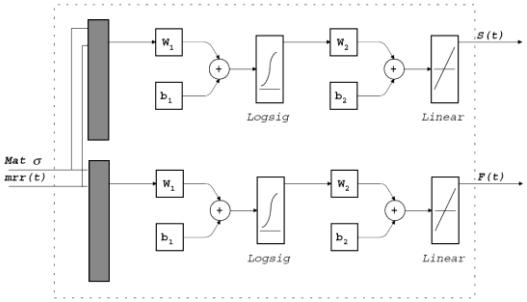


Fig. 4. Artificial neuronal network structure that implements the adapter

2.1.2.2 *Identifier*

The first step in the development of cutting force Controller was the development of Identifier from plant to control , because it is this which will in the future, adapt (train) the Controller to different processes.

The operation of the plant of our system, with non-linear characteristics, must be represented by a discrete equation type

$$Fc(t) = f(Fc(t-1), Fc(t-2),..., Fc(t-n), mrr(t), \\ mrr(t-1),...., mrr(t-m), d(t), d(t-1),...., d(t-k)) \quad (1)$$

That can be resolved in order to *mrr(t)*

$$mrr(t) = f^{-1}(Fc(t), Fc(t-1),..., Fc(t-n), mrr(t-1), \\ mrr(t-2),...., mrr(t-m), d(t), d(t-1),...., d(t-k)) \quad (2)$$

The developed identifier attempts to replicate the plant's function f^{1}. The adapter, from the Identifier point of view, is encompassed in the plant, and the Identifier must be able to adapt to errors which may be generated by the adapter.

The Identifier had to be developed on the basis of an Artificial Neuronal Network type TDNN – Time Delay Neural Network, since the plant to control is a nonlinear dynamic and complex system. In this case the network structure and the number of regressors were defined experimentally by using the test of several different structures. This does not mean that we have defined the best structure, but you have set yourself surely one of the best.

It was also proved, with data collected from the machine, that the problem's description was in fact in agreement with what is actually happen on the machine. The results obtained from the machine only differ from theoretical results because they have some noise and some disturbances that were not identified. In **Fig. 5** is represented the Artificial Neuronal Network structure, in which relied on the Identifier block.

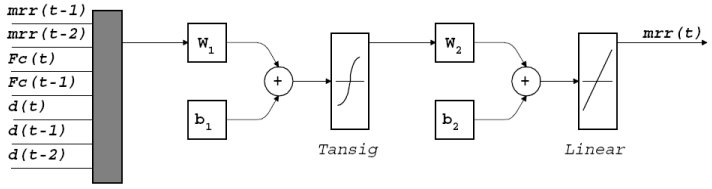


Fig. 5. Network structure that implements the Identifier

2.1.2.3 *Controller*

Contrary to what seems at first glance, the main problem in maintaining constant the cutting force, it is not due to the properties of the transfer function of the plant, but what we call the disturbance.

Disturbances are the changes which the cutting force suffers due to modifications of the machining process conditions (tool wear, part geometry changes, temperature changes, etc.). Generally, our Controller will have to maintain constant the cutting force and equal to the value of the reference force, from the feedback of a sensor that is able to detect the system cutting force.

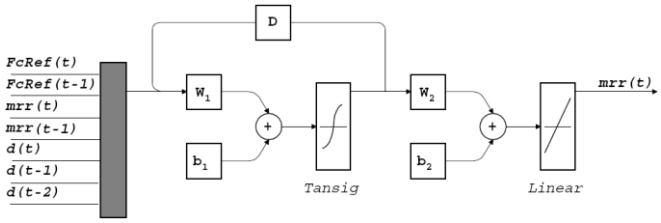


Fig. 6. Neuronal Network structure that implements the controller

Fig. 6, represents the Artificial Neuronal Network on which the control block is based. This neuronal network is of type recurrent (Elman Network) and was used the training algorithm type *backpropagation*.

3 Training and testing of neural networks

The training of all networks carried out in this section, have only intended to define the structure for the network that:

- It is possible to be implemented,
- Converge, i.e. a train that can at later on lead to desired results to network,
- Have initial knowledge that permits to be connected to the system.

The stop condition was, in general case, the satisfaction of these conditions. The training may have other special conditions to stop and in this case it will be described, locally, if they occur.

3.1 Adapter

Fig. 7-*a)* allows us to analyze what happened during training. In the small graph, inserted inside, you can check how the error curve has evolved along the 5000 iterations. The larger graph compares the values estimated by Artificial Neuronal Network (dashed line) with actual values used in the training (continues line).
Fig. 7-*a)* represents training of neuronal network that estimates the forward speed, whereas in **Fig. 7-*b)*** represents training of neuronal network that estimates the spindle speed. In **Fig. 8** we can see the error concerning two networks combined, in to the form of *F/S*

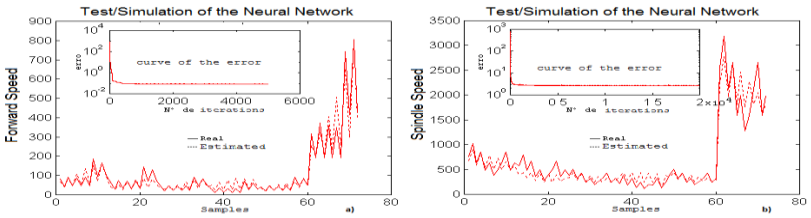


Fig. 7. Result of the training of neural networks that implement the Adapter

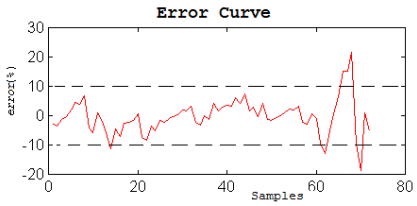


Fig. 8. Error curve relative to the value of entry *mrr (t)* and the output value of the adapter in the form *F/S*

3.2 Identifier

We were trained several settings for the neuronal network, networks with more or less neurons, and with more or less entries (regressors), however we can conclude, that the structure of the network that had better results was described above. After we trained the Identifier's artificial Neuronal network we obtain multiple results, which can be analyzed in the graphs that follow.

In **Fig. 9-*a)*** we have represented the learning rate of Artificial Neuronal Network. As we can analyze the error defined as (SSE), the sum of squared error, decreased along the iterations, reaching a value of 0.4 after 15000 iterations.

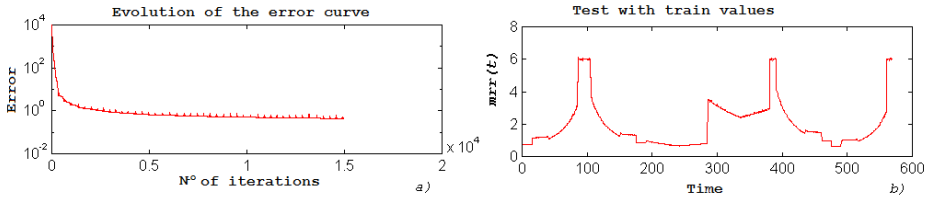


Fig. 9. Result of the neuronal network training ***a)*** evolution of square error ***b)*** the comparison between the value of *mrr(t)* estimated by network and the true value of *mrr(t)*

In **Fig. 9-*b)***, we can verify that the result of the identifier (*mrr (t)* estimated) approximates the signal used for training (*mrr (t)* real).

Fig. 10-a) represents the possible fluctuations in the value of the output of the plant (more exactly the value of the reference), due to the changes of reference force,

which will be applied to the Identifier. To this set is also added a set of possible disturbances to system ***Fig. 10-b)*.**

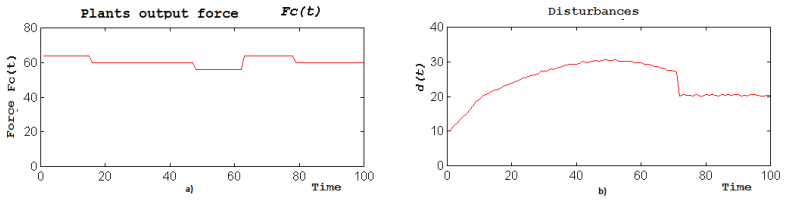


Fig. 10. *a)* Cutting force (*Fc*) measured at the output of the plant and *b)* Disturbances presented to the system

The value estimated by Artificial Neuronal Network that implements the controller (*mrr(t)*) and its actual value, can be seen in **Fig. 11**, in which the solid line represents estimated values, and another line (+) represents the actual value.

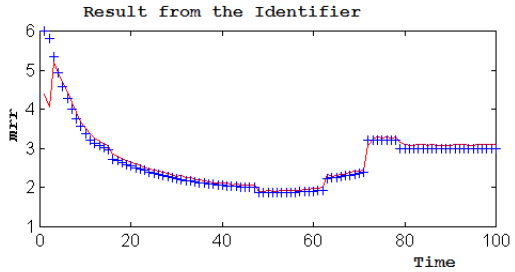


Fig. 11. Comparison between the true signal (represented by crosses) and the signal estimated by identifier (represented the solid)

In reality, we are only transmitting the initial knowledge to the Identifier that may continue to evolve throughout its live increasing its precision. As we described here it becomes clear it is not very important that training of the identifier must present error values too low, or that will be important the number of iterations performed.

3.3 Controller

The Controller, such as the Identifier, will have to act over time which means that its current status depends on variables involved and the immediately preceding State. The controller is also a network of type TDDN.

The examples used in controller's training were the same that we use to train the identifier, obviously adapted.

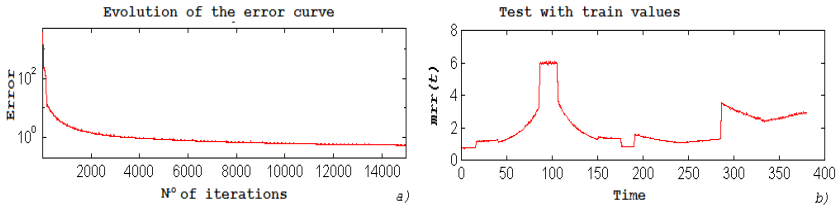


Fig. 12. *a)* - Learning curve of the controller. *b)* -simulation using training values

In **Fig. 12-*a)***, we can observe the error curve, which reached a value of 0.55 at the end of 15000 iterations. In **Fig. 12-*b)***, we have the curve representing the output values, used in training, almost completely overlapped to network response curve (estimated value) for the same input.

Once again, the result of the estimated value and the actual value are very close, being impossible to distinguish them in the chart.

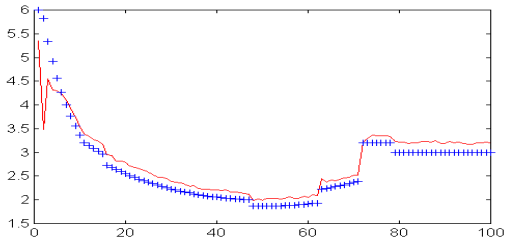


Fig. 13. Neuronal network Test, using different values from that used in training

Noted that the Controller should generate the same results as the Identifier. Somehow the Identifier oversees the results of the Controller. In **Fig. 13** values represented by the symbol ' + ', are the target values of the network, and the others represented by the solid line are the values estimated by the network. As we can see, these values are approaching.

The Controller we have defined could be a basis for a real implementation. The defined structure will probably have to be amended and adapted to reality. However, we will have a base that enables us to believe that it is possible to implement the controller in a real system.

3.4 System Tests

After we have implemented and tested individually each of the components of the Actuator, the next steps was to training and test the set. **Fig. 14-*b)***, represents the values to the output of the adapter when receives the input value *mrr (t)* generated by the Controller and the information that the material to be machined has a tensile strength equal to 20.

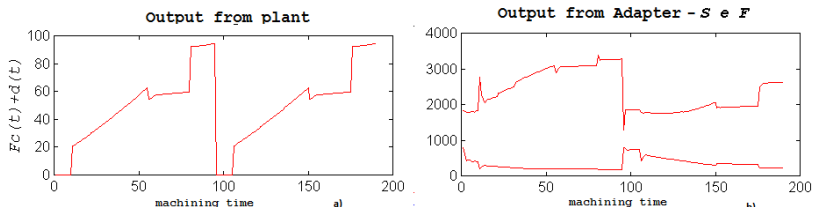


Fig. 14. *a)* Force measure at the plant output, *b)* values for *F* and *S* at the adapter's output

During the simulation the values generated by the adapter were used as information for the Identifier, which in turn, set the value of *mrr (t).* This value was then used to adapt (train) Controller.

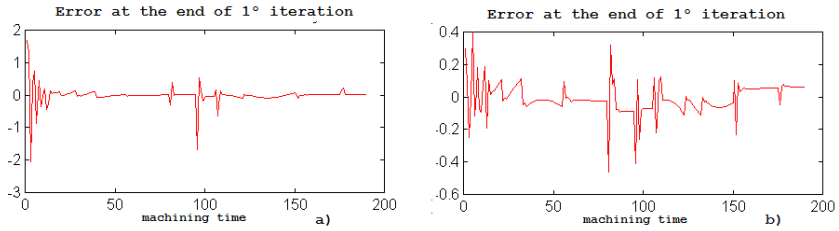


Fig. 15. Evolution of the error of the first iteration for the twenty

In **Fig. 15-*a)***, we can see the error at the end of the 1st iteration. With the increase of the number of times the Controller repeats the same operation the error decreases, being the end of the 20th piece represented in **Fig. 15-*b)***.

4 Conclusions

The most important conclusion at this time, as shown in the results, is that is possible to implement this type of control based on neural networks. Although we had some difficulties and the problem is not totally resolved, all data suggest that is possible such implementation.

Objectively, the whole work served to demonstrate that it is possible to implement the control of the cutting process based on Artificial Neural Networks. When we implement the system, all the considerations made about cutting force, linearity, disturbances, etc. have little importance if we are able to create the mechanisms that enable the system evolve through learning.

We note that the basic idea is to create a system capable of collecting information of process, so that is able to identify the error and adapt. In this system the main components are the Controllers and the Identifiers, which interacting to generate information that will enable both evolve.

The structures of Artificial Neural Networks defined for the different types of blocks should be sufficient to address the problems inherent in each of the tasks.

However, if is justified we can change the structures so that it can contemplate the new information.

We think it will be important, as a continuation of the work done so far, implement and test models of Artificial Neural Networks in real systems (machines). This would validate the structure of the Artificial Neural Networks used to implement the different blocks, and train them with real values.

At the present time the tool machines producers still trying to solve a set of problems linked to the integration of machines in flexible manufacturing cells. However will be inevitable that all brands producing CNC machines, will walk in the direction of the on-line control of cutting force. The competitive market will force it.

References

1. Acácio Teixeira da Rocha; Tecnologia Mecânica - Volume III; Coimbra Editora; 1977.
2. J. Paulo Davim; Princípios da Maquinagem; ALMEDINA; 1995.
3. Vitor Polónia; A Escolha da máquina - Ferramenta; FEUP - DEP. MAT. PROC. TEC.; 1983.
4. Jesus A. D. Rivera; Input/Output Linearization of Control Affine Sistems Using Neural Networks; A dissertation submitted to the cybernetics department in partial fulfilment of the rewuirements for the degree of doctor of philosophy; July1996.
5. Ole Ravn, Paul H S2rensen, Magnus N2rgaard; *What is Adaptative Control*; Neural Network Project; IAU; 1996.
6. Astrom & Wittenmark; *Adaptative Control*; Addison-Wesley; 1995.
7. W.S. Mclullch & W; Pitts *A logical caqlculars of the ideas immanent in neurons activity*; bulletin of mathematical biophisics 5 (115 -133); 1943
8. Jacek M. Zurada; *Introduction to artificial neural systems;* West Publishing Company; 1992
9. Eric Davalo and Patrick Naïm; *Neural Networks*; Macmillan; 1991
10. CCP & IAI; *MATIMAC- Machine Tool: Intelligent Monitoring and Control;* 1996
11. J.A. Feldman & D.H. Ballard; *Connectionist Models and their properties;* Cognitive Science, 6; 1982
12. Elman, J. L.; *Findingstructure in time*; Cognitive Scince, 14; 1990.
13. Steve Lawrence, C. Lee Giles, Ah Chung Tsoi; *What Size Neural Networks Gives Optimal Generalization? Convergence Properties of Backpropagation;* Technical Report, NEC Research Institute; 1996.
14. Ben J. A. Kröse & P. Patrick Van Der Smagt; An introdution to neural network;University of Amesterdam Faculty of Mathematics & Computer Science; 1993
15. D. O.Hebb; The Organization of Behaviour; Wiley; 1949
16. M. Minsky & S. Papert; Percetrons: An Itroduction to Computational Geometry; MIT Press; 1969
17. J. J. Hopfield; Neural Network and Physical Systems with Emergent Collective Computational abilities; Proceedings of ationalAcademy of Sciences 79; 1982

Learning Vehicle Traffic Videos using Small-World Attractor Neural Networks

Mario Gonzalez[1,2], David Dominguez[2], and Angel Sanchez[3]

[1] Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
mario.gonzalez@fe.up.pt
[2] EPS, Universidad Autónoma de Madrid, 28049 Madrid, Spain
david.dominguez@uam.es
[3] DCC-ETSII, Universidad Rey Juan Carlos, 28933 Madrid, Spain
angel.sanchez@urjc.es

Abstract. The goal of this work is to learn and retrieve a sequence of highly correlated patterns using a Hopfield-type of Attractor Neural Network (ANN) with a small-world connectivity distribution. For this model, we propose a weight learning heuristic which combines the pseudo-inverse approach with a row-shifting schema. The influence of the ratio of random connectivity on retrieval quality and learning time has been studied. Our approach has been successfully tested on a complex pattern, as it is the case of traffic video sequences, for different combinations of the involved parameters. Moreover, it has demonstrated to be robust with respect to highly-variable frame activity.

Keywords: Attractor network, Small-world, Sparse-coding, Correlated patterns, Temporal sequence, Video retrieval, Intelligent Transportation Systems, Vehicle Traffic analysis.

1 Introduction

Video Analysis involves processing information from sequences of digital images which are highly-correlated in time. In some cases, the video sequences are captured with a single camera, and its analysis exploits the temporal correlation from one frame to the next one in the sequence. In other situations, the sequences are obtained from several cameras, and the processing may involve reconstructing three-dimensional scenes from two-dimensional sequences captured by each camera. Many applications involving video analysis have been presented in domains like surveillance, manufacturing, video games, among others [1].

The application of video-based analysis to traffic surveillance [2] is an area of growing interest with the aim to detect both global events (i.e. number of vehicles in a road region) and local events (i.e. detection and tracking of a specific vehicle). As large amounts of video data are stored for analyzing the involved events on them, it becomes very important to develop efficient storage and retrieval techniques for these traffic videos. In general, these videos are sequences

of frames where the involved patterns (i.e. moving vehicles) are highly correlated in time, specially in traffic congestion scenes. Most of existing works for this problem use an approach based on scene segmentation followed by vehicle tracking [3]. In it, the vehicles are first detected in the dynamic scene using adaptive-background techniques [2] [4] and specific features like texture, color or shape [5], are extracted from the segmented targets for classification. Later, these vehicles are tracked using different techniques like optical flow [6], Kalman filters [7] or particle filters [8], among others. Segmentation and tracking tasks become more difficult on realistic traffic situations like possible vehicle congestions, variability of weather and/or illumination conditions. Moreover, the vehicle tracking results along time are highly dependent on a good segmentation of them. To avoid the need of segmentation and tracking individual vehicles, some holistic representations for the storage and retrieval of traffic videos have been proposed. Chan and Vasconcelos [3] propose a dynamic texture representation to model the motion flow in the scene. They use the Kullback-Leibler divergence and the Martin distance to retrieve and classify traffic videos without need of segmentation. Xie et al. [9] present another holistic method for traffic video retrieval using Hierarchical Self-Organizing Maps (HSOM). They extract the motion trajectories of the vehicles present in the video and these activity patterns are stored by the neural network, later this learned knowledge is combined with a semantic indexing stage to retrieve traffic sequences based on queries by keywords.

The aim of our work is to learn and retrieve a sequence of patterns that are highly correlated over time, obtained from a traffic video sequence. We use a Hopfield-type of Attractor Neural Network (ANN) with a small-world connectivity distribution. It is known that, for uniformly distributed (i.e non-correlated) patterns, the most efficient arrangement for storage and retrieval of patterns as a whole (global information) by an ANN is the random network. However, small-world networks with a moderate number of shortcuts can be almost as computationally efficient as a random network while saving considerably on wiring costs [10]. Furthermore, for non-uniformly distributed patterns, networks with spatially distributed synapsis are more efficient [11].

In order to achieve this objective one must face some typical problems found in the literature on ANN [12,13]. First, in real-world applications, such as video compression/retrieval, where patterns present high correlation, one has to deal with sparse coding patterns. Sparse-coding is the representation of items by the *strong activation* of a relatively small set of neurons [14]. This is a different subset of all available neurons when the patterns are uncorrelated. On the one hand, this sparse-coding gives the model a biological plausibility since the brain suggests a general sparse-coding strategy. This is physiologically relevant, because the amount of energy the brain needs to use to sustain its function decreases with increasing sparseness [15]. Sparse-coding is also favorable to increase the network capacity, because the cross-talk term between stored patterns decreases. On the other hand, it is difficult to sustain a low rate of activity in ANN and a control mechanism must be used [16].

Second, learning a sequence of time-correlated patterns is required by our application. The noise induced by the overlap between patterns is much larger for correlated patterns than for random patterns [17]. This implies that the network capacity drops down to an asymptotically vanishing value. Correlations between the training patterns, as it happens for a video sequence, worsens the performance of the network since the cross-talk term can yield high values in this case [18].

The contribution of this paper is twofold. First, we introduce a variant of the pseudo-inverse approach to learn/retrieve a sequence of correlated cyclic patterns (as it is the case of a video sequence) using a sparse-coding ANN with a small-world topology. Second, to demonstrate the feasibility of our approach for the storage and retrieval of traffic videos. The rest of the paper is organized as follows. A general solution to the problem based on the pseudo-inverse approach is detailed in the Section 2. The proposed model for this problem avoids the segmentation and tracking of the involved targets and also some closely related difficulties. Section 3 presents the experimental framework for complex traffic videos taken at a distance, where many vehicles appear in the scene of a Kiev crossroad and another video from a roundabout with light traffic. Results are presented and analyzed for different parameter settings. Finally, Section 4 concludes the paper.

2 Proposed model

This section introduces the topology and dynamics of the proposed ANN model where a variant of pseudo-inverse is used to compute the learning weights. The information measures used to determine the network performance, and the proposed threshold strategy to retrieve patterns with a low activity, are also described.

2.1 Neural Coding

We consider a network with N neurons and a fixed number of $K < N$ synaptic connections per neuron. At any given discrete time t, the network state is defined by the set of N independent binary neurons $\boldsymbol{\tau}^t = \{\tau_i^t \in [0,1]\,; i = 0, \ldots, N-1\}$, each one active or inactive denoted respectively by the state 1 or 0. The aim of the network is to retrieve a sequence of correlated patterns (in our case, the consecutive frames of the video sequence) $\{\boldsymbol{\eta}^\mu,\ \mu = 1, \ldots, P\}$ that have been stored during a learning process. Each pattern $\boldsymbol{\eta}^\mu = \{\eta_i^\mu \in [0,1]\,; i = 1, \ldots, N\}$ is a set of biased binary variables with sparseness probability:

$$p(\eta_i^\mu = 1) = a^\mu, \qquad p(\eta_i^\mu = 0) = 1 - a^\mu. \tag{1}$$

The mean activity for each pattern μ is $a^\mu = \sum_i^N \eta_i^\mu / N \equiv \langle \eta^\mu \rangle$. The neural activity for any time t is given by the mean: $q^t = \sum_i^N \tau_i^t / N \equiv \langle \tau^t \rangle$.

2.2 Network Topology

The synaptic couplings between the neurons i and j are given by the adjacency matrix $J_{ij} \equiv C_{ij}W_{ij}$, where the topology matrix $\mathbf{C} = \{C_{ij} \in [0,1]\}$ describes the connection structure of the neural network and $\mathbf{W} = \{W_{ij}\}$ is the matrix of learning weights. The topology matrix contains two types of links: the local and the random ones, respectively. The local links connect each neuron to its K_l nearest neighbors in a closed ring, while the random links connect each neuron to K_r others uniformly distributed in the network. Hence, the network degree is $K = K_l + K_r$. The network topology is then characterized by two parameters, the *connectivity ratio* γ and the *randomness ratio* ω, which are respectively defined by:

$$\gamma = K/N, \qquad \omega = K_r/K, \tag{2}$$

where ω plays the role of a rewiring probability in the *small-world* model [19,20]. Fig. 1 shows a topology example of the considered ANN.

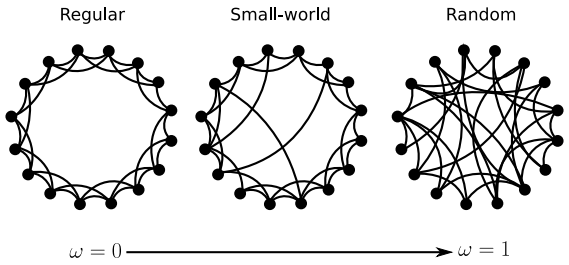


Fig. 1. An schematic representation of a small-world topology (Watts-Strogatz model) with $N = 16$, $K = 4$ and $\omega = 0.0$ (left), $\omega = 0.05$ (middle) and $\omega = 1.0$ (right).

The storage cost of this network is $|\mathbf{J}| = N \times K$ if the matrix $\mathbf{J}$ is implemented as an adjacency list, where all neurons have K neighbors.

2.3 Retrieval Dynamics

The task of the network is to retrieve the whole learned sequence of patterns (i.e., the full video sequence) starting from an initial neuron state $\boldsymbol{\tau}^0$ which is a given seed frame or a state close to it. The retrieval is achieved through the noiseless neuron dynamics:

$$\tau_i^{t+1} = \Theta(h_i^t - \theta_i^t), \tag{3}$$

$$h_i^t \equiv \frac{1}{K}\sum_j J_{ij}\frac{\tau_j^t - q_j^t}{\sqrt{Q_j^t}},\ i = 1,\ldots,N, \tag{4}$$

where h_i^t denotes the local field at neuron i and time t, and θ_i is its firing threshold. The local mean neural activity is $q_i^t = \langle \tau^t \rangle_i$, and its variance is $Q_i^t = Var(\tau^t)_i$. The local mean is given by spatial averaging: $\langle f^t \rangle_i \equiv \sum_j C_{ij} f_j^t / N = \sum_{k \in C_i} f_k^t / K$, for any given function f of the neuron sites. Here we used the step function:

$$\Theta(x) = \begin{cases} 1, \ x \geq 0 \\ 0, \ x < 0. \end{cases} \tag{5}$$

For convenience, we use in the paper some normalized variables, where the site and time dependence are implicit:

$$\sigma \equiv \ \frac{\tau - q}{\sqrt{Q}}, \ q \equiv \langle \tau \rangle, \ Q \equiv Var(\tau) = q(1-q) \tag{6}$$

$$\xi \equiv \ \frac{\eta - a}{\sqrt{A}}, \ a \equiv \langle \eta \rangle, \ A \equiv Var(\eta) = a(1-a), \tag{7}$$

where a and q are the pattern and neural activities, respectively. The averages computed in this work run over different ensembles, and are indicated in each case. These variables can be directly translated to those used in most works found in the literature for uniform (non-biased) neurons [17], in the case of $a = 1/2$.

2.4 Learning Dynamics

To state the proposed learning rule for storing cyclic patterns which are highly correlated, as it is the case of a video sequence, we will recall the expression of the weights for the standard case (static and uncorrelated patterns), and then two straightforward extensions: static and correlated patterns, and cyclic and uncorrelated patterns. Cyclic patterns correspond to sequences of patterns of variable activities, with periodic conditions [21], that means, the next to the last pattern is the first one, then $\xi^{\mu+P} = \xi^{\mu}$.

If the network learns a set $P = \alpha K$ of static and uncorrelated patterns, $\langle \xi^{\mu} \xi^{\nu} \rangle = 0$, these are stored by the network couplings W_{ij} using the classical Hebbian rule [22] for the Hopfield model:

$$W_{ij} = \frac{1}{N} \sum_{\mu=1}^{P} \xi_i^{\mu} \xi_j^{\mu}. \tag{8}$$

This rule for learning the weights can be generalized introducing a $P \times P$ matrix $A_{\mu\nu}$ in the following way:

$$W_{ij} = \frac{1}{N} \sum_{\mu,\nu=1}^{P} \xi_i^{\mu} A_{\mu\nu} \xi_j^{\nu}. \tag{9}$$

The standard case, given by Eq. **(8)**, is obtained by using an identity matrix $A_{\mu\nu}^I = \delta_{\mu\nu}$.

For the situation of learning static and correlated patterns, the pseudo-inverse approach [17] is a standard method to orthogonalize (i.e. to extract) the correlated patterns, and the matrix $A_{\mu\nu}$ is computed as follows:

$$A^C_{\mu\nu} = O^{-1}_{\mu\nu}, \quad O_{\mu\nu} \equiv \frac{1}{N}\sum_{i}^{N} \xi^{\mu}_{i}\xi^{\nu}_{i}, \tag{10}$$

where O is the $P \times P$ patterns overlap matrix.

For the case of learning cyclic (sequential with periodic conditions) and uncorrelated patterns the former Hebbian rule, Eq. **(8)**, combined with a row-shifting schema of the identity matrix can be applied [22]:

$$A^S_{\mu+1,\nu} = \delta_{\mu\nu}, \ \forall\mu \in [1,..,P-1], \quad A_{1,\nu} = \delta_{P,\nu} \ \forall\nu \in [1,..,P], \tag{11}$$

In the case of video sequences, we have cycles (or sequences of patterns) where there is a high temporal correlation between the successive frames. For this reason, we propose a heuristic where the learning weights are computed by combining the pseudo-inverse approach with a row-shifting schema, as the one used for cyclic patterns. The proposed heuristic for this case (i.e. cycles and correlated patterns) has the following four steps:

1. Obtain the pattern overlap matrix O.
2. Compute its inverse matrix O^{-1}.
3. Rotate forwards cyclically the rows of O^{-1} to obtain a new matrix M.
4. Substitute matrix A by the new matrix M in Eq. **(9)** to compute the weights matrix W for the video sequence to be learned.

The previous stages are detailed next. First, the $P \times P$ overlap matrix O, describing the video sequence is computed by Eq. **(10)**, and its inverse matrix O^{-1} is obtained next. This approach is thought to get fixed point solutions. However, if one is seeking a limit cycle solution (i.e. retrieving the whole sequence of frames cyclically), then one must benefit from the interactions between one frame and the next one in the video. Therefore, the elements of the O^{-1} matrix are shifted as shown schematically in the following equations:

$$A^V_{\mu+1,\nu} = O^{-1}_{\mu\nu}, \ \mu \in [1,..,P-1], \quad A^V_{1,\nu} = O^{-1}_{P,\nu}, \ \forall\nu \in [1,..,P], \tag{12}$$

obtaining the matrix A^V. The previous rule takes into account the dominant terms in the infra-diagonal positions of the matrix A^V. The sub-dominant terms account for the orthogonalization of the matrix O^{-1}. It is worth to note that the pseudo-inverse rule is a not local matrix, because the connections between every two neurons depend on the other neurons; it is also a non iterative rule, all patterns must be learned at the beginning of the retrieval process.

The learned weight matrix $\mathbf{W}$ is now calculated according to the rule in Eq. **(9)**, where $A_{\mu\nu}$ is computed by applying the row-shifting schema given by Eq. (12). The learning stage displays slow dynamics, being stationary within the time scale of the faster retrieval stage, as shown by Eq. **(3)**. A stochastic macro-dynamics takes place due to the extensive learning of $P = \alpha K$ patterns, where α is the load ratio.

2.5 Threshold Strategies

In order to retrieve patterns with low activity, it is necessary to use an adequate threshold of firing. If firing is not controlled, the neural activity could be higher (lower) than the pattern activity, whenever the threshold is too small (large).

The more sparse the code is, the more sensitive is the interval where the threshold can move [16]. On the one hand, one could use an optimal manually-chosen threshold, where for each learned pattern and initial condition, the retrieval is maximized. This is not a realistic strategy, since the neural network is not supposed to know the patterns during the retrieval process. Thus, a simple and convenient solution is to use a fixed value for the threshold. The value of $\theta_i = 1$ for the threshold was obtained experimentally for a sparseness ratio of $a \sim 0.1$, which is the mean sparseness of the frames in the analyzed videos.

2.6 The Information Measures

In order to evaluate the network retrieval performance, two measures are considered: the global overlap and the load ratio. The overlap is used as a temporal measure of information, which is adequate to describe instantaneously the network ability to retrieve each frame of the video. In this case, the overlap m_μ^t between the neural state σ^t at time t and the frame ξ^μ is:

$$m_\mu^t \equiv \frac{1}{N} \sum_i^N \xi_i^\mu \sigma_i^t, \tag{13}$$

which is the normalized statistical correlation between the learned frame η_i^μ and the neural state τ_i^t at a given iteration t in the sequence cycle. One lets the network evolve according to Eqs. **(3)** and **(4)**, and measures the overlap between the network states and the video frames running over a whole sequence cycle of the learned video. The neural states $\{\boldsymbol{\tau}^t,\ t = 1, \ldots, P\}$ are compared cyclically with the learned frames $\{\boldsymbol{\eta}^\mu,\ \mu = 1, \ldots, P\}$. The network starts in an initial condition close to a given frame, say $\tau^{t=1} \sim \eta^{\mu=1}$, so that the time and frame label are synchronized, and the overlap for each frame at cycle $c = 0, 1, 2...$ is:

$$m_\mu^c \equiv \frac{1}{N} \sum_i^N \xi_i^\mu \sigma_i^{\mu+cP}. \tag{14}$$

The global overlap is defined as:

$$m^c = \langle m_\mu^c \rangle \equiv \frac{1}{P} \sum_{\mu=1}^P m_\mu^c \tag{15}$$

and it measures the network ability to retrieve the whole sequence of patterns. After a transient period of time, the network dynamics converges to a stationary regime where the global overlap m^c doesn't change in the next cycles. When this global overlap between the whole set of patterns (i.e. the video sequence)

and their corresponding neural states is $m = 1$, the network has retrieved the complete sequence without noise. In this case, all the network states correspond perfectly to the frames of the video. When the global overlap m is zero, the network carries no macroscopic order. In this case, the video can not be retrieved. For intermediates values of m, where $0 < m < 1$, the video can be partially recovered with a given level of noise (when m increases, a higher number of frames can be perfectly retrieved).

Besides the overlap, we are also interested in the load ratio $\alpha \equiv P/K$, that accounts for the storage capacity of the network. This ratio depends on the size of the video, which is $P \times N$ (i.e. the number of frames by their spatial resolution, where this resolution coincides with the number of neurons), and the amount of physical memory necessary to store the video, which is $K \times N$ representing the adjacency lists sizes (see the network topology subsection).

When the number of stored patterns increases, the noise due to interference between patterns also increases and the network is not able to retrieve them. Thus, the overlap m goes to zero. A good trade-off between a negligible noise (i.e when $1 - m \sim 0$) and a large video sequence (i.e. a high value of α) is desirable for any practical-purpose model.

3 Experimental evaluation

The learning times to store our traffic video sequences were very high for the network considered. In our experiments, this time was highly dependent on the parameter K, as well as the number of learned patterns P, and it varies between 100 min and near 2000 min depending on the network degree considered. In fact the learning time is of order $O(N \times K \times P^2)$, according to Eq. **(12)**. That is why we have only used two video sequence for our experiments: the first one, *Kiev*, corresponds to a densely transited crossroad zone in Kiev, Ukraine; and the second one, *roundabout*, corresponds to a roundabout area in a Spanish city. Different model parameter configurations were tested for both sequences to get more insight on how the network behaved during the learning and retrieval of correlated cyclic frames. The *Kiev* video sequence was captured by a live camera demo site from Axis company:
http://www.axis.com/es/solutions/video/gallery.htm.
It was recorded by an Axis Q1755 Network Camera as an AVI video and consisted of 1835 frames at 25 frames per second, that is 73.4 seconds of recording. The original *roundabout* video sequence consisted of about 15 min of video which was recorded with a conventional camera at 30 fps with frames and we used only 650 frames, that is 21.7 seconds of video for our experiments.

For the two analyzed sequences, the video pre-processing included three stages:

(1) The frames of the initial color video sequence were converted into binary patterns and stored as PNG images with dimension 384×356 black-and-white pixels for the *Kiev* sequence and 640×480 pixels for the *roundabout* sequence.

(2) The *Kiev* frames were resized to $96 \times 89 = 8,544$ pixels and the *roundabout* frames to 80×106 pixels, in order to get a reasonable network size for the simulations.
(3) A new subsequence of frames was created by uniformly sub-sampling the sequence obtained in the previous stage using a natural factor f, where $f \geq 1$ (i.e., we build the video subsequence with original frames: 1, 1+f, 1+2f, ...). The goal is to ensure that the network is able to recover the whole stored sequence of frames. Consequently, we start testing with f=1, then f=2, and so on, until the condition holds.

For the simulations we have used a system with an Intel Core 2 Duo CPU E6750 at 2.66GHz and with 2GB of physical memory. The Octave image package [23] was used for processing the image files into text files with the 0/1 binary format as the neuron states required. The network parameters used in the *Kiev* simulations were $N = 8544, K = 4250, \theta_i = 1.0$ for a sparseness $a = 0.10$. For this network size, it has been recovered the video sequence each $f = 5$ frames, that is: $\frac{1835}{5} = 367$ frames. For the *roundabout* simulations a similar network were used with $N = 8480, K = 4240, \theta_i = 1.0$ for a sparseness $a = 0.07$, recovering the video sequence each $f = 5$ frames, that is: $\frac{650}{5} = 130$ frames. The video output comparing the original with the retrieved frames and the frames in text format can be found at: ftp://amaethon.ii.uam.es/video/video5/ for the *Kiev* sequence and at ftp://amaethon.ii.uam.es/video/roundabout/ for the *roundabout* sequence.

Fig. 2 and Fig. 3 show some sample post-processed frames of the stored and successfully retrieved video sequences for the *Kiev* and *roundabout* sequences, respectively. In Fig. 2 the seed used to start the retrieval was a noisy frame (top-left panel), with initial overlap $m_{\mu=1}^{c=0} = 0.5$. During the first cycle, the network is correcting the wrong pixels, (frame numbers 1, 21, 41 and 61 are presented in the top panels) $m^{c=0} \sim 0.93$, see Fig. 4. After a complete cycle the overlap reaches the stationary value of $m^{c=1} \sim 0.99$ (the same frames are shown in the bottom panels for the second cycle).

For the *roundabout* sequence in Fig 3 the seed was a noisy frame (top-left panel), with initial overlap $m_{\mu=1}^{c=0} = 0.4$. The frame numbers 1, 11, 21 and 31 are presented in the top panels for $m^{c=0} \sim 0.97$, and bottom panels for the second cycle with a stationary value of $m^{c=1} \sim 0.98$, see Fig. **??**.

3.1 Influence of the topology on the global overlap and the learning time

Using the previous network parameter setting (N,K,θ_i,a), Table 1 shows the dependence of global overlap and processing time on the random connections parameter ω at the learning stage.

As it can be observed, there is no significant difference between the processing time for learning the video with different values of ω and m parameters. This slight difference is only due to the larger times to construct random networks than to construct local networks. The retrieval time for all cases was the same,

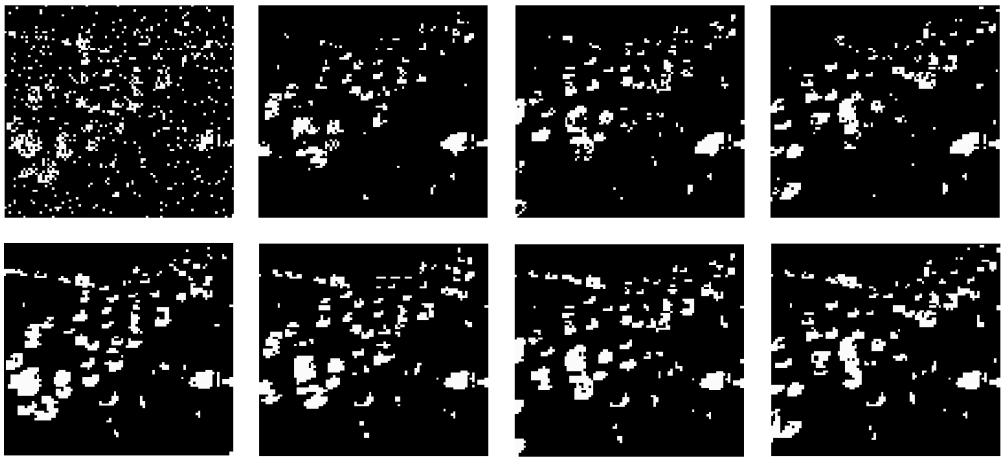


Fig. 2. Some retrieved sample frames (from left to right, frame numbers 1, 21, 41 and 61) of the Kiev crossroad traffic video sequence for $f = 5$. Initial overlap $m^1 = 0.5$. Top panels: first cycle. Bottom panels: second cycle.

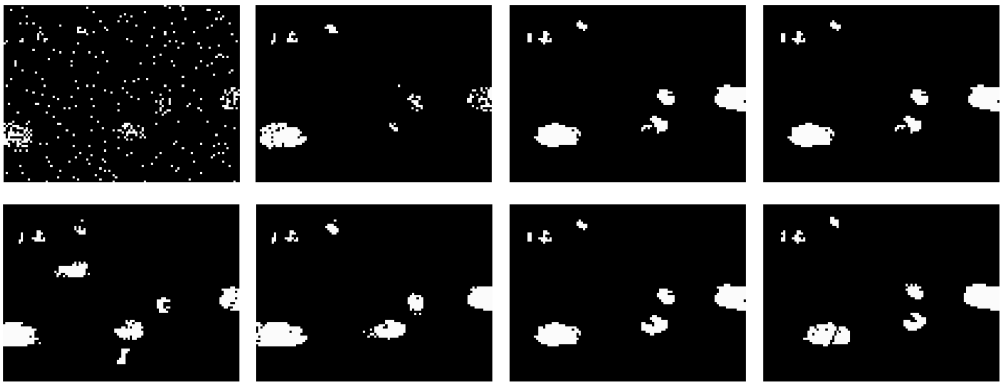


Fig. 3. Some retrieved sample frames (from left to right, frame numbers 1, 11, 21 and 31) of the roundabout traffic video sequence for $f = 5$. Initial overlap $m^1 = 0.4$. Top panels: first cycle. Bottom panels: second cycle.

Kiev crossroad			Roundabout		
ω	m	learning time	ω	m	learning time
0.0	0.32117	499m39s	0.0	0.32060	100m19s
0.3	0.33677	500m08s	0.4	0.20104	101m01s
0.4	0.99751	500m25s	0.6	0.05548	101m43s
0.5	0.99767	501m30s	0.7	0.98344	102m50s
1.0	0.94742	504m27s	1.0	0.99732	104m51s

Table 1. Randomness ratio versus global overlap and learning time for the Kiev crossroad and roundabout video sequences.

around 5 minutes for the *Kiev* and 1 minute and a half for the *roundabout* sequences. In all cases, the respective memory usages for the learning and retrieval stages are about 14.3% and 10.4% of the whole computer memory, respectively.

One can conclude that, with a network with a randomness value of $\omega = 0.4$, the retrieval of the *Kiev* video sequence is possible and it saved considerably on wiring costs as the small-world topology suggests. It is also interesting to remark in Table 1 that the transition from confusion state (i.e. $m \sim 0$) to the retrieval state (i.e. $m \sim 1$) for *Kiev* traffic video happened around $\omega = 0.35$. This is related to an effective percolation of the information over all the network. Although the network is always connected, for smaller values of the randomness parameter, the synaptic strengths are not strong enough to percolate the information from some pixels to every region of the neuron states. For the *roundabout* video sequence, the randomness value for the transition from the confusion to the retrieval state, $\omega = 0.7$, is higher than in the *Kiev* video. This effect could be due to temporal correlation between frames which is smaller for the *roundabout* video.

We also experimented with a simpler "shifted-diagonal" Hebbian learning matrix [21] replacing the pseudo-inverse rule(see Eqs. **(9-11)**). The maximal number of frames which could be retrieved for the *Kiev* video with $N = 8544, K = 4250, \omega = 0.5$ and with $m \sim 1$, was about $P = 16$. This choice is surely not appropriate for strongly correlated patterns and other learning rules like covariance rule [24] or the Bayesian rule [25] have been proposed to maximize the signal to noise ratio for a class of associative memories. A comparison with these models might be studied in a future work.

3.2 Robustness of the model with respect to the frame activity

We tested the robustness of the model (i.e. how overlapped the curves of average pattern and neural activities are along the frames of the video sequence) for a given network configuration: $N = 8544$, $K = 4250$ and $\omega = 0.4$. Fig. 4 shows that the model is robust against a variable frame activity level, where the normalized activity (i.e. sparseness) of the frames a^μ/a varies in the range $0.4 < a^\mu/a < 1.6$). This graphic can be partitioned in three regions according the numbering of the frames. In a first region, where m (black line) varies from 0.55 to around 0.95 (from first frame to around frame 20), the average pattern (red line) and the neural (blue line) activities are uncorrelated and pattern activity is much higher than temporal neural activity. In a second region, where the value of m remains stable around 0.95 (from frame 21 to frame 225), the average pattern and neural activities are highly correlated but pattern activity is slightly larger than temporal neural activity. Finally, in the third region, where m equals to one from 226 to the end of the video, the pattern and neural activities are exactly coincident despite the significant changes in frame activity over time. The global overlap for the cycle is $m^c = 0.93$.

A similar curve for the roundabout sequence is presented in Fig. **??**. for a network configuration: $N = 8480$, $K = 4240$ and $\omega = 0.4$. The overall behavior is similar to the *Kiev* sequence.

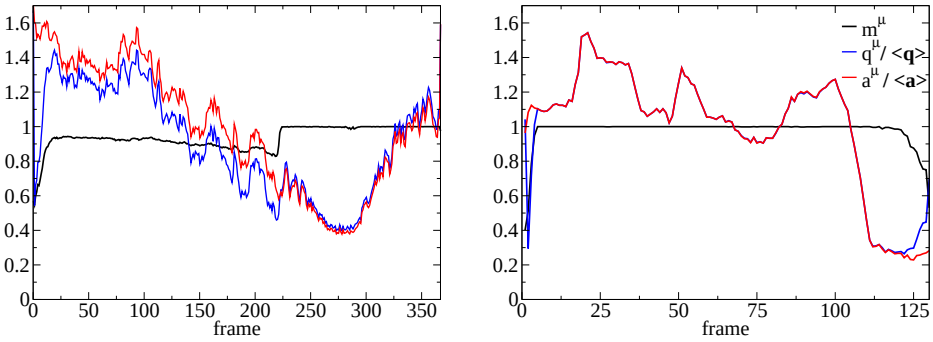


Fig. 4. Left: Kiev crossroad sequence: Plot of overlapped pattern and neural activities against frames for $N = 8544$, $K = 4250$ and $\omega = 0.4$. Initial overlap $m_{\mu=1} = 0.5$. Right: Roundabout: Plot of overlapped pattern and neural activities against frames for $N = 8480$, $K = 4240$ and $\omega = 0.7$. Initial overlap $m_{\mu=1} = 0.4$. (Color on-line)

4 Conclusion

We used a Hopfield-type of Attractor Neural Network (ANN) with a small-world connectivity distribution to learn and retrieve a sequence of highly correlated patterns. For this network model, a new weight learning heuristic which combines the pseudo-inverse approach with a row-shifting schema has been presented. The influence of the random connectivity ratio on retrieval quality and learning time has been studied. Our approach has been successfully tested for different combinations of the involved parameters on a complex traffic video sequence. Moreover, it was demonstrated to be robust with respect to highly-variable frame activity.

Another additional conclusion of our study is that the more spatially correlated the frames are in average, the smaller is the range of the interaction (randomness parameter ω) which optimizes the retrieval of the video. The opposite also holds: the less spatially correlated the patterns are, the higher should be the value of ω. For instance, if there are large regions in the frames with high activity (i.e., a huge truck or bus in the corner) in a bulk of still background of the frame, then it is strongly spatially correlated. On the other hand, the threshold strategy used in the model is fundamental, since the dependency of θ with the neural activity (as well as with the pattern activity) is set in such a way that the network dynamics is self-controlled and it does not need from any human participation. For example, with the typical activity value used $a = 0.1$ in our traffic video, we set $\theta \sim 1$ in most of the network. For a uniform activity degree in the frames (i.e. $a = 1/2$), no threshold is needed ($\theta \sim 0$). Finally, for extremely sparse code (where $a \to 0$), the threshold increases to $\theta \sim 1/\sqrt{a}$.

Automatic video-based traffic monitoring systems are an alternative to loop detectors. Video-based systems provide updated global information on the analyzed traffic scene and also specific informations of the tracked vehicles. An interesting application of such systems is content-based traffic video retrieval,

where using a query video it is possible to retrieve another similar video from a database using some types of extracted features from the videos (i.e. textural information, motion trajectories of cars, etc). This can be useful for surveillance applications where we are interested in detecting certain events on the video (i.e. accidents, congestions, etc). To achieve this goal, most approaches follow a feature-extraction approach which needs to segment the cars in the video and to track them individually. In a different way, using a holistic method like the proposed in this paper we can retrieve a complete video from a query frame if this frame represents a noisy scene of the video.

As our approach is holistic in the sense that no segmentation and feature extraction from the vehicles is required, we have to consider other holistic approaches applied to traffic videos for comparison purposes. The mentioned papers by Chan and Vasconcelos [3] and Xie et al. [9] do not segment the vehicles in the video, but they extract some global features from it (in particular, the complete motion information contained in the video), which are used for the retrieval task. They retrieve instances of traffic patterns using query videos; while in our approach the video can be retrieved using only a unique (possibly noisy) query frame. Moreover, the two compared papers do not quantitatively measure the video retrieval quality as we do using the global overlap.

Up to our knowledge, this is the first application of small-world ANNs and a row-shifting pseudo-inverse method to this specific content-based video retrieval problem. Our proposed solution is suitable for the mentioned traffic application since it produces accurate retrieval results at reasonable time. However, the required learning times are still very large and the system needs improvement to be competitive with respect to those classical methods which segment the scene and track the moving targets. Moreover, our proposal can be now suited only to those traffic applications where the learning stage can be carried out off-line. Consecuently, the use of complementary more-efficient strategies to compress the amount of memory required to store the patterns vectors like look-up tables [11] or hashing techniques like LSH [26] will be considered as future work.

5 Acknowledgements

This research has been partially supported by the Spanish projects TIN2008-06890-C02-02 and TIN-2007-65989. M. Gonzalez thanks EM ECW Lot 20 for financial support. We thank F. B. Rodriguez for useful discussion.

References

1. A. Bovik and J. Gibson (eds), *Handbook of Image and Video Processing*, Academic Press , (2000).
2. V. Kastrinaki, M. Zervakis and K. Kalaitzakis, A survey of video processing techniques for traffic applications, *Image and Vision Computing* **21**, 359-381, (2003).

3. A.B. Chan and N. Vasconcelos, Classification and Retrieval of Traffic Video using Auto-Regressive Stochastic Processes, *Proc. IEEE Intelligent Vehicles Symposium*, (2005).
4. Y.K. Jung and Y.S. Ho, A Feature-Based Vehicle Tracking System in Congested Traffic Video Sequences, Proc. PCM'01, LNCS 2195, 190–197, (2001).
5. E. Bas, Road and Traffic Analysis from Video, *Master Thesis*, Koc University, Turkey, (2007).
6. B. Li and R. Chellappa, A generic approach to simultaneous tracking and verification in video, *IEEE Trans. on Image Processing* **11**, 530–544, (2002).
7. J.W. Hsieh, S. Hao, Y.S. Chen and W.F. Hu, Automatic Traffic Surveillance Systems for Vehicle Tracking and Classification, *Proc. IEEE Conf. on Intelligent Transportation Systems*, vol. 7, 175–187, (2006).
8. B. Ristic, S. Arulampalam and N. Gordon, *Beyond the Kalman Filter: Particle Filters for Tracking Applications*, Artech House, (2004).
9. D. Xie, W. Hu, T. Tan and J. Peng, Semantic–based traffic video retrieval using activity pattern analysis, *Proc. Intl. Conf. on Image Processing (ICIP'04)*, 693–696, (2004).
10. L. Morelli, G. Abramson and M. Kuperman, Auto-associative memory in a small-world neural network, *Eur. Phys. J. B* **38**, 495–500, (2004).
11. A. Knoblauch, G. Palm and F.T. Sommer, Memory Capacities for Synaptic and Structural Plasticity, *Neural Computation* **22**, 289–341, (2010).
12. K. Koroutchev and E. Koroutcheva, Bump formation in a binary attractor neural network, *Phys. Rev. E* **73**, 026107, (2006).
13. D. Dominguez, K. Koroutchev, E. Serrano and F.B. Rodriguez, Information and Topology in Attractor Neural Networks, *Neural Computation* **19**, 956–973, (2007).
14. B.A. Olshausen and D.J. Field, Sparse coding of sensory inputs, *Current Opinion in Neurobiology* **14**, 481–487, (2004).
15. P. Foldiak and D. Endres, Sparse coding, *Scholarpedia*, 3(1):2984, (2008).
16. D. Dominguez and D. Bolle, Self-Control in Sparsely Coded Networks, *Phys. Rev. Lett.* **80**, 2961, (1998).
17. J. Hertz, J. Krogh and R. Palmer, *Introduction to the Theory of Neural Computation*, Addison-Wesley, (1991).
18. T.P. Trappenberg, *Fundamentals of Computational Neuroscience*, Oxford University Press, (2002).
19. D.J. Watts and S.H. Strogatz, Collective dynamics of small-world networks, *Nature* **393**, 440–442, (1998).
20. D. Dominguez, M. González, E. Serrano and F. B. Rodríguez Structured information in small-world neural networks, *Phys. Rev. E* **79**, 021909, (2009).
21. C. Molter, U. Salihoglu and H. Bersini, Storing static and cyclic patterns in an Hopfield neural network, Technical Report, Université Libre de Bruxelles, (2005).
22. D.J. Amit, *Modeling Brain Function: The World of Attractor Neural Networks*, Cambridge University Press, (1989).
23. Octave GNU Homepage, http://www.gnu.org/software/octave/, (2006).
24. P Dayan and DJ Willshaw,:253–265. Optimising synaptic learning rules in linear associative memories, *Biol. Cybernetics* **65**, 253–265, (1991).
25. A. Knoblauch, Optimal Synaptic Learning in NonLinear Associative Memory, *IJCNN* **167**, 3205–3211, (2010).
26. A. Gionis, P. Indyk and R. Motwani, Similarity Search in High Dimensions via Hashing, *Proc. of the 25th Very Large Database Conference (VLDB'99)*, (1999).

Towards the next-generation traffic simulation tools: a first evaluation

Zafeiris Kokkinogenis, Lúcio Sanchez Passos, Rosaldo Rossetti, Joaquim Gabriel

FEUP, University of Porto, Rua Dr. Roberto Frias, 4200-465, Porto, Portugal
{pro08017, pro09026, rossetti, jgabriel}@fe.up.pt

Abstract. Future Urban Transport (FUT) describes all desired features that are currently being envisaged within the umbrella of Intelligent Transportation Systems. With advances in computer technology and communication, as well as elevating the user to a central concern, rather than favouring performance only, both the scientific community and practitioners are in search for adequate ways to model and assess new performance measures brought about by FUT's requirements. After identifying such requirements, we'll try to propose a taxonomy on the basis of diverse criteria to assess how suitable currently available simulation packages are to assess future urban transport. Some tools are compared and their ability to suit these needs is discussed resulting as a first appraisal of suitability of traffic simulation tools. On the basis of the agent metaphor and the concept of multi-agent systems, we suggest ways in which to follow up this work.

Keywords: Future Urban Transport, Agent-Based Simulator, Comparison Taxonomy.

1 Introduction

Computational simulation has the advantage of allowing the assessment of system's behaviour before it is developed or produced in real life. This technology helped man to go to the moon and, now, has a wide use on engineering in general in various sectors aiming to improve products quality, characterize system's output without any real implementation. They are based in mathematical models, which take into account responses and constraints of the system to be simulated.

This technique only achieves its optimal applicability when some conditions are observed. First, the problem cannot be easily solved using common sense, simple calculation, analytical methods, and direct experiments. The model needs to consider the system complexity that in many cases is hard to capture. If the system's simulation is appropriate, we can then provide practical feedback to real systems, time compression or expansion, higher control, and lower costs.

Thus, one question emerges: why is it necessary to simulate traffic (road network)? In order to give an answer to this query we have to keep in mind the system's complexity. Traffic simulation is necessary because this kind of application

domain is inherently complex, usually formed of diverse entities (vehicles, traffic controllers, pedestrians etc.) that present different interactions reflecting social behaviours (e.g. competition, collaboration). In such case, mathematical analyses are complex and deals with traffic as a whole, using flow equations to describe vehicles and pedestrian movements. Moreover, simulation can provide comparison studies between new infrastructures, controls without interfering in the real system neither spending resources. Concerning to the latter, it can be used as training set for real systems, because of the time compressing characteristics that condense information and create hypothetical situations, in addition to all advantages aforementioned.

Nevertheless, a new generation of mobility systems became evident with the advent of what has been coined Intelligent Transportation Systems (ITS). Embedded systems, wireless communications, and artificial intelligence are integrated to provide a new experience to the user. Not only, within the concept of Future Urban Transport systems (FUT), the notion of mobility system overcomes its limit, from a simple process of transportation of good and persons becomes more conscious in terms of environment, accessibility, equality, security, and sustainability of resources. However, long path is necessary to be traversed to accomplish this achievement. Scientific community studies how to create virtual scenarios to address such issues.

In this work we basically recall many of the different aspects involved in the FUT system and identify some requirements for a simulation platform to support it. First, we propose a taxonomy covering FUT requirements. We also select potential simulators that can fulfil these features, either directly or indirectly, taking into account the accessibility to them. Therefore, all simulators were analysed by the taxonomy. Nonetheless, we identify existing characteristics and lacking features of the simulators in exam and further recognize potential application of the simulation tools in next generation transport systems. Our main goal is to evaluate the chosen simulators with respect to the FUT platform, and create interesting and challenging issues that can be discussed.

The remaining part of this work is organized as follows. In section two, we shall present a definition of FUT related concepts and its requirements. We also present an overview of Artificial Transportation System (ATS) concept. The differences between abstraction levels used in traffic simulation, as well as an overview of all the chosen simulators are discussed. Keeping all these concepts in mind, in section four we present the proposed taxonomy to compare the simulation tools, aiming to fulfil FUT requirements. Finally, in section five, we present the simulators analysis followed by some conclusions and suggestions of trends for our future work as well as issues to be further addressed.

2 Future Urban Transport

In the last century, since the economy became based on trading, the transport turned into an essential component of our society. These systems have become rather complex and extremely large, being geographically and functionally distributed, both in that respecting structure and management. Thus, first contemporary transportation

revolution began with the introduction of electromagnetic communication, allowing greater integration through exchanging information more quickly and efficiently. Then, low-cost digital systems started playing an imperative role in transportation, improving efficiency of traffic control and coordination, as well as transport planning. Nevertheless, as transportation systems are becoming very large, both in terms of structure and dimension, the whole process of acquiring information from all sources, processing the essential data and providing adequate responses timely is rather a very arduous task.

Finally, the third revolution puts the user as a central aspect of transportation system, forcing architectures to become adaptable and accessible by different means so as to meet different requirements and a wide range of purposes. This novel scenario needs new technologies, methodologies, and paradigms that practitioners and the scientific community are hardly working on. On the other hand, discussions are still fostered by current ambitions to the Future Urban Transport systems, even more conscious in terms of environment, accessibility, equality, security, and sustainability of resources. Some of the main features are as follows [1].

- *Automated computation, Flexibility and Freedom, Accuracy, Intelligent Infrastructures, New communication technologies, and Distributed architecture.*

By observing all FUT requirements, a question comes to our mind: How are we going to evaluate FUT systems? With the emergence of new paradigms, it is also needed to define new ways to measure if and how these requirements are being fulfilled. Current metrics, such as flow, time spent, storage utilization, and others, are not enough to totally evaluate these systems. Thus, as this new concept is user-centred, metrics also have to reflect user satisfaction (e.g. services availability, flexibility, and scalability). Also, many of the current measurements will be interpreted in a correlated form to extract system's nuances.

Artificial transportation systems according to Wang et al [2]: "*is a generalization of the traffic simulation, which integrates the transportation system with other urban systems, such as logistic systems, social and economic systems, etc., to behave as a coordinate tool for transportation analysis, evaluation, decision-making and training*." Due to the high complexity and uncertainty of the transportation systems, traditional simulators are not able to capture the dynamics that characterize them. Persons can choose whether to travel or not, can change in any moment their planned routes, their choice can be affected by social or economic or environmental phenomena. Current transportation solutions are achieving high degree of autonomy and starts interacting with the user in a different dimension, as their peers. This turns the society into a system formed by multiple heterogeneous components with its own idiosyncrasy.

Simulation is a key component in this new step of mobility systems, due to the increased complexity in the test and validation task, which is especially more byzantine due to real-time constraints and the presence of heterogeneous participating entities (vehicles, urban infrastructures, traffic infrastructures, pedestrians etc.). In our

view the new platform must support, natively or by extension, distributed and autonomous decision-making capabilities, that is the Multi-Agent paradigm, different types of communication techniques, it should simulate various types of heterogeneous entities providing as realistic as possible easy results and last should offer on-line simulation visualization, in order the user can extract the desired information.

3 Simulation of Traffic and Transportation Systems

The section bellow will give a small introduction to traffic simulation, starting with a description of different level of abstraction that can exist on it. In sequence, a brief overview of all simulators is presented so the reader can understand the tool's focus and functionalities.

3.1 Traffic Simulation Approaches

Traffic simulation is largely studied and can be classified in four types: macroscopic, mesoscopic, microscopic and nanoscopic. Macroscopic simulation models the flow of traffic using high-level mathematical models often derived from fluid dynamics, thus it is continuous simulations. This type of simulation handles every vehicle in the same way and as group. It uses aggregate input and output variables such as speed, flow and density. Macroscopic simulators are most useful for the simulation of wide-area traffic systems, which do not require detailed modelling, such as motorway networks and interregional road networks. This approach is not very realistic because in real life there are many different types of vehicle driven by different individuals who have their own styles and behaviours. However, it is fast and accurate but is not well suited to urban models in general.

Microscopic simulators model individual entities (e.g. vehicle, driver etc.) separately at a high level of detail, and are classified as discrete simulations. Here, interactions are usually governed by the car-following and lane-changing logics. Thus, traffic flow details, usually associated with macroscopic simulation are the emergent properties of the microscopic simulation. These simulators can model traffic flow more realistically than macroscopic simulators do, due to the extra details added in modelling the simulated entities individually. Microscopic simulators are widely used to evaluate new traffic control and management technologies as well as performing analysis of existing traffic operations.

On the other hand, mesoscopic simulators fill the gap between macro and micro simulators. They normally describe traffic entities at a higher level of detail, than macroscopic models but their behaviours and interactions are in a lower level of detail. In mesoscopic model, vehicles can be grouped in packets, which are routed throughout the network and are treated as one entity. Other paradigm is that of individual vehicles that are grouped into cells to control their behaviour. The cells traverse the link and vehicles can enter and leave cells when needed, but not overtake.

A new trend of traffic simulation is the nanoscopic model that extends the vehicle vision, dividing it in parts. It is mostly used in autonomous driving and is in a strict

relationship with automated robotic, because of the need to simulate sensors. Figueiredo et al [3] observed a great potential use of robotic simulators in the autonomous driving field, motivating an information exchange among robotic and traffic study groups.

3.2 Simulators Overview

In this paper we concentrate our attention to the microscopic type of simulation. There is a huge amount of traffic simulators available nowadays, with different features and choosing a certain tool depends very much on the project's requirements. In our study these are the FUT requirement, so we would like to find simulators that can support these characteristics. A preliminary study was done to filter and avoid lost of work. Thus, as first step, we try to improve the simulators list presented in Algers et al [4] and we select seven possible options most used by practitioners and researcher, cited below.

VISSIM

VISSIM [5], is developed by PTV in Germany. Its application ranges from traffic engineering, public transport, urban planning over fire protection to 3D visualization. Further, VISSIM uses a microscopic flux model of discrete, stochastic, and based on time step traffic. This simulator considers as one the pair vehicle/driver, also content a psychophysical model to car-following and lane changing based on rules. The package has two programs to form a traffic simulator: flux model for microscopic traffic and state generator (e.g. based on small time steps to get data directly from the simulator). These parts interact in a 1 second frequency.

PARAMICS

PARAMICS [6] is a microscopic traffic tool developed by SIAS Ltd and Quadstone Ltd of Scotland and is designed for a wide range of applications where traffic congestion is a predominant feature. Its modules combine together to improve usability, integration and productivity allowing users and clients to get added value from the modelling process. It is produced by Quadstone and has a package for software models to be used with a microscopic simulation as simple as an intersection, or complex traffic networks. The toolkit for developers provides access to data from infrastructure, control, communication, and other application, also create and improve behavioural models, independent of its complexity.

AIMSUN

AIMSUN [7] is developed and marketed by TSS. It is used to improve road infrastructure, reduce emissions, cut congestion and design urban environments for vehicles and pedestrians. Three simulation types are present in this tool: the traffic distribution and allocation, a mesoscopic and microscopic simulator. The

microscopic model is based in car-following, lane changing and gap acceptance algorithms. Thus, mesoscopic offers an additional option to model big nets and is less restrict in terms of calibration than the microscopic.

MITSIM

MITSIM [8] is a simulation tool that was developed for evaluating the impacts of alternative traffic management system designs at the operational level and assisting in subsequent design refinement. Examples of systems that can be evaluated with MITSIM include advanced traffic management systems (ATMS) and route guidance systems. MITSIM was developed at MIT's Intelligent Transportation Systems Program. MITSIM is a synthesis of a number of different models and has the following characteristics: represents a wide range of traffic management system designs; models the response of drivers to real-time traffic information and controls; and incorporates the dynamic interaction between the traffic management system and the drivers on the network.

SUMO

"Simulation of Urban Mobility" (SUMO) [9] is an open source, highly portable, microscopic road traffic simulation package designed to handle large road networks. The simulator is developed in the Institute of Transportation Systems at the German Aerospace Center. It is licensed under the GPL. Its features include: collision free vehicle movement, multi-lane streets with lane changing, fast execution speed, dynamic user assignment, and other.

MAS-T2erLab

MAS-T2er Lab [10], a tool developed by MAS-T2er Lab Group in the Artificial Intelligence and Computer Science Laboratory (LIACC), is an integrated multi-agent system that applies a methodological approach that allows for the assessment of today's intelligent transportation solutions through the metaphor of agents. So, the application domain is conceptualized in terms of agents and three basic subsystems are identified, namely the real world, the virtual domain, and the control strategies inductor.

ITSUMO

ITSUMO [11] was developed by MASLAB TRAFFIC from Universidade Federal do Rio Grande do Sul (UFRGS) in Brazil. It can use both off-line and on-line information (e.g. traffic flow). The information regarding the network is stored in a XML file. In addition to the cellular-automata approach, one can also define other driver decision-making procedures via a special, optional, module. A visualization module retrieves data originated from the microscopic simulation and exhibits a

graphical representation of the traffic simulation. Four distinct modules thus compose the ITSUMO system: the data module, the simulation kernel, the driver definition module, and the visualization module.

4 Proposed Taxonomy

As stated before, the goal of this study is to find out the ideal structure of a microscopic simulator that can be applied (or at least be adapted) to FUT simulation. Based on Section 2.1 we create a new taxonomy to compare the chosen tools, cited above. First, the simulator should be extensible so we can create our own scenario, techniques, and entities. Parallelism/distribution is a recommended feature for large computational enforcement ad realistic scenarios. The supported simulated entities (e.g vehicles, pedestrian, traffic and urban infrastructure among others) and the simulation approaches are, also, basic points to observe as well. In that case, agent-orientation, is a main feature; Multi-Agent System (MAS) can simulate very closely behaviour of comprehensive heterogeneous systems where another approach of simulation fails. Microscopic traffic simulators based on MAS can model traffic system in realistic manner. Finally, is important the simulation tool to be user-friendly. Accounting for these basic criteria, the taxonomy suggested in this work is formed by the following six items, as described below.

Extension - How extension is made? Which strategy? How deep can be the modifications? - A simulator is composed by a kernel and aggregated modules. We want to see in this item if the number of modules is extensible and, also, which properties of the kernel can be changed.

Parallelism/Distribution - Does the simulator use parallel and/or distributed techniques? - To simulate large scenarios, the tool can take maximum advantage of the computer performance. In this item, we want to observe the used technique by the simulator to perform complex analysis in short time.

Simulated Entities - Which are the simulated entities? Is it extensible? - Different actors must be considered in a simulation depending in the constraints. All entities must be enumerated in this item, including infrastructure (because in FUT it has to be intelligent), and if the user can add his/her entity.

Agent-oriented - Can the simulator support agents? How this can be done? - Agent paradigm can support FUT distributed characteristics. To provide that a platform have to deals with local information and be able to acts locally. These features will be analysed in this item.

Simulation type - What is the simulation type performed by it? - As explained early, transportation can be simulated in different levels to accomplish different goals. FUT platform, ideally, has to be able to work with all levels, and in this item we enumerate the support abstraction levels.

Visualization - How can we visualize the simulation and its results? Is the visualization tool integrated in the simulator core or a different application? - Albeit this item seems to be dispensable, it is not because we need to analyse and see result, not only simulate the system. We see the type of visualization and if the visualization module is or not integrated with the simulator's kernel, to provide remote access.

We believe the aforementioned criteria are relevant and imperative for us to decide whether a simulation tool is adequate and enough to assess the new performance measures brought about by FUT's requirements. Of course many other criteria might be pointed out. However, we have fixed the above ones for this first appraisal, leaving further assessments as future work following up this paper.

5 Simulators Analysis

Starting with *extension* VISSIM, it uses COM port to communicate with external components, providing full control over some parts, such as the network topology, signal control, path flows, vehicle behaviour, and evaluation data. It allows one to program large applications using Visual Basic, Visual C++ or other applications' macro and script languages (e.g. MS EXCEL).

PARAMICS counts on a powerful tool named PARAMICS Software Development Kit (SDK), which allows users to augment tool engine with new functions that can override or replace simulator's model. The access is total to the simulator core, existing two types of functions: call-back (used for providing information about the attributes of vehicles and their environment), and control (as the name say, can control entities).

A collection of tools is offered by AIMSUM, called AIMSUN NG. It is composed by three programming possibilities. The AIMSUM API user can code extensions using C++ or Python, modifying the current simulation by changing, for instance, driver parameters, control timing, implementing powerful traffic management actions. Other option is Scripting, mainly used to quick, and not so deep, extension on simulator core. For last, Software Development Kit, aimed to C++ programmers, offers access both the Kernel and the Graphical User Interface allowing the inclusion of new functionalities at both levels, adding new models, new graphical elements, new editors and dialogues.

On the other hand, MITSIM does not provide any extension functionality or interface. Nonetheless, accounting for the fact it is an open-source project, the users can arguably modify its core and extend it. Another open-source project name SUMO uses the TraCI layer to control it through TCP connections. However, it has limited functions to control the simulation process and requires that information from the simulator be gathered though sensors spread out over the network. MAS-T2er Lab uses UDP connection to provide extension and user cannot access internals algorithms and vehicle states. Not enough information has been found about the ITSUMO

package, although its extension seems to be limited to the creation of driver and semaphore controller agents by the user.

To compare simulators with respect to the second item, first we must establish the differences between parallel and distributed systems. To implement these two techniques, multiple processors are needed, further the distinction is the memory used. Parallel systems have shared memory among all CPUs, albeit in that regarding distributed systems, there exists a local memory per CPU that communicates data between processes. So, VISSIM, AIMSUN, MITSIM, MAS-T2er Lab, and ITSUMO are just parallel. SUMO due to its simplicity is neither distributed nor is parallel and just PARAMICS distributed.

Table 1. Implemented entities in the microscopic traffic simulators

	Car	Bus	Truck	Train	Bicycle	Pedestrian	Vehicles + Pedestrians	Others
VISSIM	Yes	Yes	Yes	Yes	Yes	Yes	Yes	TL,PCL,Detectors
PARAMICS	Yes	Yes	Yes	Yes	Yes	Yes	Yes	TL,PCL,Detectors
AIMSUM	Yes	Yes	Yes	Yes	Yes	Yes	Yes*	TL, Detectors
MITSIM	Yes	Yes	Yes	No	No	No	No	TL,IL
SUMO	Yes	Yes	Yes	Yes	Yes	No	No	TL,IL
MAS-T2er Lab	Yes	Yes	No	No	No	No	No	TL
ITSUMO	Yes	No	No	No	No	No	No	TL, Detectors

* It is possible with a plugin

LABELS	
TL	Traffic Light
PCL	Pedestrian Crossing Light
IL	Induction Loop
Detector	IL, Cameras, others

In Table 1, we can see that commercial simulators have various types of entities. MAS-T2er Lab and ITSUMO are small projects and have a small set of entities. SUMO and MITSIM implements cars, buses, and trucks to emulate basic traffic situations. All of them, except MAS-T2er Lab, have sensors to gather information and signalling to act in the system.

MAS provide great potential of application on transportation systems, for simulation can be agent-oriented. Nevertheless, MAS is not widely used, because of the increase on system complexity and the needed computational enforcement to simulate agents. Related to this feature, VISSIM, PARAMICS e AIMSUN are not agent-oriented, but have enough extensibility to support it. Also, MITSIM cannot work with agent, due to the lack of extensions and SUMO needs a intermediary layer to work with agents. On the other hand, MAS-T2er Lab and ITSUMO are agent-oriented, providing multi-connection and local information.

All simulators presented here were microscopic, just AIMSUN is also mesoscopic. Thus, all have 2D visualization and MITSIM and SUMO do not have 3D visualization. VISSIM, PARAMICS and AIMSUM have realistic 3D that is great for real scenario presentation. The Graphical User Interface (GUI) is always a different process from the simulator, but can be seen as a part of the simulator core.

Each simulator has advanced features, for instance, VISSIM has parameters and function flexibility, PARAMICS adapts to use all the distributed machine resource available, AIMSUN provides different forms to extend it, MITSIM has various types of controllers available for use, SUMO architecture flexibility, and MAS-T2er Lab and ITSUMO are originally agent-oriented.

Further, generally commercial simulators have more functionality than open source ones and do not exist a complete agent-oriented simulator. Most of simulators focus on traffic management, but nowadays are moving its focus to MAS paradigm. From this initial study, AIMSUN seems to be the best to work given our requirements, however a commercial license is required to use, even so the implementation from scratch the hardest and best choice.

6 Related Works

Two main areas were involved in this work: traffic simulation and agent-based simulators, being them expensive areas and cite some of the most important works in the field adds great value to our comparison.

Even though is out of the paper's scope, worth to mention SMARTEST final report [4] where, it presents the maybe most complete review of microscopic simulation models. An important contribution of the study is the identification of the gaps between model capabilities, Intelligent transportation system (ITS) modelling, high quality performance, execution speed among the others, and users requirements. The authors reached the conclusion that a good microscopic simulation model should provide capabilities not only dealing with common traffic/mobility problems, but also modelling various ITS features as well.

A more recent comparative study of microscopic and macroscopic traffic simulators is found in Ratrout et al [12] were the authors reviews the features of various, traditionally used, traffic simulators. Comparative analyses between the simulation software packages are presented as well, in this work. In Boxill et al [13] an evaluation is made about which simulators are suitable for real world ITS applications. Yet, in [14] a comparative evaluation on the car following model was done in a number of traditional microscopic simulators, namely VISSIM, PARAMICS and AIMSUN, measuring the performance of the car-following behaviour implemented in each simulator.

Xiao, et al. [15] has proposed a methodology framework for selecting a microscopic simulator. The comparative study was conducted using the AIMSUN and VISSIM simulation software packages. In [16] is presented a comparison of various microscopic traffic simulators. Here, the features of the software taken into account

are the ability to simulate large networks, creating traffic networks and the associated vehicles patterns, CPU and memory performance among the others.

As we point out in the previous paragraph FUT, expressed mostly as traffic and transportation systems, are made up of many autonomous and intelligent entities.
The agent or better Multi-agent metaphor is well suited to represent systems whose entities exhibit autonomy and some degree of interaction with one another.

Burmeister et al [17] presents the fact that traffic and transportation domain has the characteristics that agent-oriented techniques are aimed to solve: traffic and transportation is highly dynamic environments and much more flexibility is requested than what can the traditional systems can provide.

A number of agent-based traffic and transportation tools have been reported in literature. Fischer, et al. [18] propose the AGENDA/MARS testbed where different MAS cooperation methods based on negotiation are developed in order to simulate solutions for scheduling problem in the transportation domain. Rossetti et al [19] present an extension to an existing microscopic simulation model DRACULA using BDI agents in an agent-based framework to assess drivers' decision making behaviour. Another open-source agent-based toolkit that can support large-scale transport simulation is the MATSim [20].

Fourie [21] compares the performance of MATSim with that of an established equilibrium assignment model (EMME/2), where MATSim agents manage to produce more realistic travel times than the traditional model, improving also the network utilisation avoiding congestion.

Panwai et al [22] presents a car-following model, which is developed using reactive agent techniques based on Artificial Neural Networks. To test the performance of the model the AIMSUN simulator has been used through the GETRAM extension. The evaluation carried out by testing the proposed model with three well-known car-following models in VISSIM, AIMSUN and PARAMICS. The result shown that the agent-based model performs better than the traditional ones.

Zhang et al [23] proposes a multi-agent framework for single-lane traffic simulation. In this work, the major entities are modelled as agents such as traffic-lights agents, driver-vehicle agents etc. Finally, a recent work [24] couples together, reviewing applications of agent technology in traffic and transportation systems.

7 Conclusions and Future Work

Traffic systems have been focus of much improvement and commuters have in general witnessed a revolution in the way a trip is planned and actually carried out in urban networks. Computer technologies as well as communication capabilities have put intelligent transportation solutions in the same level as users, as they now feature some degree of autonomy and intelligence as well, sometimes even deliberating with end users the best alternative to improve system's performance. In such a contemporary scenario, ITS-based solutions and users are peers and present a rather social interaction, which brings about new performance measures that must be assessed somehow. Furthermore, as user is central now, traditional traffic simulation

packages fail to model and represent all aspects of human behaviour in a detailed way.

In this work we carry out a first attempt at evaluating current available simulation environments an their ability to model and simulate future urban transport. We have started by idealising a transport system featuring all desired characteristics FUT, where not only performance is essential but also the user entity is a key aspect playing an imperative whole in all social interactions taking place in such a complex domain. Basically, we must take into consideration that current transportation systems are now able to explicitly interact with end users, allowing them to be rather greener, accessible, cheaper, more efficient (both in terms of resource consumption and performance), and environment friendly. Moreover, privacy and safety are other important issues that must be addressed at first hand.

Having identified those characteristics, we have devised a taxonomy that was used to assess currently available traffic simulation packages. Our taxonomy includes criteria such as extension capabilities, computational processing approach (parallel/distributed), entities simulated, agent-orientation, simulation approach, and visualisation capabilities. As for the assessment carried out, we can conclude that with respect to the proposed taxonomy, the item *extension* was difficult to define for each simulator because it demanded much user's knowledge, so deeply depends on who is analysing. Furthermore, *parallelism/distribution* and *simulated entities* require some work due to the lack in the tool's documentation (in the case of *parallelism/distribution* some tests needed to be realized). Finally, all items left were easy to define and compare, especially to define the integration level between GUI and simulator's core in *visualization* item.

In general, most simulators follow a microscopic approach as an attempt to improve the representation of human behaviour. However, a very few arguably implement the concept of agents, although some authors claim their representation are based on the agent metaphor. Even so, entities present a very basic behaviour, being only able to perform car-following and lane-changing interactions. As for deliberation and other more cognitive characteristics of the decision-process performed by humans, they are basically ignored in most packages. Nonetheless, extensibility in some tools are quite promising, allowing the user, with good programming skills to implement the desired features to support FUT assessment.

There are basically two extensions that could follow up the present work. Firstly, we intend to increase the number of analysed simulators, i.e., we have restricted the number of tools based on the microscopic modelling approach. However, different types of abstraction level will be needed in the FUT simulation platform. Also, as seen in related works there are many platforms that are claimed to be agent-oriented from scratch, meaning they have been devised and implemented with the agent metaphor in mind from the very beginning of their conceptualisation. Thus, our taxonomy will certainly need be adjusted to contemplate specific characteristic of agents, such as which social behaviours they implement, which level of cognition agents are able to perform (are they just reactive or are they cognitive?) and so forth.

Another important aspect to have in mind is the communication abilities of entities, so that new standards being currently applied in transportation can be tested

and evaluated accordingly. As for the simulated entities, this aspect must be improved in order to allow us to consider new users' devices and intelligent infrastructures. Finally, after having a proper and complete appraisal of those features, the very next step is to devise and specify an artificial transportation platform on the basis of the agent-oriented paradigm, which we believe is the right way to support FUT's modelling and assessment in all levels.

References

1. Passos L., Rossetti, R.J F.: Intelligent Transportation Systems: a Ubiquitous Perspective. 14th Portuguese Conference on Artificial Intelligence, Aveiro, pp.27-38 (2009)
2. Li, J., Tang, S., Wang, X., Wang, F.: A software architecture for artificial transportation systems - principles and framework, Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE, pp.229-234 (2007)
3. Figueiredo, M.C., Rossetti, R.J. F., Braga, R.A.M., Reis, L.P.: An approach to simulate autonomous vehicles in urban traffic scenarios. Proc. 12th International IEEE Conference on Intelligent Transportation Systems, pp. 322-327 (2009)
4. Algers, S., Bernauer, E., Boero, M., Breheret, L., Di Taranto, C., Dougherty, M., Fox, K., Gabard, J.: Smartest: Review of micro-simulation models. Technical report, The SMARTEST Project, Leeds University (1998)
5. Fellendorf, M.: VISSIM: A microscopic simulation tool to evaluate actuated signal control including bus priority. Proc. 64th ITE Annual Meeting (1994)
6. Bertini, R.L., Lindgren, R., Tantiyanugulchai, S.: Application of paramics simulation: at a diamond interchange. Technical report, Oregon Department of Transportation, Portland (2002)
7. TSS-Transport Simulation Systems. Aimsun microscopic traffic simulator: A tool for the analysis and assessment of its systems (2005). Available in http://www.aimsun.com/aimsun-overview-hccmeeting.pdf.
8. Ben-Akiva, M., Cortes, M., Davol, A., Koutsopoulos, H., Toledo, T.: Mitsimlab: Enhancements and applications for urban networks. Technical report, Intelligent Transportation Systems Program, Cambridge (2001)
9. Krajzewicz, D., Hertkorn, G., Rossel, C., Wagner, P.: Sumo (simulation of urban mobility): An open-source traffic simulation. Proc. 4th Middle East Symposium on Simulation and Modelling, pp. 183-187 (2002)
10. Ferreira, P.A.F.: Specification and implementation of an artificial transport system, Master Thesis, Faculty of Engineering, University of Porto (2008)
11. Silva, B.C., Bazzan, A., Andriotti, G.K., Lopes, F., Oliveira, D.: Itsumo: An intelligent transportation system for urban mobility. LNCS Springer, no. 3473, pp 224-235 (2006)
12. Ratrout, N.T., Rahman, S.M.: A comparative analysis of currently used microscopic and macroscopic traffic simulation software. The Arabian Journal for Science and Engineering, vol. 34, no. 1B, pp. 121-133 (2009)
13. Boxill, S. A., Yu, L.: An evaluation of traffic simulation models for supporting ITS development. Technical, Transportation Training and Research, Texas Southern University, USA (2000)

14. Panwai, S., Dia, H.: Comparative evaluation of microscopic car-following behavior. IEEE Transactions on Intelligent Transportation Systems vol. 6, no. 3, pp. 314-325 (2005)
15. Xiao, H., Ambadipudi, R., Hourdakis, J., Michalopoulos, P.: Methodology for selecting microscopic simulators: Comparative evaluation of AIMSUN and VISSIM. Research, Center for Transportatin Studies, University of Minnesota (2005)
16. kotuseyski, G., Hawick, K.A.: A review of traffic simulation software. Technical, Computer Science, Institute of Information and Mathematical Sciences, Massey University, Albany, Aukland, New Zealand (2009)
17. Burmeister, B., Haddadi, A., Matylis, G.: Application of multi-agent systems in traffic and transportation. IEE Proc-Software Engineering vol. 144, no. 1, pp. 51-60 (1997)
18. Fischer, K., Chaib-draa, B., Muller, P.J, Pischel, M., Gerber, C.: A Simulation Approach Based on Negotiation and Cooperation Between Agents: A Case Study. IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews vol. 29, no. 4, pp. 531-545 (1999)
19. Rossetti, R.J.F., Bampi, S., Liu, R., Van Vliet, D.: An agent-based framework for the assessment of drivers' decision making. IEEE Intelligent transportation Systems. Dearborn, MI, pp. 387-392 (2000)
20. MATSim. Retrieved from http://matsim.org
21. Fourie, P.J.: Agent-based transport simulation versus equilibrium assignment for private vehicle traffic in Gautenggauteng. 29th Southern African Transport Conference. Pretoria, (2010)
22. Panwai, S., Dia, H.: A reactive agent-based neural network car-following model. IEEE Conference on Intelligent Transportation Systems. Vienna, Austria (2005)
23. Zhang, F., Li, J., Zhao, Q.: Single-lane traffic simulation with multi-agent system. IEEE Conference on Intelligent Transportation Systems, pp. 56-60 (2005)
24. Chen, B., Cheng, H.H: A review of the applications of agent technology in traffic and transportation systems. Trans. Intell. Transport. Sys., vol. 11(2), pp. 485-497 (2010)

SESSION 3

ARTIFICIAL INTELLIGENCE

Chairman: Ali Azarian

Crowd Simulation Modeling Applied to Emergency and Evacuation Simulations using Multi-Agent Systems

João E. Almeida[1,2], Rosaldo Rosseti[1,2], António Leça Coelho[3]

[1]LIACC – Laboratório de Inteligência Artificial e Ciência de Computadores
[2]FEUP – Faculdade de Engenharia da Universidade do Porto
[3]LNEC – Laboratório Nacional de Engenharia Civil
Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal
joao.almeida@engenheiros.pt

Abstract. In recent years crowd modeling has become increasingly important both in the computer games industry and in emergency simulation. This paper discusses some aspects of what has been accomplished in this field, from social sciences to the computer implementation of modeling and simulation. Problem overview is described including some of the most common techniques used. Multi-Agent Systems is stated as the preferred approach for emergency evacuation simulations. A framework is proposed based on the work of Fangqin and Aizhu with extensions to include some BDI aspects. Future work includes expansion of the model's features and implementation of a prototype for validation of the propose methodology.

Keywords: Crowd Simulation, Modeling, Evacuation, Emergency Planning, Multi-Agent Systems, MAS.

1 Introduction

Crowd and group simulations are becoming increasingly important in the computer games industry and in emergency simulation. Applications range from the entertainment to more serious use like pedestrian behavior in the real world or in panic situations. This paper summarizes a synthesis of what has been done in recent years in this field, discussing the various aspects involved, from social sciences to the computer implementation of modeling and simulation using Multi-Agent Systems. A framework is proposed based on the work of Fangqin and Aizhu with extensions to include some BDI aspects. Future work includes expansion of the model's features and implementation of a prototype for validation of the propose methodology.

1.1 Crowd behavior

Studying crowd behavior in emergency situations is difficult since it often requires exposing real people to the actual, possibly dangerous, environment. Fire drills (fig.1) are a possible approach but hardly recreating the truly panic conditions, people tend to

take it not seriously. A good computational tool that takes into consideration the human and social behavior of a crowd could serve as a viable alternative.

Fig. 1. During a fire drill at "Casa da Musica"[1] Oporto.

Computer models for emergency and evacuation situations have been developed and most research into panics has been of empirical nature and carried out by researchers from social sciences [1],[3],[5],[6].

1.2 Normal pedestrian behavior

Pedestrian crowds have been empirically studied for the past decades [1],[2]. The evaluation methods applied were based on direct observation, photographs, and time-lapse films. Apart from behavioral investigations, the main goal of these studies was to develop computer animated realistic applications, for the game industry, design elements of pedestrian facilities, or planning guidelines for architectural building and urban design.

In their common environment pedestrians tend to show some basic attributes. For example people always try to find the shortest and easiest way to reach their destination. If possible they avoid detours, even if the shortest way is crowded. The basic principle is the "least effort principle", which means everyone tries to reach their goal as fast as possible spending the least amount of energy and time.

1.3 Individual and crowd behavior in emergency and panic situations

Most of the normal behavior vanishes when pedestrians face an emergency situation (it does not always have to be an emergency situation, similar effects can be observed for example in crowds trying to get the best seats at a concert or consumers running for sales). Observations made for pedestrian crowds in emergency situations feature typically the same patterns. As people try to leave the building as fast as possible, the desired velocity increases which leads to some characteristic formations. As nervousness increases there is less concern about comfort zone and finding the most convenient and shortest way.

[1] Photos taken by the author in 2008

It is observable, for example, that if people have to leave a building in an emergency situation and they don't know the structure of the building well enough, they would run for the exit they used as an entrance, even if other exits might be easier to reach or even safer.[1]

They also might lose the ability to orient themselves in their surrounding and thus show herding or flocking behavior [3]. Not only do they lose certain abilities, they also start to exhibit new behaviors like pushing or other physical interactions. Nonadaptive crowd behaviors are recognized to be responsible for the death and injury of most victims in crowd disasters. Nonadaptive crowd behaviors refer to the destructive actions that a crowd may experience in emergency situations, such as stampede, pushing, knocking, and trampling on others.

1.4 Herding or Flocking

Herding tries to describe a human group dynamics visible in emergency situations (fig.2). When people get nervous and feel panic, they lose the ability to act logically and to decide on their own. As a result of this lack of independence, people tend to follow others in the assumption they could get them out of the dangerous area. On one side this could actually help people to escape faster, but if for example smoke is reducing the visibility or the person leading the group does not know the structure of the building well enough, it could also reduce the chance to find an exit. So instead of people wandering around on their own, more and more flocks of people start to form with increasing anxiety or nervousness. As simulations have shown none of the extremes (people walking around on their own or as a single large group) results in optimal evacuation time [1],[5].

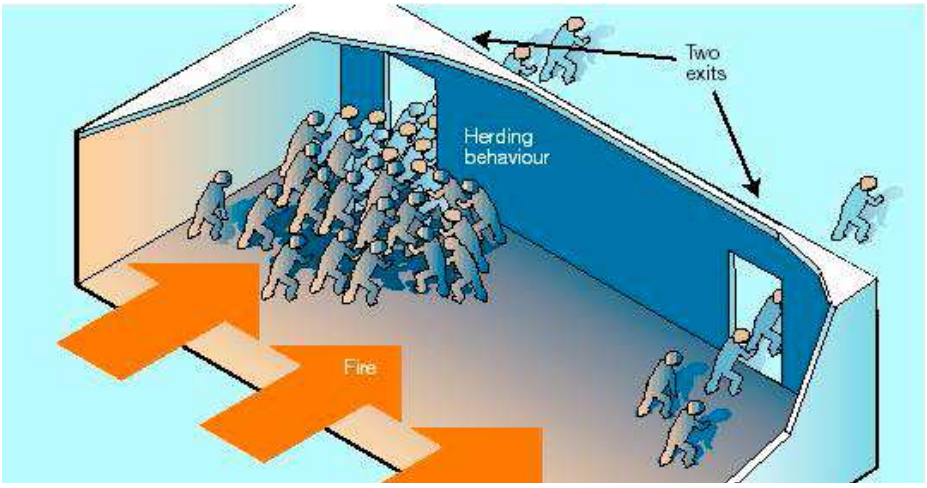


Fig. 2. Crowd trying to escape from smoke-filled room [1].

1.5 Arching and Clogging

Observations have shown a phenomenon called arching, which appears when a big crowd with a high desired velocity tries to pass through a door. Instead of passing through the door in less time, or giving the oncoming pedestrians a chance to pass through the door, the door gets clogged and the crowd gets arch-shaped (fig.3).

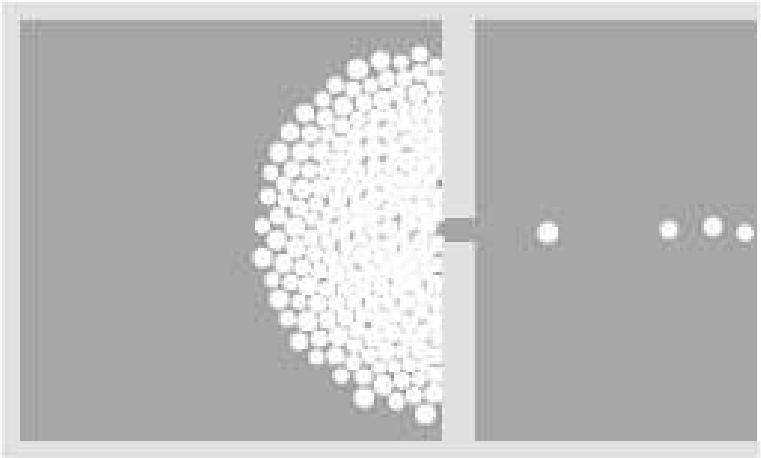


Fig. 3. Arching and clogging [1].

2 Related work

There are three main reasons for developing computer simulation for crowd behaviors: first to test scientific theories and hypotheses; second, to test design strategies; third, to create phenomena about which to theorize [8]. A full understanding of crowd behaviors would require exposing real people to the specific environment for obtaining empirical data, which is difficult since such environments are often dangerous in nature. In addition to studying crowd behavior based on observations and historical records, computer simulation is a useful alternative that can provide valuable information to evaluate a design, to help the planning process, and for dealing with emergencies.

Human behaviors are complex emergent phenomena, which are difficult to capture into computers as mathematical equations. There are several techniques to model crowds. Existing models can be categorized into one of the following groups.

2.1 Flow-based modeling

Flow-based models use the density of nodes in continuous flows. The basic principle is the analogy with fluid and particle motions. Often are called *macroscopic models*. Characteristics are defined beforehand thus all particles behave in the same way.

In this kind of models the simulated physical environment is defined as a network of nodes. The nodes represent physical structures, such as rooms, stairs, lobbies, and hallways that are all connected and comprise a single structure from which an evacuation is executed. The nodes contain people. Certain nodes are designated as destination nodes identifying the possible exits. For each node, the usable area must be calculated and allowance is made for the presence of closets, equipment, and other such items, as well as the space which persons place between themselves and a wall. Besides nodes, the model also requires the provision of specification for arcs. Arcs are passageways between building components with two variables: traversal time or the amount of time it takes to cross the passageway, and an arc flow capacity which delimits the amount of human occupants that can cross the passageway per unit time. One example of this type of modelling is EVACNET4 [4],[18].

2.2 Cellular Automata

In this kind of modeling space is discretized. A matrix is created plotting areas in a two-dimensional array. The simulation technique uses a time-frame pre-defined in which the occupants can move from one position or node to another, assuming it is free or it is not an obstacle. Each element can have several values: empty, occupied by a person, occupied by some object, or part of the limits (wall). The movement occurs at every step of the defined time-frame when occupants can move to one of the adjacent nodes. Each person can only move to an empty node, and directions are limited to the eight possible nearby nodes. Microscopic and macroscopic analyses are both permitted.

This type of model is simple to implement but fails when trying to replicate the erratic movement of people in real life, since only limited movement is allowed. Also it is not easy to model different speeds and interaction between people, due to the grid shape of space. Nevertheless this is the most used type for crowd modeling in games and more serious applications. One popular example is Exodus in early versions [4]. Another example is EGRESS [8].

2.3 Agent-based

Multi-Agent Systems (MAS) approach to this problem is probably the most realistic solution since it allows to model each individual person with their own unique characteristics, but related with all surrounding persons, thus recreating the real world interactions among human beings.

SIMULEX [4] was the first application to use MAS. Exodus latest's versions and PedGO also use MAS [9].

3 Multi-Agent Systems in Evacuation Simulators

In recent years MAS has been used as the preferred method to simulate crowd movement in different scenarios [7],[8],[9],[11],[16]. The enormous complexity of agent modeling, the need of data and rules to feed the system and the computational time needed (although according to Moore's Law computers' processing power keeps increasing) have created some difficulties to this approach.

However, investigation is going on and new papers describing work in this field are becoming more and more common. The possibilities offered by MAS are immense, as long as social rules and interaction knowledge among people is known and fed to the model. Social knowledge from researchers of other fields besides modeling and computational areas are welcome.

3.1 MAS Model

The model must be as complete as possible with all variables supplied to the virtual environment and then made available to the agents.

Human individuals are modeled as autonomous agents who interact with a virtual environment and other agents according to the individual's characteristics (which may vary from person to person) using global rules derived from the world where the system is created. Each agent has a limited vision of the world. Depending on the environment and the behavioral levels of individuals and their relationships with the group (or the crowd), the agent could interact and react in a collaborative or competitive manner. In contrast to agent-based systems for design applications, there is no global system control in the simulation model. In fact, the objective here is to be able to observe the random dynamics among the individuals (agents) in the simulation environment. To simulate human cognitive processes agents continuously sense and assess the surrounding environment making decisions based on their own decision model. The crowd social behaviors are collectively observed as emergent phenomena.

3.2 BDI Agents

MAS can use different levels of complexity and implement social-like behavior, using the BDI technique (*Beliefs*, *Desires*, *Intentions*) where agents are driven by *Desires* (the goals), according to certain *Beliefs* (set of knowledge of the world) and *Intentions* (actions) to fulfill the *Desires*. For instance, in an emergency evacuation simulation, agents' *Desires* are to leave the place where they are, due to fire or other hazard, as quickly as possible, using the fastest and safest path (following the *Beliefs*) and taking the necessary actions (*Intentions*).

Social forces such as comfort zone, pushing and fighting for space, should also be modeled and interactions between agents tested and validated.

BDI agents will implement more complex decision making processes and interaction among them can help scientists find new relations and derive modeling mathematical rules to understand crowd behavior in normal and emergency situations. This would help designers to build safer buildings, planners to prepare better

emergency plans and educators to find the best strategies for emergency plans. Although much work has been done, much more is needed to achieve realistic results.

4 A Framework for Crowd Simulation in Emergency Situations

In order to implement a good framework for crowd simulation in emergency situations, the steps to follow are:

- select the best methodologies to implement;
- use the best modeling data available;
- use of MAS for crowd modeling;
- create an open-source framework ready to accept new add-ons;
- use modular development, OO languages and off-the-shelf technology;
- allow easy inclusion of newer algorithms.

4.1 Model structure

The model structure proposed by Fangqin and Aizhu (fig.4) is a good starting point. This model proposes the use of readymade and available software thus saving much developing time.

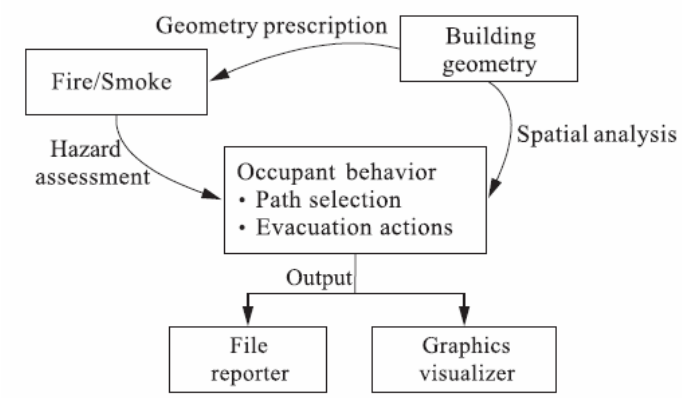


Fig. 4. Model flowchart [7].

Building geometry and *Fire/Smoke* models are based on existing and tested software, like FDS [19],[20] and PyroSim, a commercial software package, for 3D CAD building design [21]. Data interchange can be done using file systems and batch processing, since computational time needed is high and real-time visualization can only happen after all calculations are complete.

Fire Dynamics Simulator (FDS) is a Computer Fluid Dynamics (CFD) software developed by NIST[2] for fire hazard assessment, freely available for scientific use[19],[20]. FDS uses geometry data based on a database structure shared by PyroSim [21].

The *Occupant behavior* module implements crowd modeling based on geometry from PyroSim and hazard data from FDS. Its main objective is to define the exit path for each occupant, considering all interactions between the agents and environment.

4.2 Agent's attributes

For the agent modeling, the attributes are described bellow (as shown on fig. 5):

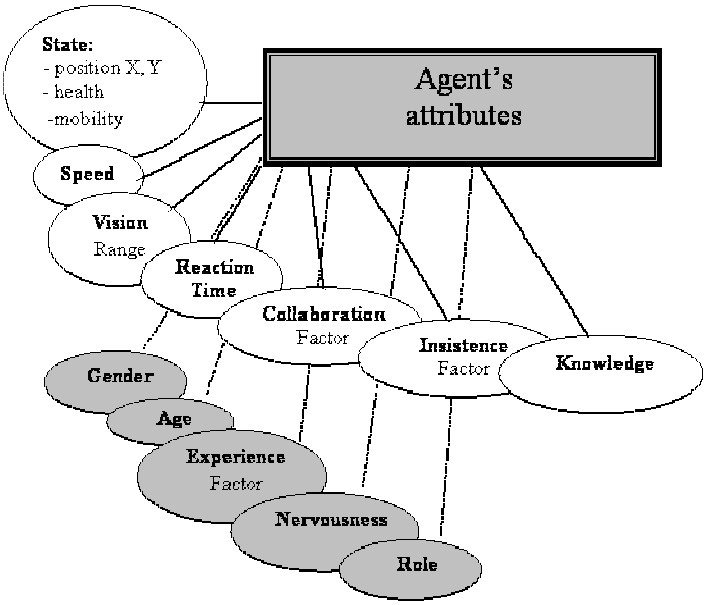


Fig. 5. Agent's attributes as proposed by Fangqin and Aizhu [7] and extensions (shaded items).

State. Physical data of the agent: current position (X,Y), being X and Y the coordinates of the discrete spatial location in the virtual space; health condition [0-1] real interval showing if the agent is alive (1), injured or ill (0 < health condition < 1) or dead (0); mobility (0-can't move; 1-normal behavior; 2-panic behavior).
Speed. Agent speed velocity (m/s) varies with health condition and mobility, ranging from 0 (when health=0 or mobility=0) to 7 m/s (when running in panic).

[2] NIST: National Institute of Standards and Technology, USA

Vision. Vision is the visible range from current location. Depends on health condition and OD[3] given by FDS. Also must be able to detect obstacles and other agents.
Reaction Time. Reaction Time (RT) is the time an occupant spends to decide what action should take. Typically this time ranges from few seconds to some minutes [6]. This attribute will be in seconds.
Collaboration. Some degree of cooperation among occupants is typically refers in all studies concerning this area [5],[6],[12],[13],[14]. The attributes here will reflect the path selection algorithm that will be implemented.
Insistence. The insistence factor defined on the interval [0,1] indicates the probability of maintaining the current evacuation strategy. When an agent is experiencing low evacuation efficiency, the attribute decreases and leads to strategy adjustments.This attribute will be used to adjust the path selection algorithm.
Knowledge. Represents the degree of familiarity of the occupant with the building. This factor varies with the knowledge of the surroundings and will increase when the agent gets more acquainted with the space.

Other attributes to implement a BDI architecture are proposed in this paper to expand the model possibilities:

Gender. Many studies refer differences between men and women reactions in panic or stress situations [5],[6],[12],[13].
Age. Another important factor determining behavior according referred studies.
Experience. Previous experience in exercises or real situations are proved to be important in the decision making process [14]. In this attribute factors such as (1) knowledge to use fire extinguishers (2) participation in fire drills (3) previous experience in emergency situations, should be taken into account.
Nervousness. Factor indicating the degree of nervousness or anxiety, of the agent when facing emergency situations.
Role. The agent role can be set as a coefficient related to the importance in evacuation scenarios. Can be a hierarchical status (director/chief/simple employee or teacher/student) indicating the importance of his/her decisions and influence for the surrounding agents. This attribute can be used to implement the leader-follower model some researchers propose with dynamic grouping [10]. Integer (0: none; 1: top level; 2: 2nd level; … n: nth level).

4.3 Occupant behavior module

This model implements the interactions between agents and the environment. Hazard information is received from the FDS module with all variables related to temperature, smoke, pressure, toxicity of air and visibility, in each of the compartments or spaces in the scenario. Geometry is supplied from the same database used by FDS and designed using PyroSim.

For the *occupants behavior* MAS will be used. The complexity of the decision making process will depend upon agent's attributes and environment data supplied

[3] Optical Density: unit 1/m, measures the visibility in smoke condition

the *Building geometry* and *Fire/Smoke* models. Rules can vary from simple reaction action to more complex BDI with interaction between agents.

The actions agent's will take into account the following aspects:

1) analysis of environment conditions (alarm, temperature, presence of smoke, etc.) and determine the need of evacuation (this will give Reaction Time);
2) observation of other agent's behavior and eventually follow their actions (depends on agent's level of hierarchy);
3) visibility conditions, knowledge of the environment, surrounding exits;
4) knowledge of the building and nearest path to safe place or exit;
5) presence of obstacles or other people clogging exits;
6) physical conditions;
7) social forces when crowd is forcing to pass through a clogged exit.

Within this context, the decision making algorithm should provide: (1) an exit path (2) adapt route whenever conditions change (3) inform other agent's of actions taken.

To improve the model, works related with social forces and human interaction should be used, like the recent studies of Moussaïd et al [17] where the decision process for pedestrians and behavior rules are mathematically modeled based on empirical observations.

4.4 Output module

The occupant behavior module actions will be the input for this module. These actions can be either saved in file, for later processing, or directly sent to graphic display using 3D software like OpenGL.

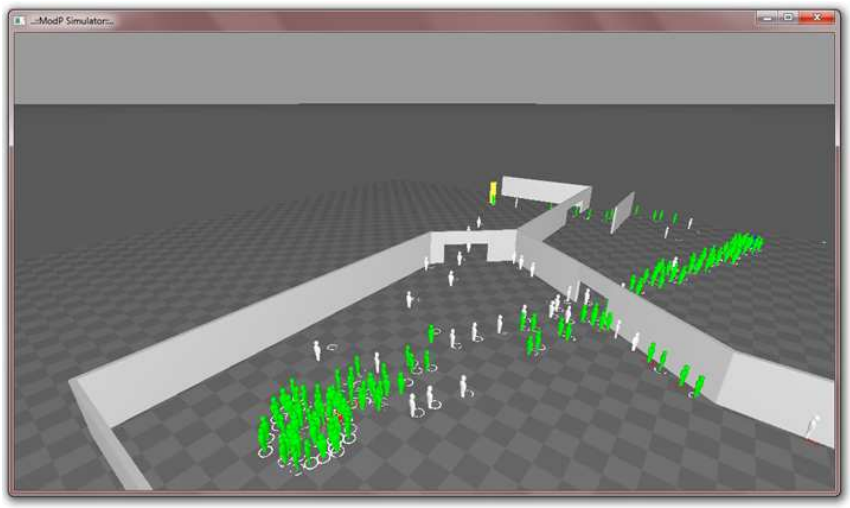


Fig. 6. Modp 3D viewer [22],[23].

Features to implement related to the animation, could be: selecting the camera position; detailing regarding the human representation; background textures; visualization of smoke and flames, etc. Exodus shows a good approach to realistic images.

One possibility would be using ModP [22],[23] for 3D pedestrian visualization module (fig.6).

5 Conclusions and future work

This paper introduces the need of crowd simulation for computer games or more serious applications like emergency evacuation. Problem overview is described including some of the most common techniques used. Multi-Agent Systems approach is stated as the preferred technique for emergency evacuation simulations. A framework for crowd simulation in emergency situations is proposed based on the work of Fangqin and Aizhu [7] with some extensions to include some BDI agent's aspects such as (1) sensors for the real world, (2) social forces and (3) interaction with other agents.

Future work includes expansion of the model's features and implementation of a prototype for validation of the propose methodology.

References

1. Helbing, D., Farkas, I., Molnar, P., Vicsek, T.: Simulating of Pedestrian Crowds in Normal and Evacuation Situations. In M.Schreckenberg, S.D. Sharma(ed.) Pedestrian and Evacuation Dynamics. Springer Verlag Berlin and Heidelberg, pp. 21-58, (2001)
2. Helbing, D., Molnar, P., Farkas, I., Bolay, K.: Self organizing pedestrian movement, in Environment and Planning B: Planning & Design (2001)
3. Reynolds, C. W.: Flocks, Herds, and Schools: A Distributed Behavioral Model. Proceedings of SIGGRAPH '87, Computer Graphics, 21(4), pages 25-34, July (1987)
4. Santos, G., Aguirre, B.E.: A Critical Review of Emergency Evacuation Simulation Models. NIST Workshop on Building Occupant Movement during Fire Emergencies June 9-10. National Institute of Standards and Technology, U.S. Department of Commerce (2004)
5. Coelho, A.L.: Modelação de Evacuação de Edifícios Sujeitos à Acção de um Incêndio (in Portuguese). Ph.D. Dissertation, LNEC, Lisboa (1997)
6. Cordeiro, E.: A Influência do Comportamento das Pessoas e suas Limitações na Evacuação dos Edifícios (in Portuguese), LNEC, Lisboa (2009)
7. Fangqin, T., Aizhu, R.: Agent-Based Evacuation Model Incorporating Fire Scene and Building Geometry. Tsinghua Science and Technology ISSN 1007-0214 21/25 708-714 Volume 13, Number 5, October (2008)
8. Pan, X., Han, C.S., Dauber, K., Law, K.H.: A multi-agent based framework for the simulation of human and social behaviors during emergency evacuations. AI & Society Volume 22, Number 2, 113-132, DOI: 10.1007/s00146-007-0126-1 (2007)
9. Jafari, M., Bakhadyrov, I., Maher, A.: Technological Advances in Evacuation Planning and Emergency Management: Current State of the Art. Report nº. EVAC-RU4474. Center for Advanced Infrastructure & Transportation (CAIT) Civil & Environmental Engineering

Rutgers, The State University Piscataway, NJ 08854-8014. U.S. Department of Transportation Research and Special Programs Administration. March (2003)
10. Qingge, J., Can G.: Simulating Crowd Evacuation with a Leader-Follower Model. IJCSES International Journal of Computer Sciences and Engineering Systems, Vol.1, No.4, October (2007)
11. Cherif, F., Djedi, N.: A Framework to Simulate the Evacuation of a Crowd in Emergency Situations. Georgian Electronic Scientific Journal: Computer Science and Telecommunications 2006 | No.1(8) (2006)
12. Kuligowski, E.D.: The Evaluation of a Performance-Based Design Process for a Hotel Building: The Comparison of Two Egress Models. Master of Science, Dissertation. Faculty of the Graduate School of the University of Maryland, College Park (2003)
13. Kuligowski, E.D.: Modeling Human Behavior during Building Fires. NIST Technical Note 1619. National Institute of Standards and Technology, U.S. Department of Commerce (2008)
14. Kuligowski, E.D.: The Process of Human Behavior in Fires. NIST Technical Note 1632. National Institute of Standards and Technology, U.S. Department of Commerce (2009)
15. Levin, B.C., Kuligowski, E.D.: Toxicology of Fire and Smoke. CRC Press (Taylor and Francis Group), Boca Raton, FL, (2005)
16. Murakami, Y., Minami, K., Kawasoe, T., Ishida, T.: Multi-agent simulation for crisis management Proceedings of the IEEE Workshop on Knowledge Media Networking (KMN'02)
17. Moussaïd, M., Helbing, D., Garnier, S., Johansson, A., Combe, M., Theraulaz, G.:Experimental study of the behavioural mechanisms underlying self-organization in human crowds. Proceedings of The Royal Society Biological Sciences, 2755-2672 (2009)
18. Kisko, T. M., Francis, R. L., Nobel, C. R.: EVACNET4 USER'S GUIDE. University of Florida (1998)
19. Fire Dynamics Simulator (Version 5) Technical Reference Guide Volume 1: Mathematical Model. NIST Special Publication 1018-5. National Institute of Standards and Technology, U.S. Department of Commerce (2008)
20. Fire Dynamics Simulator (Version 5) User's Guide. NIST Special Publication 1018-5. National Institute of Standards and Technology, U.S. Department of Commerce (2008)
21. PyroSim User Manual. Thunderbird Engineering 403 Poyntz Ave. Suite B Manhattan, KS 66502-6081 785-770-8511, USA (2010)
22. Esteves, E.F.: Utilização de agents autónomos na simulação pedonal em interfaces multi-modais (in Portuguese), Master Dissertation, Engineering Faculty of Porto University, Porto, 2009
23. Aguiar, F.H.M.: Crowd Simulation Applied to Emergency and Evacuation Situations, Master Dissertation, Engineering Faculty of Porto University, Porto, 2010

Implementation of Autonomous Robotic Cooperative Exploration and Goal Navigation

Nuno Saleiro[1]

[1] Laboratório de Inteligência Artificial e de Ciência de Computadores
Departamento de Informática
Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias s/n,
4200-465 Porto, Portugal
nuno.saleiro@fe.up.pt

Abstract. This paper approaches a set of techniques that allows a robot team to search effectively for a position (called beacon), whose location is unknown inside a scenario containing obstacles and barriers. Implementation was done in a multi layer perspective, allowing a decision algorithm to select the right action to make facing the data received from the sensors. Sensorial information was stored to provide the robot with a global perspective of the world. Robots were allowed to change their strategies and possible goals of exploration. It was expected that the interoperation of the software layers would allow the robot to find a goal (beacon) quicker, either via the individual agent intelligence or via the coordinated shared environment exploration. Results were measured and collected of the performance from different perspectives: in the implementation architecture (pure-reactive v.s. subsumption), in the A* Algorithm plan effectiveness and finally in the cooperation gain. The final application was able to achieve success on intelligently searching/mapping of the environment and better likelihood of finding the goal position in a shorter time.

Keywords: Robotic Autonomous Navigation, Robotic Cooperation, AGV, A* Algorithm, Mapping, Planning and Control, Path Planning, Cell Decomposition, Ciber-Rato Simulator Contest, Robotic Navigation, Robot Maze Solving, Group Exploration, Group Search.

1 Introduction

Standard environments where humans operate, like offices, industries, and residences, are filled with objects, obstacles and architectonic barriers that exist between the subject current location and its desired target. Humans perform locomotion in a autonomous perspective, they are able to select where they are going, change their plans in the middle of the road, avoid new unexpected obstacles efficiently, process vision, perceive possible obstacles from a far distance ahead and to do plans using spatial representations from their site experiences stored in their memory.

Robots however, are still far from possessing the complete human versatility. Autonomous Mobile Robots or Autonomous Guided Vehicles to act intelligently

without human guidance, have to mimic same processes as humans perform less consciously. When the operational environment is unknown, the autonomous robot should build an internal representation of its surroundings. External information, collected trough local spectrum sensors is unable to provide instant complete information of the global state of the world. As such, mapping requires timed framed information on obstacles, goal proximity hints and other terrain relevant information is integrated and stored, providing the robotic software an overview of the situation [1].

Ciber-Rato is a simulator platform, who mimics a homogeneous real robot competition, being its purpose to promote the development of autonomous software control agents. The official Ciber-Rato competition is won by the robot that finds quicker a target position marked by the presence of a beacon device that emits a tracking signal within a low range. Once the robot is next to the beacon it must mark its discovery by raising the finish flag. The simulation field contains the presence of high and low walls, which respectively block and allow the beacon signal to pass trough. The robot basically has to go from a start position to an unknown end position, using 3 frontal Infra-Red Sensors and a collision detection ring to sense.
The simulator platform supports academic research on agent's architecture that supports localization, mapping and sensorial fusion. Ciber-Rato simulator is very similar to the Robot Fire Fighter Competition, and other problem solving competitions that have known educational properties [2].

The goal position discovery procedure is the main process, which is affected by not only internal (decision based), but external factors: the environment configuration, the individual robot decision process and the cooperation/antagonism actions resulting from other moving objects (mostly other intelligent robots).

The main objective of each robot is to place itself in the goal position:

$$position(r, s_t^{local}) = goal \tag{1}$$

It would be desired that the robot would have a global perspective of the world that surrounds it however the robot perceives local situations in discrete time events. Bellow problem formalization is presented:

$$p(r_{robot\_number}, s^{world}) = \sum (p_{local} + p_{local\_other\_robots} - p_{other\_robots\_locations}) \tag{2}$$

The ideal would be to know the world by considering all local perceptions on the several states of the world, obtained in different timings from either the local robot (p_{local}) or the other cooperative robots ($p_{local\_other\_robots}$) disregarding perceptions obtained and caused by the presence of another cooperative robot in a local analyzed state ($p_{other\_robots\_locations}$).

Local data gathered in P_{local} is a collection of data acquired in different states of time (s_t^{local}) by the local robot (r_{robot_number}):

$$p_{local}(r_{robot\_number}, s^{world}) = \left[\sum_{local=1}^{max\_l} \sum_{t=0}^{max\_t} p(r_{robot\_number}, s_t^{local}) \right] \tag{3}$$

Preferably the purpose in exploration would be to minimize max_t , or the time that we use to discover the maze and to maximize max_l, to know all the possible locations. Other robots presence in the simulator is the only mutable factor $\forall t,t2\, p(s_t^{local}) \neq p(s_{t2}^{local}), t \neq t2$, this situation is solved by considering the most recent state as the most important and believable state. In the normal case, however since we are in a maze with certain properties like enclosed (inaccessible spaces), makes it unfeasible to navigate on it sequentially, or to acquire perceptions of all the local states ($\exists s^{local}, p(s^{local}) = \varnothing$), so it might be impossible to have completeness on $p(s^{world})$. The robot cannot be in same place where there is an obstacle $\forall t1,t2, \text{obstacle}_{t1}^{local} \rightarrow \neg \text{robot}_{t2}^{local}$ or a different robot $\exists robot1, robot2 \forall t1,t2 robot1_{t1}^{local} \rightarrow \neg \text{robot2}_{t2}^{local}, robot1 \neq robot2$, and that collisions that have penalties associated should be minimized:

$$\forall local_1, local_2, \min \sum_{t}^{max\_t} (\text{distance}(\text{obstacle}_t^{local_1}, \text{robot}_t^{local_2}) = 0) \quad (4)$$

In the Ciber-Rato simulator, moving the robot from a position to another with precision must take into account several variables, being that those low level operations have to execute faithfully the decisions done at the higher levels of decisions. Lower level decision layers deal autonomously with operations issues like avoiding eminent dangers.

The exploration problem is complemented with mapping an unknown map, coverage of the space to locate the target, path planning to the goals and navigation to the target. Physical navigation to a location must be an efficient process promoting the selection of the shortest path to a goal [3]. In our work we assume that there is a precise GPS available and no errors can occur in localization.

Previous work in the area consists of developing exploration techniques either to map the ground inside our homes [4] or to operate in unstructured environments for space exploration purposes [5]. Simmons et al. [6] refer to the creation of maps of the environment as a challenge in mobile robotics, and use multiple robots to the job. Other researchers [7] try to perform cooperative map integration from multiple robots.

The current document has the following structure: Chapter 2, starts with the description of the problem, the space mapping and cell decomposition technique that allows data to be transformed to feed into the Path Planning algorithm, followed by a description of the reactive architecture that is behind the reflexes of the robot (Chapter 3). Chapter 4 describes the algorithm used to perform path planning. Plans generated were then executed step by step in an execution approach shown in Chapter 5. The Communication capabilities are mentioned in Chapter 6, and a compilation of the Results is presented at Chapter 7, followed by the respective Conclusions displayed in Chapter 8.

2 Space Mapping and Cell Decomposition

The Ciber-Rato world consists of: a robot start position, walls and a goal that the robot must get to. As the robot travels trough the environment, it records measures from its IR sensors, after these measures reach a certain threshold, there is a very strong indication of the presence of an obstacle inside the angle covered by the sensors sweep area.

The mapping of the states of the world can be complemented by other detailed information coming from the sensors. The robot presence on a place can indicate that this place is open to be traveled upon and empty of obstacles, the radio presence of the beacon, the beacon and the unknown areas can be marked in the map (Fig. 1).

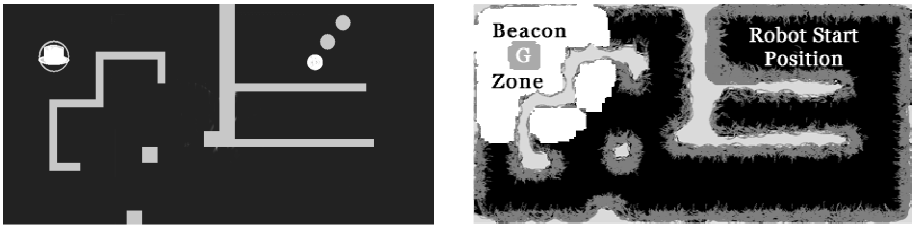


Fig. 1. a) A scenario b) Discovered internal representation of the scenario acquired by the robot. Goal position is referenced with the letter "G".

The thin grained information of the mapping process contains a large amount of information to be supplied into a path planning algorithm, whose inputs typically represent nodes in a graph. Real time decision implies scaling the size of the information to the processing algorithm. Many pixels would give less valid solutions, some plans go trough a narrow place thinner then the body of the robot, others could come across too much tightly near the walls promoting collisions. A variable sized grid was implemented, since the Ciber-Rato competition rules states that all the angles are 90 degrees and the minimum size of an opening has the size of the robot, we could allow a grid size smaller then the robot size (Fig. 2), however we considered to be more practical in terms of cooperation and overheads (to locate the robot accurately) to adjust the grid to the size of the robot.

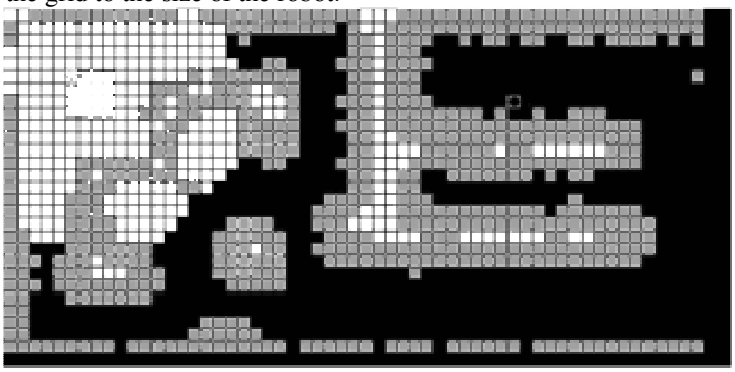


Fig. 2. A quadrant generated image of the maze

We generate an occupancy grid[8], which is created from a process similar to a rude coverage mapping. The grid states assigned are dynamic according to a percentage of thin grained markers measured (Table 1). Blocks that were traveled in the past are marked to remain as a viable option. in the future, even if their cost is set higher then normal blocks.

Table 1. **Minimum parts of sensor measured particles to qualify the quadrant.**

Quadrant Qualifier	Thin Grain Percentage Threshold
Unknown	80%
Free Space	70%
Near Beacon Zone	40%
Obstacle	8%
Beacon	*3%*

3 Reactive Subsumption Architecture

Humans try to analyze behavior as a sign of intelligence. We used the proven Brooks Subsumption Architecture [9], as a way to introduce a layered behavioral approach whose goal was to balance and make decisions on the right behavior to be taken (Fig. 3). By using it we could promote certain behaviors in adequate situations and mitigate the same in non suitable conditions.

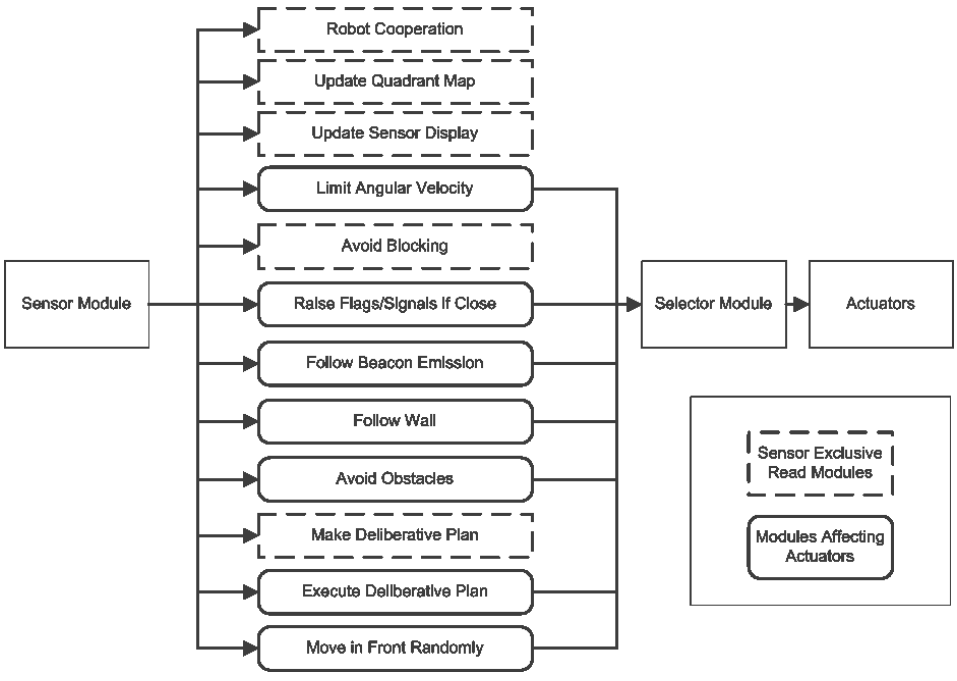


Fig. 3. Adopted Subsumption Reactive Hierarchy of Modules, (on conflict *top* modules have priority over the ones at the *bottom*).

There are several objectives or behaviors that must/or can be done like: avoiding that the robot's "*rotation velocity*" increases beyond control, go to the beacon (if located), avoid bumping into obstacles, explore parts of the maze, follow walls, and display data found from the robot to the user for monitoring purposes. Subsumption requires that those behaviors are prioritized so that some of them can override others. The hierarchy gives priority to act on the modules that are on top, over the ones that are bellow, the priority (inhibition) is given at the *selector* module. *Modules that affect sensors* act based on the freshest information coming from the sensors, other modules (related to map building) process all events, because they may catch some recorded information whose content is useful.

To assure a quick response reflex response the robot must process immediately the sensor data in fast algorithms to make its important decisions, and less to reason about past events. In these cases the selected algorithm is the main factor for achieving a good decision to do exploration and to find the hidden beacon.

4 Deliberative Path Planning

The path chosen should be the quickest route considering known obstacles and ignoring unknown contents of the search space. The A* algorithm is a path plan algorithm that uses as inputs: several nodes, a starting position, an end position and the routes among nodes. The algorithm uses the formula F=G+H to be able to pick the next node to be traveled. G measures the amount of space traveled from the starting point of the maze to the current position, and H is an estimation function whose result is a heuristic function containing the expected distance to the end position. H is in general the Euclidean distance, but other distances have been used due to performance gains. The most popular distance is the Manhattan distance, which represents an aspect under which the traveled path on several interactions is equal to the sum of the sides of a square where one vertex is on the origin and the opposite vertex is on the destination.

The implementation the A* algorithm was based on an algorithm produced by James Macgill [10], and was adapted to meet the possibility of diagonal traveling in the grid. Diagonals have a cost of $\sqrt{2}$, coming from the fact that distance traveled is equal to the hypotenuse of a triangle with unitary legs size. We obtained a graphical representation of the different possibilities on picking a quadrant and the cost of that quadrant (Fig. 4).

To avoid multiple recalculations of the A* a new path was only calculated when the robot moved to a different quadrant, or when the robot sensed that there was an obstacle in contact with the robot or near its external sensors. An unexpected obstacle in the precalculated path, is an example of a situation where a new plan to go to the target is needed.

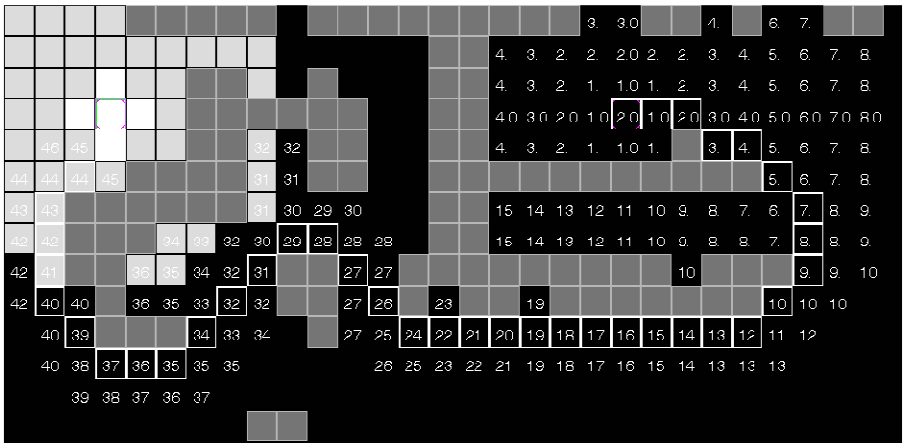


Fig. 4. An application of the A* using the Manhattan distance Heuristic. Planning is denoted with white pronounced squares. Quadrants contain the distance from the start point.

The angle that the robot needs to rotate is calculated by the expression in (5).

$$angle_{toTarget}=atan((Target_Y-Source_Y)/(Target_X-Source_X)) . \quad (5)$$

The result needs to be fitted to the exact angle interval that goes outside the]-Π/2, Π/2[scope. Since the screen coordinates grow in the opposite way of the trigonometric triangle ($Target_Y$-$Source_Y$) sign should be inverted.

4.1 Heuristic A*

The Heuristic A* algorithm uses the Manhattan distance as H function heuristic (6).

$$Distance_{Source,Target}= weight \times (Abs(Target_Y-Source_Y)+ Abs(Target_X-Source_X)) \quad (6)$$

The algorithm has two lists. One is called the open list, which contains the nodes that will be traveled in the future. The closed list is the one that contains the nodes already considered.

The algorithm starts with the start node in the open list, which is analyzed and all its adjacent nodes are added to the open list to be analyzed in the next steps in a cascading procedure. As the nodes are traveled, they are added to the closed list, so that they aren't again reconsidered.

4.2 One Tail A*

The One Tail A* Algorithm doesn't use a heuristic initially to establish the next participant in the solution. First it travels to all the adjacent nodes in a chain process, adding the distance traveled from the start to the current node. Since almost all the nodes are traveled, the exploration is heavier then the other approaches. After it has

established that it is possible to achieve the final node, the algorithm travels the path backwards from the end to the start, always choosing the neighbor that has a less cost, using the precalculated values.

4.3 Fudge

Amit Patel's Fudge [11] is a different heuristic function, with a search algorithm similar to the Heuristic A*. It attempts to cope with the fact that the A* algorithm (in open scenarios) produces too many similar paths with the exact same length, and usually only one is necessary. To prevent ties between similar solutions, the weight of H is slightly increased in the F=G+H formula. The algorithm, whose search selects the minimum F, behaves in broadness search near the start point and changes to depth search as it explores closer to the end point.

The Fudge Heuristic (7) gives preference to paths that tend to search in a straight line towards the goal. This algorithm takes in account the distance between two points and the distance from one of those points to the start point.

$$Distance_{S,T} = ((Cx-Tx)*(Sy-Ty))-((Sx-Tx)*(Cy-Ty)) \quad (7)$$

The computed cross vectorial product between 3 points (the Start, the Current location and the Target destination), measure vectorial alignment. If the vectors aren't aligned the resulting difference is larger, what makes this formula prone to obtain solutions that operate on a straight line. The heuristic however shows inefficiency on obstacle courses, doing over exploration before the obstacle.

5 Execution of the Path Planning

Path Execution was implemented by programming the effectors, to rotate and move to the first position provided by the A* plan. If due to an unexpected situation the robot travels to a middle plan position the plan is automatically trimmed to start from that position. The A* plan considers that diagonals can be taken very near corner walls, but that decision is executed poorly in course grain maps, resulting in collisions against the corner. Special cases were solved by an execution of the plan in a different manner, promoting only vertical or horizontal motion.

6 Communications and Strategies

Robots communicate and share their current location to other robots in the group. Team members add knowledge that they acquire from other teammates in their internal memory representations and also avoid creating plans that may collide with other robots. More important robots share their intentions on the destination that they wish to arrive. The exploration destination information is used for the current robot to know if it's necessary to ask to other robot to change its search strategy, and travel to

a different location instead. Livelocks, occurring from situations where two robots suggest at the same time a "*strategy change*", are minimized by random time deferral.

6.1 Selecting what to explore

Maze exploration in order to locate the beacon and be efficient should avoid going to places already visited in the past [12]. If the robots are doing exploration they tend to pick squares that have a lot of other hidden squares around them, this mitigates the behavior of the robots having to travel to a remote location just to uncover a single square. The seeking mechanism also promotes looking to unknown squares that have near beacon zones.

If however the main goal is available on the knowledge base, the attempts will be focused in travelling to that location, otherwise exploration is necessary.

6.2 Exploration Strategies

Three exploration strategies were implemented. The first "Explore Near" seeks the most promising quadrant that is nearest from the current robot position. The second one "Explore Far", explores a quadrant that is far from the robot position and "Explore Random" chooses a random quadrant from those unknown. The available quadrants are selected using a utility function that provides ranking to the available choices[13]. In most maps the starting position of the robots is very near, so they tend to pick the same quadrants to explore. Once the communication establishes that situation, one of the robots asks the other conflicting robots to change their search strategy, allowing for a wise shared exploration using the capabilities available.

7 Results

We analyzed a complex set of results, reactive subsumption architecture, the Path Planning A* Algorithm efficiency, and the Cooperation gain on the beacon localization.

7.1 Reactive Subsumption Architecture

The development of this architecture has led to a processing pipeline, whose sensorial data traverses trough the modules, until a decision is made upon what instructions should be sent to the actuators. It's important that this pipeline acts quickly, because the robot is limited to respond to the world at its internal speed.

In the pipeline, the instructions shouldn't be processed in order of arrival inside all the modules, because those would be sending out of late responses that are unsuitable to the current state of the world. This has led to the optimization of the pipeline: modules that decide upon the actuators don't process all the received sensorial information sequentially, but just the most recent information available. It was

possible then allowing the robot to act sooner to the current reality. The module that provided the actuators with instructions operated in a continuous loop, this forced an action to be given to the actuators from the decision making modules in a short time window and avoided situations that would imply modular synchronization.

Modules which process readings from the sensor module and produce instructions for the actuators were implemented as running threads. In theory this made all the modules run faster because their acting wasn't interdependent and complex processing modules didn't reduced the output of modules whose implementation was simple. Reasoning threads when not in use, released their CPU time to other threads, however the selector module wasn't granted idle CPU time due to the reasons mentioned in the previous paragraph.

7.2 A* Path Planning Heuristics Performance Analysis

The time taken by the robot to fully execute the path was measured under different Heuristics. The results we present are the average of 6 executions (Table 2).

Table 2. **Average time taken to reach the goal.**

Heuristic Type	Time to achieve the goal (sec)	Average Time to the execute (ms)
HeuristicAStar	129.4	18.09
OneTailedAStar	103.9	219.09
Fudge	128.9	16

Resulting algorithm execution times were as expected worse for the OneTailedAStar. Executing from the start position, took 13 times more to locate the goal then the other heuristics. The Fudge algorithm proved also in the same conditions and for this particular scenario, to be marginally the fastest to execute. Performance considerations are important when considering that these algorithms are summoned multiple times during the execution (Fig. 5).

The OneTailedAStar even if it's inefficient in the procedure of analyzing all the quadrants, it has taken statistically less time in the execution displacement (going from the start of the maze to the end position), this can indicate that the plan has slightly better quality (less collisions, avoids walls and conflicts of actions).

Achieving the goal in 73 sec. could be obtained using lighter CPU loads. Most of the algorithms prefer traveling in the diagonal when the path is shorter. The A* Star algorithm considers correctly that the shortest path can be in the diagonal, but it doesn't considers the time that the effectors will take to rotate the robot. Other algorithms were slower because they performed stair-case motion, on the linear diagonals (rotating in both directions alternatively), in those cases, traveling trough the Pythagorean triangle legs can be more efficient.

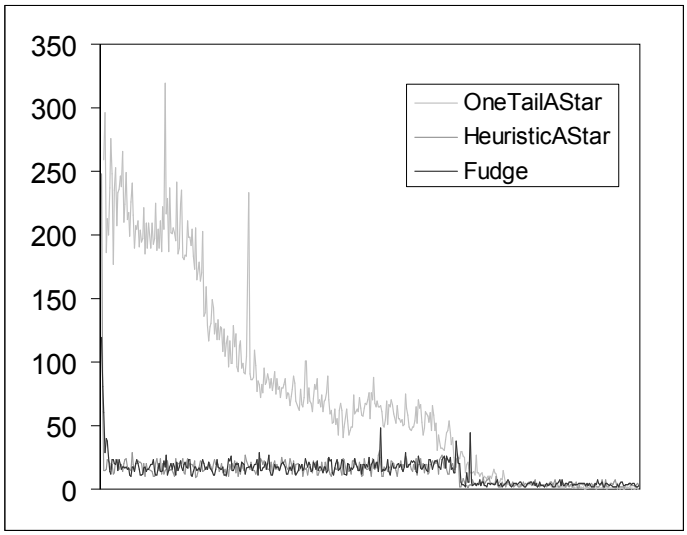


Fig. 5. Duration of the A* Heuristic Algorithm (in ms) in the beacon approach procedure

7.3 Cooperation

After doing several runs measuring the time of arrival of the first robot (with two robots) we measured average times of 336 sec., and 508 sec. for a team of 2 robots and a single robot respectively. It was expected that with the double of means to explore (robots) the time would be reduced by half. It was obtained a 44% of reduction, meaning that search cooperation could still be optimized. We assume that those losses occurred when a robot places itself on the same path as another, stalling its plan execution and also be caused by the lack of processor time of the robotic control processes.

8 Conclusions

The Pure-Reactive implementation was definitely faster then the Subsumption architecture, but the quality of the second, produced better solutions and allowed a complex code architecture and subsequent management of the same. Three path-finding heuristics where compared and their performance. The results weren't proportional and consistent with the plan execution times, we believe that's due to the A* plans consider only spatial distances and disregard penalizing turn operations as a temporal cost. Spending a lot of effort in planning generates better options (avoiding conflicts of being close to obstacles), those cost saved and were compensated by a faster arrival to the desired location. It's impressive to see how humans do navigation everyday in such a simple and intuitive manner. To allow our robots to perform part

of this required a lot of parallel processing, filter errors and to keep the plans with the current world state, up to date.

The software built, was adapted to be less dependent on the world sampling rate, allowing it to decide in a more historical perspective. It's very likely that some applications in the field seem to execute with more agility because their programming language allows them to extract samples of the world faster, but that doesn't mean that they would be able to be ported to other environments.

Optimizing processor flow, software architectures that are able to deal with indecision, studying path planning algorithms whose main factors are non-spatial would require further study. Developing algorithms to control a robot in a multi level architecture is possible and allows seeing how promising these technologies are in the promotion of an autonomous cooperation to achieve common goals in less time.

Acknowledgments. I would like to thanks Prof. Dr. Armando Sousa and Prof. Dr. Luis Paulo Reis and Prof. Dr. Eugénio Oliveira from the Faculdade de Engenharia Universidade do Porto, I also would like to thanks Prof. Dr. Artur Pereira and Prof. Dr. Nuno Lau from the Universidade de Aveiro for their suggestions.

References

1. Rahim, M., Nawi, I.: Path Planning Automated Guided Robot. In: Proceedings of the World Congress on Engineering and Computer Science 2008, San Francisco (2008)
2. Pack, D., Avanzato, R., Ahlgren, A.: Fire-Fighting Mobile Robotics and Interdisciplinary Design-Comparative Perspectives. Vol. 47 nr. 3 IEEE Transactions on Education, IEEE, 369--376 (2004)
3. Martinengo, A., Campani, M., Torre, V.: Complex Tasks and Control Strategies of Robots. IEEE 1050-4729, 861--866 (1994)
4. Forlizzi, J. : Service Robots in the Domestic Environment: A Study of the Roomba Vacuum in the Home, HRI '06 Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction (2006)
5. Gifford, C., Webb, R., Bley, J. et al. :Low-Cost Multi-Robot Exploration and Mapping, Proceedings of the IEEE International Conference on Technologies for Practical Robot Applications, 74--79 (2008)
6. Simmons, R., Apfelbaum, D., Burgard, W. et al.: Coordination for Multi-Robot Exploration and Mapping, American Association for Artificial Intelligence (2000)
7. Fox, D., Ko, J., Konolige, K. et al.: Distributed Multi-robot Exploration and Mapping, Proceedings of the IEEE 94(7), 1325--1339 (2006)
8. Yamauchi, B.:Frontier-based exploration using multiple robots, Proc. of the Second International Conference on Autonomous Agents, 47—53 (1998)
9. Brooks, R.A.: A robust layered control system for a mobile robot. In: IEEE I. Robotics and Automation 1, 14--26 (1986)
10. A* Demonstration, http://www.vision.ee.ethz.ch/~cvcourse/astar/AStar.html
11. A*'s Use of the Heuristic, http://theory.stanford.edu/~amitp/GameProgramming/Heuristics.html#S12
12. DaeEun, K.: Evolving Internal Memory for T-Maze Tasks in Noisy Environments, Global Grid Forum (2002)
13. Burgard, W., Moors, M. , Stachniss, C. et al. : Coordinated Multi-Robot Exploration, Vol. 21 issue 3, IEEE Transactions on Robotics, 376—386 (2005)

Humanoid Clock-Turning Gait Synthesis based on Fourier Series And Genetic Algorithms

Nima Shafii[1,2], Luís Paulo Reis[1,2], Nuno Lau[3,4]

[1]DEI/FEUP – Informatics Engineering Department, Faculty of Engineering of the University of Porto, Rua Dr. Roberto Frias s/n, 4200 465 Porto, Portugal
[2]LIACC – Artificial Intelligence and Computer Science Lab., Porto, Portugal
[3]UA – University of Aveiro, Campus Universitário de Santiago, 3810 193 Aveiro, Portugal
[4]IEETA – Institute of Electronics and Telematics Engineering of Aveiro, Portugal
nima.shafii@fe.up.pt, lpreis@fe.up.pt, nunolau@ua.pt

Abstract. In humanoids robots path planning, Turn-in-place or Clock-Turning is a basic motion. However, in biped locomotion studies, it can be considered as one of the most complicated tasks. It needs to use all joints movments of legs in three planes of the transverse(axial) ,frontal(lateral) and sagittal. This paper presents an approach based on fourier series to generate all joint's angular trajectories of the legs which are moving in these three planes. Our emphasis is how to make a robot to do "turn-in-place" motion more stably and faster. In this regard Genetic Algorithms is utilized to optimize produced trajectories by Fourier Series (FS). The effectiveness of the proposed method can be shown through simulation and experimental results.

Keywords: Bipedal Locomotion, Gait Generation, Turning, Gait Optimization.

1. Introduction

In recent years, planning of biped walking has been one of the fundamental issues in robot locomotion researches [1-5]. Planning of walking follows different goals, such as ; reduction of energy consumption [2], walking on stairs [3],walking on slopes [4] and rough surfaces. In future Humanoid robots are supposed to be able to operate in any environment that human operate [6].Therefore studies of robot locomotion can't be limited to robot walking and all other motions should stay in research domain regarding robot locomotion. Similar to walking, other motions like beside walking and turning are essential motions for humanoid robots. Humanoid robots sometimes must change their direction in their paths in order to adapt their environmental features and avoid obstacles. Kuffner [7] presented a footstep planning for humanoid walking, which can avoid obstacle by using Omni-directional walking. Their method is more like a method for obstacle avoidance in the field of mobile robots, where the balance and locomotion of the wheeled robots is not considered. Kajita et al. [8] proposed a simple method to control humanoid walking direction. When changing the walking direction, the robot's reference frame at a foot rotates to a specified angle

accordingly. However, their method was based on a simplified 3D inverted pendulum model, which overlooked substantial dynamics of robots and the constraints between the DOFs were not clear. In [9], a turning gait planning, in which the objective of turning is to follow a curve with certain radius is proposed. This is called forward-turning or curving the aim of clock-turning is to change robot's direction at a fixed position. These two kinds of turning are suitable in different situations, due to the fact that in forward-turning ,there are no switches between different motion gaits during the whole obstacle avoidance. In this case, robots requires more prediction and consideration of accurate position and shape of obstacles. But, clock-turning is able to combine turning and walking to adapt to different shapes of obstacls. Therefore, clock-turning has lower requirements for the humanoid sensor system. Although the clock-turning has been used as a basic motion for humanoid global path planning [10], there is only one previous work on it [11]. In 2007 Tang et al. tried to investigate turn motion by using the ZMP[1] indicators and inversed kinematics. Like other model based approaches, One of the main issues of the inversed kinematics method is generating a result pattern that has "bent-knee" posture, In order to avoid singularities problem when solving the inverse kinematics [12]. It also needs lots of computation and is not suitable for real time tasks. In contrast, Model-Free approaches do not have these kinds of issues.

In this paper a new model-free approach for turning-in-place is proposed which has an acceptable speed and is stable. In this method, Fourier series with genetic algorithms is used to control the biped robot for turn in place motion. This paper is organized as follow. In Section 2, a humanoid model is described. Section 3 describes how to model angular trajectory generator by using Fourier series. Optimization of model using genetic algorithm and results is presented in section 4. Finally, the conclusions and future works are presented.

2. Simulator and Robot Model

We test our method on a simulated humanoid robot. The robot model in this study is NAO robot and the simulation is performed by Rcssserver3d .we are using a generic three-dimensional simulator, which is based on Spark and Open Dynamics Engine (ODE). The simulator environment with robot Model is shown in "fig. 1". The robot model has 22 DOFs with a height of about 57cm, and a mass of 4.5kg.Our humanoid body like human moves in three plane of the transverse (axial) ,frontal(lateral) and sagittal [10] which these correspond to the yaw/heading, lean/roll and pitch degrees-of-freedom, respectively(Fig. 2).

Fig. 1. simulated Nao robot in rcssserver3D environment

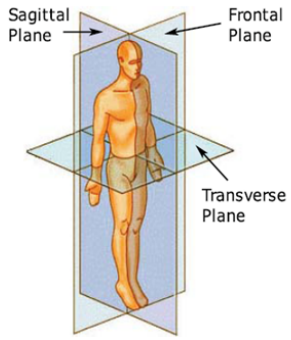


Fig. 2. The planes which human body moves on

In Fig. 3 Schematic view of our humanoid robot is shown.DOFs 2,3, and 4 move on Sagital plane and DOFs 1 and 5 move in Frontal plane and also DOFs 6 move on transverse plane.

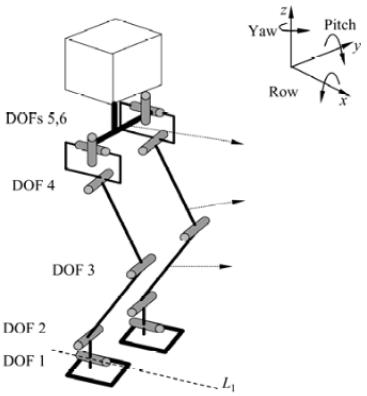


Fig. 3. Schematic of the lower body of the humanoid robot

3. Turn Angular Trajectories

We aim to generate a turning gait of a humanoid robot that looks like human one. Therefore analyzing human turn pattern is used for acquiring beneficial information about this motion. We can investigate human turn motion from many aspects; turning trajectory being one of them. The turning trajectory is divided into several types. Positional trajectory and angular trajectory are two of them. In angular trajectory, the angle of each joint is plotted at a certain time slice. Therefore, the angular trajectory is obtained by angular variation of each joint. Biped angular trajectory of two joints; hip and knee in sagittal, frontal and transverse captured from human turning are shown in Fig 4 and 5. [11]

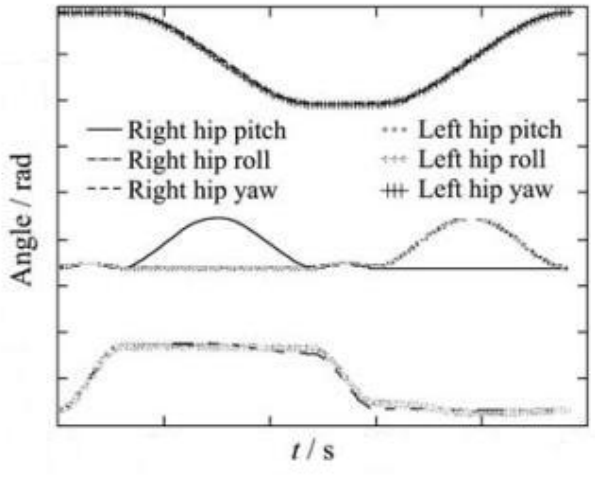


Fig. 4. Hip angle trajectories in sagittal,frontal and transerve planes[11]

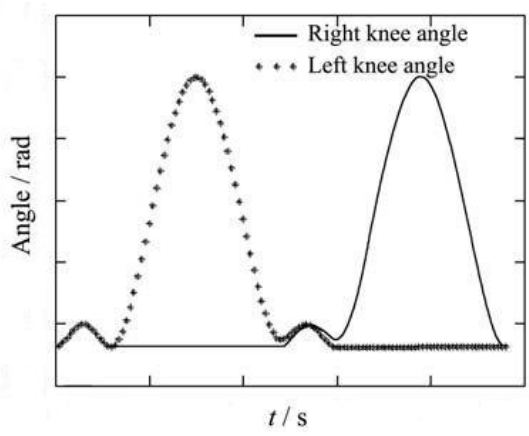


Fig. 5. Knee angle trajectories in sagittal plane[11]

To generate human like motion, we start from analyzing the angular trajectories results from the previous works, which can be found in [11],in order to achieve the shape of angular trajectory of Gaits. Since A gait is a cyclic, periodic motion of the joints of a legged robot, requiring the sequencing or coordination of the legs to obtain reliable locomotion. In other words, gait is the temporal and spatial relationship between all the moving parts of a legged robot. [13] Therefore all gaits angular trajectories are periodic and Fourier Series can be used to generate mentioned trajectories. Finally it is enough to use an optimization algorithm to find the optimum parameters of Fourier series to achieve a stable turning motion for humanoid robot.

There are six DOFs in each leg; two in the hip, two in the ankle and one at the knee. An additional DOF that exists at each leg's hip for yaw, causes the legs to rotate outward and inward and this moves the leg in transverse plane. In order to move legs on the frontal, sagittal Planes, It is found that all other 5 DOFs mentioned are involved in Turning.

In the following sections we will explain how to generate references trajectories for these DOFs in mentioned planes separatly to achieve a stable turn-in-place motion for humanoid robot.

4. Movements in Sagittal Plane

As mentioned in previous sections There are three DOFs in each leg which move on sagittal plane; one in the hip, one in the ankle and one at the knee. The model for generating angular trajectories in sagittal plane is very similar to Truncated Fourier series (TFS) model which was presented in 2009 [13]. In that work, similar to [14], foot in sagittal plane was kept parallel to the ground by using ankle joint. This is done in order to avoid collision. Therefore, ankle trajectory can be calculated by hip and knee trajectories and ankle DOF parameters are eliminated.

In that model, and also according to “Fig.4” and “Fig.5” each signal has an offset. In addition, the amount of shift phase of the two legs trajectories signal in this plane is half of the period of each signal so by producing trajectory of one leg the other leg's trajectory can be calculated. The trajectories for both legs are identical in shape but are shifted in time relative to each other by half of the turning period. The Fourier Series (FS) for generating hip and knee trajectories in sagittal plane, are formulated as below .

$$\Theta_{hx} = A_{hx} \,.\, sin(w_{hx}\, t) + C_{hx}\ ,\ W_{hx=} 2\pi/T_{hx}$$
$$If(\Theta_{hx} < C_{hx})$$
$$\Theta_{hx} = C_{hx} \qquad (1)$$

$$\Theta_{k} = A_{k} \,.\, sin(w_{k}\, t) + C_{k}\ ,\ W_{k=} W_{hx}$$
$$If(\Theta_{k} < C_{k})$$
$$\Theta_{k} = C_{k}$$

In these equations, A_{hx} and A_k are constant coefficients for generating signals. The h_x and k index stands for hip and knee in sagittal plane respectively. Also C_{hx} and C_k

are signal offsets and T_{hx} is assumed the period of hip trajectory. As it is illustrated in Fig. 4, all joints exceptt "hip yaw" joint have equal movement frequency and the frequency of FS for "hip yaw" joint is half of other joints. Therefore, the $W_{hx=} W_{k=} 2\pi/T_{hx}$ equation can be concluded. Fourier series parameters in sagittal plane are $A_{hx,}$ A_k , $C_{hx,}$ C_k and W_{hx} .

5. Movements in Frontal plane

There are two DOFs in each leg which move on Frontal plane; one in the hip and one in the ankle. Fig.7 illustrates the movement of leg's in frontal plane. θ is assumed as the maximum of legs movement. Again similar to movements in sagittal plan, Foot is kept parallel to the ground by using ankle joint in Frontal plane in order to avoid collision and with considering the fact that just one of each hip's joint moves each time, the angle of ankle is equal to the hip's angle of the opposite leg. So the trajectory of ankle in frontal plane can be generated by trajectory of hip angle in frontal plane.

According to Fig.5 to apply captured trajectory in our humanoid robot and generate it by FS we consider the trajectory of hip joint in frontal plane as is shown in Fig.6.

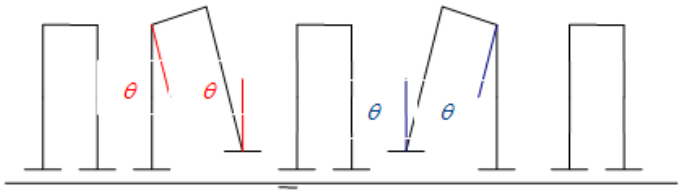


Fig. 6. Movement of legs in frontal plane

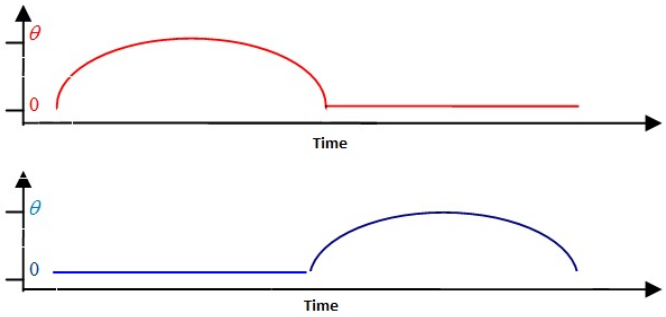


Fig. 7. Modified trajectory of left and right leg to use in our humanoid robot

Fig. 7 can be assumed as the hip angular trajectory in one period of Turning. It is a sinusoidal signal that has a lock phase at zero degree. The trajectories for both legs are identical in shape but are shifted in time relative to each other by half of the turning period. Therefore in order to produce proper angular trajectories to move the hips in frontal plane , proper parameters of following equation must be obtained(2).

$$\Theta_{hy} = A_{hy} . \sin(w_{hy} t) \quad , W_{hy} = W_{hx} = 2\pi/T_{hx}$$
$$If(\Theta_{hy} < 0) \qquad (2)$$
$$\Theta_{hy} = 0$$

In the above equation, the *hy* index stands for hip in frontal plane. A_{hy} and W_{hy} are assumed as amplitude and frequency parameter of signal respectively and T_{hy} is assumed the period of hip in frontal plane. As mentioned before, ankle trajectories can be calculated from hip trajectories .As mentioned, the period of the hip signal in frontal plane and the period in *sagittal* plane are equal. Therefore W_{hy} is eliminated in this method for producing the proper abduction and adduction, the proper value of A_{hy} parameter must be found.

6. Movements in transverse plane

There is one DOF in each leg which moves on tranverse plane; the one in the hip. According to Fig.4, the signal of this joint has an offset. Also the signal for this joint in both legs is the same so it is enough to achieve the parameters of one leg. The Fourier Series for generating hip trajectory in transverse plane are formulated as below (3):

$$\Theta_{hr} = A_{hr} . \sin(w_{hy} t) + C_{hr} \quad W_{hr} = W_{hx}/2 = \pi/T_{hx}$$
$$If(\Theta_{hr} < C_{hr}) \qquad (3)$$
$$\Theta_{hr} = C_{hr}$$

In these equations, A_{hr} is constant coefficient for generating signal. The *hr* index stands for hip in transverse plane. Also C_{hr} is signal offset. As it is mentioned, the movement frequency of FS for hip yaw joint is half of other joints, so the $W_{hr} = W_{hx}/2$ equation can be concluded. Therefore, W_{hr} can be achieved from $W_{hx,}$ and It is eliminated from parameters to be found . Consequently, A_{hr} and C_{hr} should be gained. So parameters of Fourier series to generate proper angular trajectories for turn-in-place motion are $A_{hx,}$ $A_{hy,}$ $A_{hr,}$ A_k , $C_{hx,}$ $C_{hr,}$ C_k ,W_{hx} And an optimization algorithms must optimize the 8 dimensions problem to find the best turn gate generator in this stage.

7. Genetic Algorithms For Optimizing Trajectories

Robot turning is a complex motion because many factors affect turning style and stability such as robot's Kinematics, collision between feet and the ground and dynamics of the robot. Therefore, for this complex motion, relation between turning gait trajectory and turning characteristic is nonlinear. Since this kind of optimization is usually difficult, a genetic algorithm is suitable to solve it. Genetic Algorithms (GA) is a stochastic searching procedure based on the mechanics of natural selection

and genetics [16]. GA is used to find the best parameters to generate angular trajectories for robot turning motion.

Using GA in our optimization problem, parameters are coded in to a finite length of string (Genes) as a chromosome. According to section 6, TFS has 8 parameters to generate all joints angular trajectories, thus each chromosome has 8 Genes. In Designing of Chromosomes gene's type is considered as double format. Population for each generation is assumed to be 100. Chromosomes are generated randomly and uniformly for the first iteration between lower and upper bound. In this study the lower and upper bound data for initialization are depicted in following table.

	C_{hx}	C_k	C_{hr}	$A_{hr,}$	A_k	A_{hy}	$A_{hx,}$	W_{hx}
Upper Bound	30	0	0	60	0	40	45	1
Lower Bound	-10	-50	-45	20	-60	0	0	0.05

TABLE 1. LOWER BOUND & UPPER BOUND

Fitness function has a critical role in GA and is used to judge whether a solution represented by a chromosome is good or bad. To learn how to turn, first angular trajectories based on each chromosome are produced by TFS, and then these angular trajectories are used by simulated robot for turning, to follow these angular trajectories all individual robot joints attempt to drive towards their target angles by using proportional derivative (PD) controllers. Finally, chromosome's fitness is calculated based on this turning.

To achieve more stable and faster turn, a fitness function based on robot's turn-in-place motion with having limited time for turning is assumed. The amount of deviation from the initial position *(dist)* is subtracted from the maximum of the degree which robot could turn, to force the robot to turn stable and in place. Fitness function is calculated when the robot falls or time duration for turning is finished. Equation 4 is assumed fitness function formulation, where the robot is initialized in x=y=0 (0,0) with aligning to the horizontal axis where *rotdegree=0* and time duration for turning is assumed as 60 s.

```
if (Test Time >= time duration for turning  or robot has fallen)
  Fitness = rotdegree - 90*dist
Fi
```

In the above equation the rot degree is the amount of turning at radian and dist is the amount of movement from initial place. For using GA as an optimizer, scattered function has been chosen as a cross over function. This function creates a random binary vector with the length of chromosome. Then the genes where the produced vector is a 1 from the first parent and the genes where the vector is a 0 from the second parent has been selected, then by combining these genes the child will be formed.

For mutation, uniform function has been used where after selecting a fraction of chromosome with same probability as mutation rate (which is assumed as 0.06), a random number from range of upper and lower bound (table I) has been selected uniformly then existing number in the fraction has been replaced by this random

number.

Selection method is roulette wheel and reproduction rate is assumed as 0.8. Termination condition is having a generation counter greater than 28. Therefore, the GA requires 2800 trials to find appropriate FS parameters.

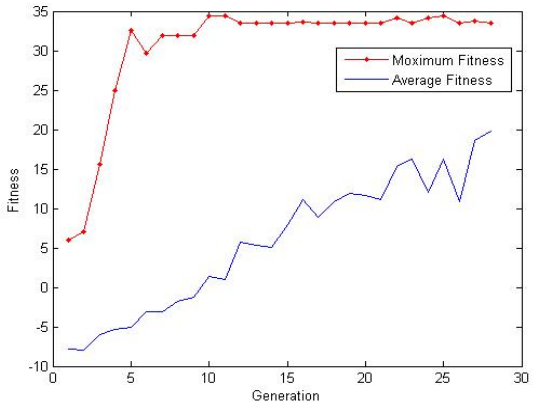


Fig. 8. Average fitness and Maximum fitness during 28 generations

9 hours after starting GA on a Pentium IV 3 GHz machine with 2 GB of physical memory, generation exceeded to 28 and the robot could turn in place 360 dgree in 6s with average body speed of around 60 degree/s. "Fig. 8" shows the average and maximum fitness values for the robot over 28 generations. However Elitism was employed, because physics simulation is not accurate and has noise, Maximum fitness was falling as well as rising from generation to generation. In fig. 9 and fig. 10 angular trajectory generated by TFS after learning process are shown. Followed trajectory is also shown in this figure.

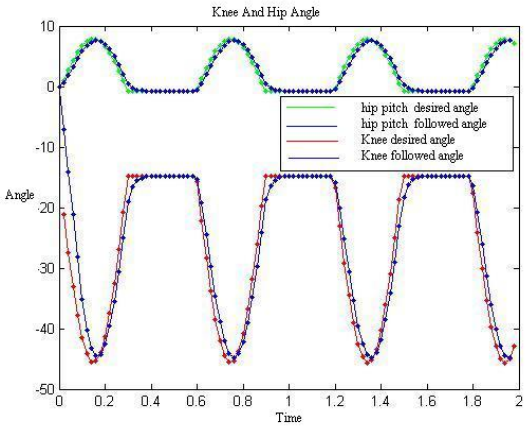


Fig. 9. Angular Trajectory generated by learned FS (Desired angle) and followed with controller

(Followed angle) for left hip and knee of simulated Nao robot.

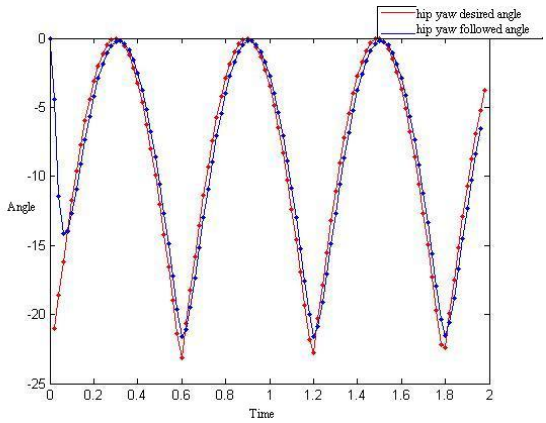


Fig. 10. Angular Trajectory generated by learned FS (Desired angle) and followed with controller (Followed angle) for left yaw hip of a simulated Nao robot.

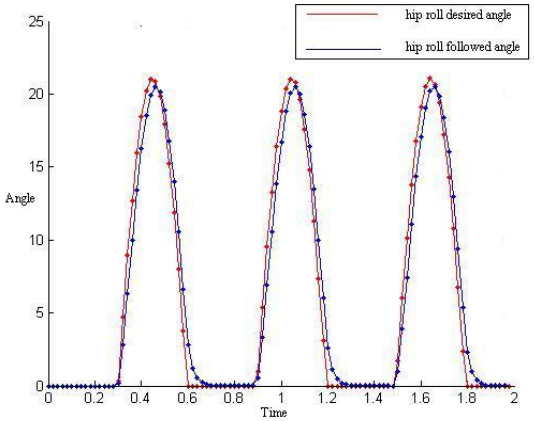


Fig. 11. Angular Trajectory generated by learned FS (Desired angle) and followed with controller (Followed angle) for left roll hip of a simulated Nao robot.

Genetics algorithm led the robot to learn how to turn in place; fig .11 exhibit biped locomotion is obtained from GA search.

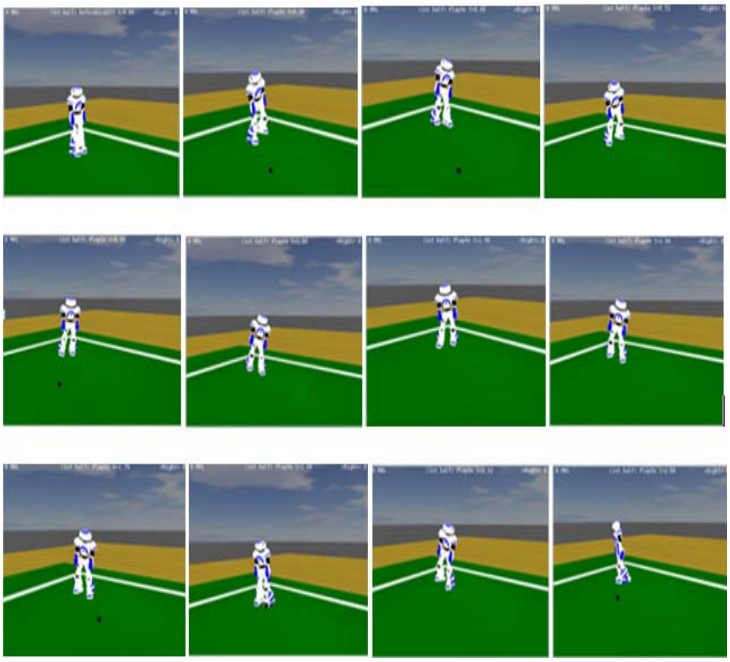


Fig.12. Learned Nominal trajectory is obtained by GA

Conclusion

Turn-in-place or clock turning is one of the principal and complicated motions in humanoids studies. For the first time, in this paper a model-free approach to generate turn-in-place motion was presented, which does not have the weaknesses of the previous model-based approaches. It is also capable of being implemented on any kinds of humanoid robots without considering its physical model. The gait generator was based on Fourier series and has 8 parameters for producing all joints angular trajectories. By using genetics algorithm, an optimum gait to produce the fastest possible turning was also found. In future works, it could be interesting to investigate the method on a real humanoid robots and try to reduce the parameters of the gait generator, since in real humanoids, optimization would be much harder and more time consuming.

Acknowledgements

The first author is supported by FCT under grant SFRH/BD/66597/2009. This work has been partially funded by FCT Project ACORD - Adaptative Coordination of Robotic Teams (PTDC/EIA/70695/2006).

References

1. Vukobratovic, M., Borovac, B., Surdilovic. D. :Zero-moment point proper interpretation and new applications. In: Proceedings of the 2nd IEEE-RAS International Conference on Humanoid Robots, pp. 237-244 (2001)
2. S. Kajita, T. Yamaura, A. Kobayashi. Dynamic walking control of a biped robot along a potential energy conserving orbit[J]. IEEE Transactions on Robotics and Automation, 1992, 8(4): 431 – 438.
3. C.-L Shih. Ascending and descending stairs for a biped robot[J]. IEEE Transactions on Systems, Man & Cybernetics – Part A: Systems & Humans, 1999, 29(3): 255 – 268.
4. Y. Zheng, J. Shen. Gait synthesis for the SD-2 biped robot to climb sloping surface[J]. IEEE Transactions on Robotics and Automation, 1990, 6(1): 86 – 96.
5. Q. Huang, K. Yokoi, S. Kajita, et al. Planning Walking Patterns for a Biped Robot[J]. IEEE Transactions on Robotics and Automation, 2001, 17(3): 280 – 289.
6. K. Hirai, M. Hirose, Y. Haikawa, et al. The development of Honda humanoid robot, Proceedings Of the IEEE International Conference on Robotics & Automation. New York: IEEE Press, 1998: 1321 – 1326.
7. J. Kuffner, S. Kagami, K. Nishiwaki, et al. Online footstep planning for humanoid robots, Proceedings of the IEEE International Conference on Robotics and Automation. Piscataway, New Jersey: IEEE Press, 2003: 932 – 937.
8. S. Kajita, F. Kanehiro, K. Kaneko, et.al. A realtime pattern generator for biped walking[C]//Proceedings of the IEEE International Conference on Robotics & Automation. Piscataway, New Jersey: IEEE Press, 2002: 31 – 37.
9. Z. Tang, C. Zhou, Y. Kong, et al. Turning gait planning for a humanoid robot[J]. Dynamics of Continuous, Discrete and Impulsive Systems, Series B: Applications and Algorithms, 2005, 12(1): 113 – 118.
10. J. Gutman, M. Fukuchi, M. Fujita. Real-time path planning for humanoid robot navigation[C] //Proceedings of the 19th International Joint Conference on Artificial Intelligence. San Francisco, California: Morgan Kaufmann Publishers, 2005: 1232 – 1238.
11. TANG. Z., SUN., Z. , LIU. H., Joo. M., Clock-turning gait synthesis for humanoid robots, Journal of Control Theory and Applications vol. 5 (1), pp. 23–27, 2007.
12. Nakanishi, J., Morimoto, J., Endo, G., Cheng, G., Schaal ,S., Kawato ,M.: Learning from Demonstration and Adaptation of Biped Locomotion with Dynamical Movement Primitives, Robotics and Autonomous Systems ,Vol. 47(2), pp. 79-91, 2004
13. Chen, L., Yang, P.,Liu, Z.: Gait optimization of biped robot based on mix-encoding genetic algorithm, Robotica (2009) volume 27, pp. 355–365, 2008.
14. Shafii, N., Javadi, M.H., Kimiaghalam B.: A Truncated Fourier Series with Genetic Algorithm for the control of Biped Locomotion. In: Proceeding of the 2009 IEEE/ASME International Conference on advanced intelligent Mechatronics, pp. 1781--1785, 2009.
15. Kagami, S., Mochimaru, M., Ehara, Y., Miyata, N., Nishiwaki, K., Kanade, T., Inoue, H.: Measurement and comparison of humanoid H7 walking with human being. Robotics and Autonomous Sys, vol. 48, pp. 177—187 (2003)
16. Shrivastava, M., Dutta, A., Saxena, A.: Trajectory Generation Using GA for an 8 DOF Biped Robot with Deformation at the Sole of the Foot, Journal of Intelligent Robot Systems, Vol. 49, pp. 67–84, 2007.

Estimating the Probability of Winning for Texas Hold'em Poker Agents

Luís Filipe Teófilo
Departamento de Engenharia Informática, Faculdade de Engenharia da Universidade do Porto, Portugal
Laboratório de Inteligência Artificial e de Ciência de Computadores, Universidade do Porto, Portugal
luis.teofilo@fe.up.pt

Abstract. The development of an autonomous agent that plays Poker at human level is a very difficult task since the agent has to deal with problems like the existence of hidden information, deception and risk management. To solve these problems, Poker agents use opponent modeling to predict the opponents next move and thereby determine its next action. In this paper are described several methods to measure the risk of playing a certain hand in a given round of the game. First, we discuss the game of poker and the expectation in its events. Next, several hand evaluation and classification techniques are described and compared in order to determine the advantages of each one. The current methods to deal with risk management can help the agent's decision. However, in future work, the integration of these methods with opponent modeling techniques should be improved.

Keywords: Poker, Opponent Modeling, Probabilities, Autonomous Agents, Strategy Games, Artificial Intelligence

1 Introduction

Poker is a game that is increasingly becoming a field of interest for the AI research community on the last decade. This game presents a different challenge when compared to other strategy games like chess or checkers. In these games, the two players are always aware of the full state of the game. Unlike that, the Poker's game state is hidden because each player can only see its cards or the community cards and, therefore, it's much more difficult to analyze. Poker is also stochastic game i.e. it admits the element of chance.

The combination of chance with incomplete information makes it essential to model opponents in order to create a competitive player. When the agents identifies the opponents' playing style it can adapt its strategy in order to beat them, by making decisions that have better probability of success [1, 2].

The goal of this work is to determine how an agent can measure the risk of its actions. This article particularly focuses on the game information that is visible to the player: its cards and the community cards. An agent, by knowing how much strong it is, can make better decisions, by managing the risk of betting.

The rest of the paper is organized as follows. Section 2 describes the game of Poker and more particularly the variant that was studied – Texas Hold'em.

Section 3 describes hand odds evaluators, that is, algorithms that are capable of estimating the probability of the player's hand being victorious. Section 4 describes hand rank evaluators which are functions that score a Poker hand which are essential to hand odds evaluators. Finally in section 5 presents the paper main conclusions and some pointers for future work.

2 Texas Hold'em Poker

Poker is a generic name for literally hundreds of games, but they all fall within a few interrelated types [3]. It is a card game in which players bet that their hand is stronger than the hands of their opponents. All bets go into the pot and at the end of the game, the player with the best hand wins. There is another way of winning the pot that is making other players forfeit and therefore being the last standing player.

2.1 Hand Ranking

A poker hand is a set of five cards that identifies the strength of a player in a game of poker. It is not necessary to have all the cards belonging to one player, because in poker there is the definition of community cards – cards that belongs to all players. In poker variations that use community cards, the player hand is the best possible hand combining his cards with community cards.

The possible hand ranks are (stronger ranks first): Royal Flush, Straight Flush, Four of a Kind, Full House, Straight, Three of a Kind, Two Pair, One Pair and Highest Card.

2.2 No Limit Texas Hold'em

No Limit Texas Hold'em is a Poker variation that uses community cards. At the beginning of every game, two cards are dealt for each player. A dealer player is assigned and marked with a dealer button. The dealer position rotates clockwise from game to game. After that, the two players to the left of dealer post the blind bets. The first player is called small blind, and the next one is called big blind. They respectively post half of minimum and the minimum bets. The player that starts the game is the one on the left of the big blind.

After configuring the table, the game begins. The game is composed by four rounds (PreFlop, Flop, Turn, River) of betting. In each round the player can execute one of the following actions: Bet, Call, Raise, Check, Fold or All-In.

In any game round, the last player standing wins the game and therefore the pot. If the River round finishes, the player that wins is the one with the highest ranked hand.

2.3 Opponent Modeling

Since Texas Hold'em is a game of incomplete information, it is essential to model the opponents in order to predict their actions. By predicting the opponents' actions, the player is able to optimize his profit.

One good example of opponent modeling is the Sklansky groups [3] which define types of players and common actions that they take for each group of cards.

3 Hand Odds Evaluation

Evaluating the odds of a hand consists in measuring the quality of the hand in a given round of the game. By evaluating the hand it is possible to determine the probability of winning or losing the current game. By using this knowledge, the agent can decide either to fold the hand or play it, based on the probability of success and the risk that the agent is willing to take.

The hand evaluation function may consider the following variables:

- Player cards;
- Number of opponents;
- Community cards;
- Possible community cards to come;
- Possible opponents cards;

The hand evaluation function returns a number in an interval, typically a real number between 0.0 and 1.0. If it returns the lower limit, this usually means that the hand will lose no matter what happens. Conversely, if it returns the upper limit, this means that victory is mathematically assured – the only way of losing is to unwisely fold the hand.

3.1 Current Hand Odds

The current hand odds algorithm calculates the probability of a player hand being better than the opponents' hands, considering that the opponents can have any remaining non visible cards and that they won't fold until the showdown.

For instance, consider the following game state:

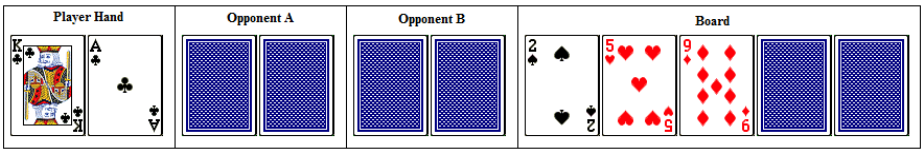


Fig. 1. Game state example with three players in Flop round

To calculate the player's current hand odds for this game state, it must be considered that "Opponent A" cards, "Opponent B" cards and the remaining invisible

board cards can be any cards other than **K♣**, **A♣**, **2♠**, **5♥** or **9♦**. For example, a valid set could be:

- Opponent A: 3♦, 3♥;
- Opponent B: 4♣, 5♣;
- Board: 2♠, 5♥, 9♦, Q♠, Q♣.

The algorithm can be represented by the following code:

```
function CurrentHandOdds (hand, board, numOpponents){
  const fullDeck = [Ac, Ad, As, Ah, Kc, Kd …]
  remainCards = fullDeck – hand – board//set operations
  ahead = behind = 0
  /* for each possible boards */
  for each case(completeBoard =
PossibleBoards(remainCards, board)) {
    handRank = Rank(hand, completeBoard)
    remainOppCards = fullDeck - completeBoard
    /* for each combination of possible cards
    for each case(opponents = Opponents(numOpponents,
remainOppCards)) {
      if (any rank(opponent in opponents,
completeBoard) > handRank) {
        behind++
      } else { ahead++ }
    }
  }
```

As it can be seen in the above code, all possible remaining card sequences are considered. For this reason, this process can be time consuming, given the large number of possible sequences of cards. For the given example (Fig. 1), the number of remaining cards is 47 (52-3-2) and the sequence size is 6. This means that the total number of card sequences is $47^{\underline{6}} = \frac{47!}{(47-6)!} \approx 7.73 \times 10^9$. This would take a lot of time to calculate.

The solution for this problem is to use Monte Carlo Method [4]. The Monte Carlo Method considers a small random subset of card sequences. This means, that the calculation is much faster. The only problem with this technique is that the obtained result is an approximation. However, this experience [5] showed that the error is very small, making this a very reliable solution.

3.2 Hand Strength

Hand strength[6, 7] is the probability that a given hand is better than any other possible hand. It consists in enumerating all possible hands that an opponent can have a checking if hand is better than the enumerated hand. By counting the number of times the player's hand is ahead, it is possible to measure the quality of the hand. It is possible to calculate the hand strength as:

```
function HandStrength(ourcards, boardcards){
  ahead = tied = behind = 0
  ourrank = Rank(ourcards, boardcards)
  /*Consider all two-card combinations of remaining
cards*/
  for each case(oppcards) {
    opprank = Rank(oppcards, boardcards)
    if(ourrank > opprank) ahead++
      else if (ourrank == opprank) tied++
        else behind++
  }
  return (ahead + tied / 2) / (ahead + tied + behind)
}
```

The main advantage of this algorithm is that it works in any round of Texas Hold'em. Moreover, it is capable of giving a quick estimate of the probability of success since the number of cycles in this algorithm is small (1225 at most). It is also possible to calculate the hand strength against a varying number of opponents by raising the found probability to that number (equation 1).

$$\boldsymbol{HS_n = (HS_1)^n} \quad \textbf{(1)}$$

3.2.1 Combining hand strength with opponent modeling

In this article [8], the author suggests that it is possible to combine the hand strength algorithm with opponent modeling, in order to calculate the hand strength taking into account the opponents. To do this, the new algorithm instead of iterating over every possible hand, it only iterates over the cards that the opponent probably has, using Sklansky Groups [3]. This approach was only tested in heads up games.

3.3 Hand Potential

Hand potential[6, 7] is an algorithm that calculates the possible evolution of the hand quality throughout the game. When the game reaches the Flop round, there are still two more board cards to be revealed. This means that the current hand rank might improve; because the hand is composed of a set of five available cards (pocket or community cards) that has the highest rank among all available cards.

This algorithm is very similar to hand strength, but instead of only considering the current available cards, it considers the possible community cards that have not been revealed yet. This algorithm also considers that the opponents hand might improve as well.

Hand potential has two components:

- Positive potential: of all possible games with the current hand, all scenarios where the agent is behind but ends up winning are calculated.

- Negative potential: of all possible games with the current hand, all the scenarios where the agent is ahead but ends up losing are calculated.

The components of hand potential can be calculated as follows:

```
function HandPotential(ourcards, boardcards){
  int array HP[3][3], HPTotal[3] /* Init to 0 */
  ourrank = Rank(ourcards, boardcards)
  /* Consider all two-combinations of remaining cards*/
  for each case(oppcards) {
    opprank = Rank(oppcards, boardcards)
    if(ourrank>opprank) index = ahead
      else if(ourrank=opprank) index = tied
        else index = behind
    HPTotal[index]++
    /* All possible boards to come. */
    for each case(board) {
      ourbest = Rank(ourcards, board)
      oppbest = Rank(oppcards, board)
      if(ourbest>oppbest) HP[index][ahead]++
        else if(ourbest==oppbest) HP[index][tied]++
          else HP[index][behind]++
    }
  }
  PPot = (HP[behind][ahead] + HP[behind][tied]/2 +
HP[tied][ahed]/2) / (HPTotal[behind]+HPTotal[tied]/2)
  NPot = (HP[ahead][behind] + HP[tied][behind]/2 +
HP[ahead][tied]/2) / (HPTotal[ahead]+HPTotal[tied]/2)

  return (PPot, NPot)
}
```

The main advantage of this formula is that it considers future rounds, which is important because some games might reach showdown, being for this reason more precise than hand strength. Regarding disadvantages, this formula can't be used in River round (because the hand can't evolve no more). This formula cannot be used in Pre Flop rounds, because we can't calculate the hand strength for a two hand card. This might be solved by combining this algorithm with Chen Formula which will be explained later.

3.3.1 Combining hand potential with opponent modeling

Just like the Hand strength algorithm, if the Hand Potential is modified to only iterate over cards that the opponents might have [8], it is possible to obtain a better estimate over the chances of winning.

3.4 Effective Hand Strength

It is possible to calculate the probability of winning by combining the Hand Strength and Hand Potential algorithms (equation 2).

$$\mathbf{Pr(win)} = \mathbf{HS} \times (\mathbf{1} - \mathbf{NPot}) + (\mathbf{1} - \mathbf{HS}) \times \mathbf{PPot} \quad \textbf{(2)}$$

By setting the Negative Pot Potential to 0, it is possible to determine the effective hand strength which is the probability of the hand is either the best or will improve to become the best (equation 3).

$$\mathbf{EHS} = \mathbf{HS} + (\mathbf{1} - \mathbf{HS}) \times \mathbf{PPot} \quad \textbf{(3)}$$

3.5 Chen Formula

Chen formula is a mathematical formula developed by the professional poker player William Chen[9]. This formula can determine the relative value of the pocket hand. The main advantage of this formula over hand strength and current hand odds is that it does not need to generate permutations of card sets. For this reason, this algorithm is much faster than the others. The disadvantage of this algorithm over the others is that it only supports two card hands thus working only in Pre Flop rounds.

The following code represents the Chen Formula algorithm.

```
function ChenFormula(card1, card2){
  baseScore = Max(Score(card1), Score(card2))
  if(card1.value == card2.value) { /* if is pair */
    baseScore = Max(5, baseScore * 2)
  }
  if(card1.suit == card2.suit) {
    baseScore += 2
  }
  gap = Abs(card1.value - card2.value)
  switch(gap) {
    case 0: break;
    case 1: baseScore++; break;
    case 2: baseScore--; break;
    case 3: baseScore-= 2; break;
    case 4: baseScore-= 4; break;
    default: baseScore-= 5; break;
  }
  return baseScore - gap;
}

function Score(card) {
  switch(card.value) {
    case Ace: return 10; break;
    case King: return 8; break;
    case Queen: return 7; break;
```

```
      case Jack: return 6; break;
      default: return card.value / 2.0; break;
    }
  }
```

The algorithm is composed by two functions. The Score function returns a real number that scores a card. For instance, any ace card has the biggest score possible (10). The ChenFormula function returns an integer that represents the value of the hand. The max value of the returned value is 20 for a double Ace hand.

4 Hand Rank Evaluation

A poker hand rank evaluator is a function that receives a set of cards (5 to 7) and returns a number that means the relative value of that hand. This function is used to calculate lower level probabilities. For instance, every hand odds function described in this article (on the previous section) uses a rank function in order to determine if the player's hand is better than the opponent's hand.

Building an algorithm to determine the hand's rank is a relatively easy task, using naive methods, i.e. using an algorithm that intuitively makes sense [10]. Naïve hand rank evaluators usually follow the next steps:

- Sort the hand by card value (Deuce has the lowest value and Ace has the highest);
- Iterate across the hand, collecting information about ranks and suits of the cards;
- Make specific tests to check if the hand is of certain type (Flush, Straight, Three of a Kind, …);

The main problem of naïve evaluators is that they are slow. The hand odds evaluation functions call hand rank functions multiple times for each call, which means that having a slow hand rank function might slow down the calculation of the probability of success. The solution to this problem is to use top-down dynamic programming algorithms in order to speed up the rank function. The following subsections will present some hand rank functions.

4.1 Cactus Kev's 5-Card Evaluator

The Cactus Kev's 5-Card Evaluator uses one of fastest hand evaluator algorithms. The idea behind the algorithm is using a pre-computed hand ranking table. The number of possible combinations can be calculated as in equation 3.

$$N = \frac{(n!)}{(r!(n-r)!)}, n = 52, r = 5 \rightarrow N = \frac{52!}{5!(47!)} \rightarrow N = 2598960 \quad (3)$$

Since the number of combinations is not that high, it is quite computationally feasible to store all hands value in a 10mb table (2598960 * 4 bytes). The problem is that all hands must be in order. Cactus Kev's 5 card evaluator uses prime numbers to order the cards, because multiplying prime values of the rank of each card in your hand will result in an unique product, regardless of the order of the five cards [11].

Cactus Kev's evaluates cards with the following prime values: Two – 2; Three – 3; Four – 5; Five – 7; Six – 11; Seven – 13; Eight – 17; Nine – 19; Ten – 23; Jack – 29; Queen – 31; King – 37; Ace – 41.

For instance a King High Straight hand will always generate a product value of 14,535,931. Since multiplication is one of the fastest calculations a computer can make, we have shaved hundreds of milliseconds off our time had we been forced to sort each hand before evaluation [11].

The only big limitation of this hand evaluator is that it can only be used to evaluate 5-card hands. This means that to use it in game variations like Texas Hold'em which needs to evaluate 7-card hands (in River round), the function had to evaluate all possible 21 combinations of 5 cards to determine which one was the best.

4.2 Pokersource Poker-Eval

Poker-eval is a C library to evaluate poker hands. The result of the evaluation for a given hand is a number. The general idea is that if the evaluation of the player's hand is lower than the evaluation of the hand of its opponent, then it loses. Many poker variants are supported (draw, Hold'em, Omaha, etc.) and more can be added. Poker-eval is designed for speed so that it can be used within poker simulation software using either exhaustive exploration or Monte Carlo [10, 12]. Poker-eval is probably the most used hand evaluator, because of its multi-portability of poker variants and its speed of evaluation. The only limitation might be the complexity of its low level API, however there are some third party classes that encapsulate the usage of Poker-Eval API, making it simpler to use.

4.3 Paul Senzee's 7-Card Evaluator

Paul Senzee's 7 Card Evaluator [10, 13] uses a pre computed hand table to quickly determine the integer value of a given 7 card hand. Each hand is represented by a 52 bit number, where each bit represents an activated card. The total number of activated bits is 7, representing a 7 card hand.

If unlimited memory was available, it would be possible to index the resulting rank value into an enormous and very sparse array. However, this would require an array with 2^52 entries which means it would not be, nowadays, computationally feasible as it would require about 9 petabytes of memory (9 million gigabytes).

To solve the problem, Paul Senzee's developed a hash function that turns the hand value into an index between 0 and roughly 133 million and computed a 266mb which is by far a much smaller table, by grouping similar hand ranks.

The limitation of this hand evaluator is that it only evaluates 7 card poker hands, therefore is not portable to other poker variants.

4.4 TwoPlusTwo Evaluator

TwoPlusTwo evaluator is another lookup table poker hand evaluator with the size of 32487834 entries with a total size of ~250mb[10]. However TwoPlusTwo Evaluator is extremely fast, probably the fastest hand evaluator there is. To get the value of a given hand, the process is just performing one lookup per card. For instance to get a 7 card hand value the code with be just (admitting HR is the lookup table):

```
function Rank(cards) {
  p = HR[53 + cards[0]]
  p = HR[p + cards[1]]
  p = HR[p + cards[2]]
  p = HR[p + cards[3]]
  p = HR[p + cards[4]]
  P = HR[p + cards[5]]
  P = HR[p + cards[6]]
  return p
}
```

The idea behind the implementation of this extremely fast evaluator is a state machine. Each entry on the table represents a state. The next state is the sum of the value of the card and the value of the state. In the final state, the value represents the hand value.

This hand evaluator has no real limitations since it supports 5, 6 or 7 card hands.

4.5 Hand rank evaluators benchmark

In order to determine the fastest hand rank evaluator, a benchmark was performed. The results are shown in the following table.

Table 1. Hand rank function benchmark

Hand rank function	Elapsed time for 100.000 trials (ms)
Cactus Kev's 5-Card Evaluator	420 ms
Pokersource Poker-Eval	300 ms
Paul Senzee's 7-Card Evaluator	380 ms
TwoPlusTwo Evaluator	85 ms

As it can be observed, TwoPlusTwo Evaluator is by far the fastest hand rank evaluator.

5 Conclusions

There is still a long way to go to create an agent that plays poker at the level of the best human players. This research presented how an agent can measure the quality of its hand in order to aid its decisions during the game. This is rather important to build a competitive artificial agent in the future, because it is impossible to play well without knowing how strong the hand is.

There are five main hand odds evaluators to check the probability of winning taking in to account the current game state. Every one of them provides relevant information in each game round. Hand potential and hand strength algorithms were also modified in order to integrate opponent modeling with this measures.

With respect to hand rank evaluators, some of them proved to be very fast, which is important in because they are called many times by the hand odds algorithms. The fastest evaluator was TwoPlusTwo evaluator, achieving the smallest elapsed time in a simple experience.

In future work, the researchers should focus on combining even more this measures with opponent modeling techniques, in order to create better strategies.

References

[1] D. Billings, *et al.*, "Opponent modeling in poker," presented at the Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence, Madison, Wisconsin, United States, 1998.

[2] A. Davidson, "Opponent modeling in poker.," Master, Department of Computing Science, University of Alberta, 2002.

[3] D. Sklansky, *The Theory of Poker: A Professional Poker Player Teaches You How to Think Like One*: Two Plus Two, 2002.

[4] N. M. S. Ulam, "The Monte Carlo Method," *Journal of the American Statistical Association,* vol. 44, pp. 335-341 1949.

[5] P. Programmer. (2006, 02-01-2011). *Poker for Programmers: Poker Algorithms and Tools for the C# Programmer.* Available: http://pokerforprogrammers.blogspot.com/

[6] D. Billigs, "Algorithms and Assessment in Computer Poker," Doctor PhD, Department of Computing Science, University of Alberta, 2006.

[7] D. Billings, *et al.*, "The challenge of poker," *Artif. Intell.,* vol. 134, pp. 201-240, 2002.

[8] D. Felix, *et al.*, "An Experimental Approach to Online Opponent Modeling in Texas Hold'em Poker," presented at the Proceedings of the 19th Brazilian

Symposium on Artificial Intelligence: Advances in Artificial Intelligence, Savador, Brazil, 2008.

[9] B. C. a. J. Ankenman, *The Mathematics of Poker*, 1 ed.: Conjelco, 2009.

[10] (28 November). *Coding the Whell: Poker Hand Evaluator Roundup*. Available: http://www.codingthewheel.com/archives/poker-hand-evaluator-roundup

[11] (2006, 01/01/2011). *Cactus Kev's Poker Hand Evaluator*. Available: http://www.suffecool.net/poker/evaluator.html

[12] (2006-2010, 01-01-2011). *Pokersource Poker-Eval*. Available: http://pokersource.sourceforge.net/

[13] (2007, 01-01-2011). *Paul Senzee on Software, Game Development and Technology*. Available: http://www.senzee5.com/2007/01/7.html

SESSION 4

COMPUTER GRAPHICS

Chairman: João Tiago Pinheiro Neto Jacob

A Conceptual, Generic and Object Independent Animation Controller

Nuno Barbosa

Faculdade de Engenharia da Universidade do Porto, Porto, Portugal
nuno.barbosa@fc.up.pt

Abstract. Computer graphics animation require multiple methods and control mechanisms. Each created model, simple or complex, needs a system to control the animation. If creating an application to support one kind of animation is hard, supporting more than one is obviously harder but imperious, since it will simplify the task for animators and programmers. This paper presents a new conceptual system for handling multiple animation types, in real-time (video games) or off-line (films) rendering pipelines, with shading support, scripting schemas and that generalizes the animation controlling process. We first compile a set of animation techniques, select the techniques we need to use in the final release and build all add-ons of the application based on the selected techniques. We also present an implementation example which suggests multiple input formats, external application integration based on the previous methodology.

Keywords: Animation, Controllers, Computer Graphics, Graphics data structures and data types

1 Introduction

Current animation technologies provides the developer multiple animation techniques. This makes it harder to develop a framework for the support of several types of animation. The lack of conceptualization for such system makes a first approach on the subject essencial. In this paper, we show an initial presentation of a conceptual system that controls animation, ignoring details up until an abstraction layer that enables the programmer or the artist to only care about some minor parameters. This article can be used as a guideline for future systems' development, either for real-time applications, such as video games, or off-line rendering systems, used in films. We employ a basic finite state automata to help us describe and conceptualize the controller, called State Manager, and state transitions, without entering too deep in the automata subject. We assume that the reader has a small knowledge of computer graphics and how animation is processed in current systems and animation applications. A practical example is also described on this article.

There exists several animation techniques and implementing all in one system is time-consuming and hard. We selected a group of techniques based on a

pre-compiled and more global set and them we build the framework using all selected tools and the more specific selected animation techniques. In this article, the system supports three kinds of animation: standard, which includes translation, rotation and scale; blend shapes, based on a linear interpolation between two or more complete meshes of the same object and commonly used for facial animation; and skeletal animation, used for full body animation[1].

The next section presents a summary on state-of-the-art animation techniques, compiling some of the main existing techniques; the third chapter makes conceptual description of the system, and the fourth shows an implementation example; we later present a conclusion of the article.

2 Animation Techniques State-of-the-Art

Keyframe Interpolation is the most basic and ancient type of character animation because it was the first ever present animation technique. It consists on specifying complete models for a given set of points in time, called key frames or key poses. Each key frame is a base pose for the model and can be used to define a movement, like traditional hand-drawing pose-to-pose techniques. The key frames consist on the original model redefined for the pose we desire. The in-between frames are generated by interpolations. A problem is that we possibly need to define several key frames' poses or we may end up with a faulty and unpredictable animation, especially on complex models, and the path of action will usually be incorrect leading objects to intersect each other[2].

Geometric Deformation is based on the idea that to simplify the manipulation of a complex object we can use a simpler one (ex.: NURB curves) to control it, by presenting a simpler and easier control interface (ex: using a NURB curve, we can manipulate all the vertices associated with the neck of a human head). It also provides a geometrical high level control over the deformation, it is commonly used in video-games because it offers fast and easy animation generation, it is easy to create a system to support it and it provides flexible and effortless model deformation's tools. The first geometric deformation technique is called non-linear global deformation[3]. A typical geometric approach is the Free-Form Deformation (FFD)[4], which deforms solid geometric models in a free-form manner with the use of trivariate Bernstein polynomials. It enables deformation of objects by manipulating the control points. It was later extended to change the shape of an existing surface either by bending it along an arbitrarily shaped curve or by adding randomly shaped bumps to it using non-parallelepiped type 3D lattices[5]. The FFD suffered more extensions[6][7] but all this extensions made the technique lose some flexibility, stability and control.

Physically-Based techniques try to simulate all the Newtonian physics laws and the elastic and viscosity properties of the model[8]. These techniques involve a large number of calculus and processing. There are three ways to create physically-based simulations: mass-springs, which is a combination of mass and springs simulations in the mesh to represent reactions of the tissue to physics laws and viscosity properties[9]; finite elements algorithms[10]; finally, a com-

bination of mass-springs and finite elements algorithms[11]. Since the goal of physically-based techniques is to simulate in animation what happens in the real world, the calculations involved makes this methods not suitable for real-time animation, especially in video games, where graphics have an important role. An interesting technique is presented in [12] where the animator crafts key frames in an interactive physical simulator, and then automatically constructs a reduced control basis using these key frames and their tangents. Optionally, the basis is augmented with natural modes of the deformable model. The basis is then used to solve a reduced space-time optimization problem with tons of controls per time step instead of thousands. Building a reduced basis for the control of deformable models is not straightforward. Key frames should be created using simulation so that they are compatible with optimization. A problem we have with this method is that we need to combine it with other methods to achieve rigid objects animation.

Animation retargeting technique tries to solve the problem of adapting the motion from a character to another with different proportions or shape. In the adaptation of the motion, we need to modify the source data so that it can serve the target model to create the animation[13]. A method that mapped video-recorded performance of one individual to another, generating a detailed 3D texture face mesh for the target identity was also presented in [13]. Another system, introduced in [14], works by using a given a model, analyzing it and creates a rig ready to be animated. The system receives, as input, two models: the first is a complete model with the rig and the complete set of attributes (skeleton, influence objects, shapes and animation scripts); the second is the target model, which doesn't have a character rig associated to it. Then, the system automatically transfers the rigs and the animation from the source model to the target model. The method follows the principals of a non-linear warp transform[15], and uses facial features landmarks, resulting in an efficient deformation technique. In cases of animation transfer of extreme characters, such as the transferring of the rig and the animation from a human model to a cartoon model, the animation may seem imperfect because, normally, the cartoon based character has more exaggerated movements. Also, some final touches from the artists are sometimes needed so that the animation fulfills the quality requirements. A method for facial animation that uses anthropometric statistics is described in [16]. It takes 3D face scans as examples in order to exploit the variations presented in the real faces of individuals. Also, it uses Principal Component Analysis (PCA) to learn a model and, for each facial feature, it computes a set of anthropometric measurements to parameterize the example meshes into a measurement space. By using the PCA coefficients as a compact shape representation, the face model problem is formulated in a scattered data interpolation framework which takes the user-specified anthropometric parameters as input.

In the next section we try to abstain ourselves from the character animation types, focusing on how we can build a framework that, in the end, will hide the animation handling difficult part.

3 Description

A system as the one described in this paper must support the most common functionalities provided by any other application of its kind and offer some kind of uniqueness that other frameworks do not provide. We are describing a conceptual system, meaning that we can create a system based on this principles but target it to a specific platform, engine or application.

We first settle the basic supported capabilities the system must have.

3.1 Basic Capabilities

Between the basic capabilities, we can enumerate some that are crucial to the final deployed framework:

- Multi Application Program Interface (API) support: this may provide cross-platform execution;
- Mathematical operations: support for vectors, matrices, translations, rotations and scaling operations;
- Shaders support: to enable top quality graphics animation, the system must support shaders;
- Good learning curve: an easy to learn and implement (or re-implement) system which is better accepted and more commonly used.

If the system does not support any of the described basic capabilities, it would lose part of its interest since it would become more suitable to accept another application that could be extended to suppress the programmers' or artists' needs and to serve as the development platform. Many libraries or graphics engines available already provide us with the basic support listed above.

The offer available on the market and, also, on open source solutions, makes ir unreasonable to develop a new API. This is why the work description in this article does not intend to deploy a new state-of-the-art API. We pretend to provide a new methodology that enables an already existing engine to be extended, supporting new functionalities and increasing its capabilities.

The described functionalities cannot be discarded in any graphics application, but we need extra tools embed on the deployed system that completes the framework, such as a scripting tool.

3.2 Scripting

The use of scripts is a great addition to any application, not only to animation applications. For animation we can use a script to load a scenario or even load a 3D model (as an example, the reader may be interested in looking to the Collada schema[1]).

We propose a script development that could load scenarios, models (characters or not) and animations. This script works similarly like a movie script:

[1] `http://www.khronos.org/collada`

it is divided by scenes; each scene may take place in a different location from any other, with possibly different characters and several animations by each character. The script may be based on a well-known schema, such as XML, or developed from scratch. It is not important the path the developer chooses but it is important that the provided solution can be easily adapted and altered during the production stage.

With the script load, we can start animating the scene with an implemented animation support on our system.

3.3 Animation Support

In Section 2 we already introduced the four main techniques for animation. Ideally, the system should support the different animation types but this is time consuming and the developer normally does not need full animation support.

The critical part on the subject is how it can be added to a system without changing the system itself. Obviously, the developer must choose the base framework for the project he is involved. A video game will require a different framework from a film. We must see the problem from a theoretical point of view and remember that we are not presenting an all-in-one graphics development application.

We can simply define multiple animation using this guideline: Hide the complexity! This guideline is truly the holy grail for a system used for fast development. Approaches like using the same function to load or control any kind of animation, simple sequence loading and control, greatly simplify the process of development, reducing the time to launch a new video game or film. This may seem obvious but there are not that many good examples of this implementation. Current systems have different handler functions for different animation types, making it harder to master each technique.

The described multi animation controller does not handle all the tasks associated with animation; part of them must be handled by shaders.

3.4 Modular Shader Support

Actual graphics applications need shaders to exist. Shaders provides the cinematic beauty of animation and are programs that are used to create rendering effects. We have two sort of shaders: the real-time shader, loaded directly to the Graphics Processing Unit (GPU), commonly used in video games but with less accurate effects; the off-line shader, which can be either loaded to the GPU or to the Central Processing Unit (CPU), used, for example, on films and physically-based simulations with more accurate results.

In order to have modular support, we need to run more than one shader at a time and, possibly, to use different shading languages. This can be easily implemented in real-time rendering engines, but it is unpractical to do in off-line rendering applications. We can understand this because, on real-time, the shaders are compiled to run on the GPU, meaning that two shaders created

with different shading languages will be prepared to have similar instructions sets for the GPU since the compiler is embed on it. This does not occur in off-line rendering, since the compilers are different and the rendering calculation may differ. The only reason for using different shaders in this case is that one may have better results in some kinds of scene than the other (ex.: a shader may return better results on closed environments than another).

Another important reason for modular shader support is the existence of different materials. We have several surfaces, which implies different properties. In other words, we have different materials for each object. This also enables new effects to be applied to a scene, like a glow or mirror effect.

Several shader parameters are passed to the shader in run-time, meaning that we need a manager that can control these parameters and all the functionalities defined in the previous sections.

3.5 State Manager

A state manager is a data structure that enables the application to control the animation. Each state is a complete action and a state transition is the action taking place. We can imagine the state manager like a finite automata.

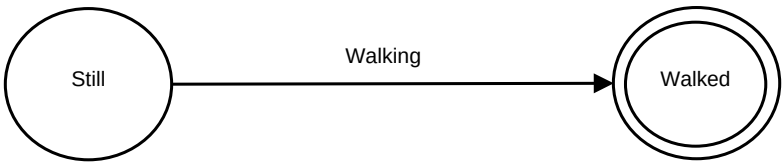


Fig. 1. Automa example for State Manager

Figure 1 shows how the state manager works. This is a very simple example where the character is "Still" and "Walks" until it stops and its state changes to "Walked". The complexity of the automata may increase depending on the number of actions to be executed by the character.

Also, each state may interact differently with a shader and can have a different shader associated with it. It is completely plausible that different characters may have different aspects (ex.: a human has skin, a robot has metal) and, to each special looking feature, we may have a different shader with different parameters. The interaction with the shader must be partially handled by the state manager, by the control of the parameters and by the rendering engine, which transmites them to the shader.

Figure 2 is a small conceptual design of the state manager and all related information to each state and the manager itself. Remember that, for each model, we have a state manager and a different number of animations, meaning we can have a variant number of states per model.

We can now give a simple demonstration of our controller, which implements all the requirements described before.

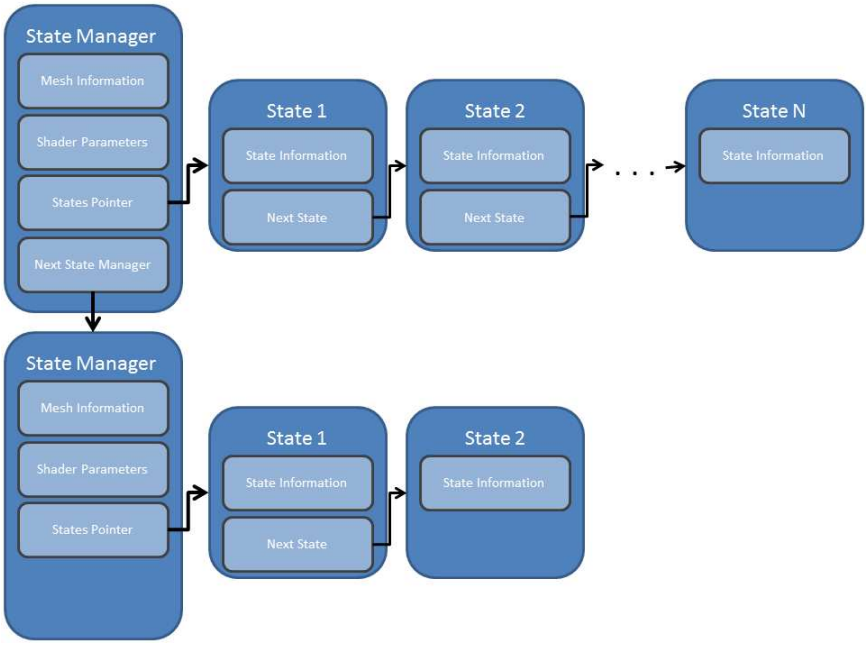


Fig. 2. Example of a State Manager conceptualization

4 Implementation Example

The example presented in this section is a clear demonstration of the described animation controller benefits. We used only open-source or freely available solutions for this demonstration. The complete description of the example system is available in [19].

4.1 Tools

Regarding the rendering engine, we have selected the popular Object-Oriented Graphics Rendering Engine (OGRE)[2], and for the scripting part we used a XML schema called dotScene[3] which was extended it to suit our needs for the system. The parsing capabilities for the described XML schema are provided by TinyXML[4] parser.

4.2 State Manager

The first stage of the application development is the creation of a list for the establishment of the required fields for the control of the animation. In the given

[2] http://www.ogre3d.org/
[3] http://www.ogre3d.org/tikiwiki/DotScene
[4] http://www.grinninglizard.com/tinyxml/

example presented in this article, we focus on three animation types: standard, which we define as solid objects with basic mathematical operations like translations, rotations and scaling; blend shape animation[17] based on key frames; and skeletal animation[18], a geometric deformation controller.

The standard and blend shape animations are based in steps, meaning that each animation will occur in a specified number of steps, since the animation is based on a linear interpolation between the end of a state and the beginning of the following. If the artist defines that an animation should be fast, then he can define a small number of steps; if the animation must be slower, then he can define a larger number of steps making it take more time to end. Skeletal animation can also be based in steps but, to demonstrate the modularity of this implementation example, we have implemented it based on time. The calculation is based on the time per frame rate. This approach makes the skeletal animation dependable on the machine, so the team must previously define the minimum required machine to run the application (the well-known hardware prerequisites) and the model detail. Greater detail makes a model's mesh more complex and harder to process by the computer.

Each manager must have the information related to the model itself (i.e.: mesh file, textures...) and also the type of animation and, for each state associated to the manager, must have the number of steps or the time, depending on the animation type. Additional information is provided to enable the interaction with a shader. The complexity of building the state manager is hidden from the programmer since the only difference between each animation type is, when creating the manager, to declare the type; the rest is up to the manager.

4.3 Script

The dotScene schema does not provide support for generating sequences of scenes or animations, meaning that it can only define a static scene. The positive side about dotScene is that it already supports many world and environment definitions, such as lights and cameras, and has already built open-source exporters for the main animation softwares as well as importers for OGRE, making the task for extending the script simpler and faster. This happens because we do not need to completely create an exporter or importer, just to add the extra information to each one of the tools to suppress our needs.

An example of the original dotScene example is showed in Figure 3. With a close analysis to the code, we notice the inexistence of animation related information. The next step is adding the required information on the correct place. Naturally, the animation is associated with a model which is declared by the XML tag "node". The animation related tags should be declared as "node" childes tags.

Figure 4 presents the new extended schema. We added new child tags to "node". The first added tag is the "manager" tag which defines the state manager as described in Section 4.2; the "states" tag and its childes define the states associated with the current model; "sequences" define the sequence on which

```
1
2  <scene id="0" formatVersion="0.2.0" sceneManager="any" minOgreVersion="0.14.0">
3      <environment>
4          <colourBackground r="0.5" g="0.5" b="0.5" a="1"></colourBackground>
5      </environment>
6      <nodes>
7          <node name="Realista" id="">
8              <position x="-1.355854681e-031" y="-1.982230048" z="-0.7742838039"></position>
9              <rotation qx="0" qy="0" qz="0" qw="1"></rotation>
10             <scale x="1" y="1" z="1"></scale>
11         </node>
12     </nodes>
13 </scene>
14
```

Fig. 3. Original dotScene schema example

```
7          <node name="Realista" id="">
8              <position x="-1.355854681e-031" y="-1.982230048" z="-0.7742838039"></position>
9              <rotation qx="0" qy="0" qz="0" qw="1"></rotation>
10             <scale x="1" y="1" z="1"></scale>
11             <manager name="Realista" id="" meshFile="realista_ShapesSkin.mesh" animationT
   y="0.0" z="0.0">
12                     <states>
13                             <state stateName="Suprise" materialName="" actionName="surpris
   g="0.067" b="0.073"></state>
                                    .   .   .
19                     </states>
20                     <sequences>
21                             <sequence actionaName="Sad" passes="25" isReverse="false"><
                                    .   .   .
32                             <sequence actionaName="Surprise" passes="25" isReverse="tru
33                     </sequences>
34                     <loopAnimation value="true"></loopAnimation>
35             </manager>
```

Fig. 4. Extended dotScene schema example

the animation should occur; finally, "loopanimation" informs the framework if the animation should loop or not.

4.4 Shading

Shading is harder to handle since different shaders have different parameters. On OGRE, these parameters are defined on the material and we can access them in real time. We created a dictionary that associates each shader parameter to an OGRE defined index. This index is what OGRE uses to send the information to the correct variable in the shader. The dictionary solution proved to be very flexible, permitting us to alter the information in run-time, which means we can change the shader associated to a material while running the application.

4.5 Integration

With some minor changes, we integrated this system in an animation application (Autodesk Maya[5]). We used it as a scene previewer for films and video games animation sequences. The application was integrated as a plug-in. The host application (Maya) exports the models and the dotScene extended file to the built plug-in and starts the viewer. After a few seconds, the artist has a full and faithful preview of the scene he is creating. The main advantage with integration is that it spares the time spent waiting for large mainframes rendering machines to generates the scene. Obviously, the quality is not the same, but it can provide a close example of what the scene will become.

5 Conclusions

We defined conceptual, generic and object independent animation controller that can abstract the programmer and the artist from caring about how the animation is created. The system can use multiple animation types, character based models or non-living models and manage the animation. This abstract definition can be inserted in any production pipeline. The down side of this definition is that it must adapted to the selected API because each API handles animation and model processing differently, taking some initial developing time. However, that can be later compensated with the animation loading and optimization process. Other limitations may arise from the implementation itself, but those limitations should be identified at an early development stage and solved right after.

References

1. Xiao, Z.; Zhang, J. J.; Bell, S., Control of Motion in Character Animation, Proceedings of the Information Visualisation, Eighth International Conference, IEEE Computer Society, 2004, 841-848

[5] `http://usa.autodesk.com/`

2. Lasseter, J., Principles of traditional animation applied to 3d computer animation, SIGGRAPH '87: Proceedings of the 14th annual conference on Computer graphics and interactive techniques, ACM Press, 1987
3. Barr, A. H., Global and local deformations of solid primitives, SIGGRAPH Comput. Graph., ACM, 1984, 18, 21-30
4. Sederberg, T. W., Parry, S. R., Free-form deformation of solid geometric models, SIGGRAPH, 1986, 20, 151-159
5. Coquillart, S. Extended free-form deformation: a sculpturing tool for 3D geometric modeling, ACM Press, 1990, 187-196
6. Coquillart, S., Jancéne, P., Animated free-form deformation: an interactive animation technique, SIGGRAPH Comput. Graph., ACM, 1991, 25, 23-26
7. MacCracken, R., Joy, K. I., Free-form deformations with lattices of arbitrary topology, Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, ACM, 1996, 181-188
8. Terzopoulos, D., Fleischer, K., Modeling inelastic deformation: viscolelasticity, plasticity, fracture, ACM, 1988
9. Debunne, G., Desbrun, M., Cani, M.-P., Barr, A. H., Dynamic real-time deformations using space and time adaptive sampling, ACM, 2001
10. Koch, R., Gross, M., Bosshard, A., Emotion Editing using Finite Elements, ETH Zurich, 1998
11. Kähler, K., Haber, J., Yamauchi, H., Seidel, H.-P., Head shop: generating animated head models with anatomical structure, ACM, 2002
12. Barbic, J., Popovic, J., Real-time control of physically based simulations using gentle forces ACM, 2008
13. Lee, Y., Terzopoulos, D., Walters, K., Realistic modeling for facial animation, ACM, 1995
14. Orvalho, V. C., Susin, A., Fast and reusable facial rigging and animation, ACM SIGGRAPH 2007 sketches, ACM, 2007
15. Bookstein, F. L., Principal Warps: Thin-Plate Splines and the Decomposition of Deformations, IEEE Trans. Pattern Anal. Mach. Intell., IEEE Computer Society, 1989, 11, 567-585
16. Zhang, Y., Prakash, E. C., Face to face: anthropometry-based interactive face shape modeling using model priors, Int. J. Comput. Games Technol., 2009, 1-15
17. Yoo, T.-K., Lee, W.-H., Blend Shape with Quaternions, Proceedings of the 2007 International Conference on Convergence Information Technology, IEEE Computer Society, 2007, 776-780
18. Forstmann, S., Ohya, J., Krohn-Grimberghe, A., McDougall, R., Deformation styles for spline-based skeletal animation, Proceedings of the 2007 ACM SIGGRAPH/Eurographics symposium on Computer animation, Eurographics Association, 2007, 141-150
19. Barbosa, N.: Dynamic Skin-Shading: MSc Thesis, Universidade do Porto, Faculdade de Cincias (2010)

Towards Adaptive Occlusion Culling in Real-Time Rendering

Vítor Cunha

FEUP - Faculdade de Engenharia da Universidade do Porto
Rua Dr. Roberto Frias
4200-465 Porto, Portugal
pro10003@fe.up.pt

Abstract. The performance of real-time graphical simulations may suffer due to the lack of resources needed to render very large and complex virtual environments. In these situations the use of occlusion culling techniques enables the removal of occluded geometry, thus decreasing the amount of work sent to the graphics hardware to generate the image presented to the user. Recent approaches use the Hardware Occlusion Query mechanism to tackle the problem of occlusion determination. Despite being appropriate to use in real-time this technique comes with a cost, especially if the tested geometry is found to be visible. Appling the principle of temporal coherence between consecutive frames can reduce the total cost of occlusion culling by reducing the number of deployed queries. This paper presents a hierarchical occlusion culling method which exploits temporal coherence by adapting the queries interval according to the geometry visibility history. Several approaches to use this method in an adaptive way of determining occlusion in real-time graphical simulations are also discussed. The presented method is conservative and to demonstrate its benefits the results of several simulations using a large and depth complex customized virtual environment are presented.

Keywords: visibility, real-time rendering, occlusion culling, hardware occlusion query.

1 Introduction

Real-time graphical simulation refers to the ability of a software application to create synthetic images on a computer fast enough so that the user can interact with a virtual environment. An example is the driving simulation in urban environments where typically wide scenes with very high geometry density are used. To achieve interactive rates when simulating these complex virtual environments, the software applications usually use techniques to remove geometry that does not contribute to the final image presented to the user. This operation is called culling. The goal is to find all geometry that should be removed considering the current point of view to the scene. This way saving resources that can be used, for instance, to generate more frames per second. To deal with the visibility problem vast amounts of research have been made and several algorithms have been proposed over the years.

One of the techniques used to cull geometry has the goal of identifying occluded portions of a scene, inside the view-frustum, which do not contribute to the final image. To solve this *occlusion culling* problem several algorithms have been proposed. With the introduction of an OpenGL extension, generally designated by Hardware Occlusion Query (HOQ), the efficiency of this set of algorithms has evolved significantly.

The OpenGL extension introduced by Hewlett-Packard [1][2] and improved [3] by NVIDIA [4] allows hardware assisted occlusion determination. With this mechanism, a software application can query the graphics hardware concerning the visibility of a given geometry, e.g. bounding volume, against previously drawn geometry. The returned query result is the number of visible pixels of the rasterized geometry. This result can be used by the application to decide whether or not to render the object inside the tested bounding volume. Although being a simple mechanism the occlusion test isn't free. The latency between the query (GPU) and the return of the result to the application (CPU) may negatively affect the performance. Furthermore, the query itself uses resources and if the tested geometry is determined as visible no profit comes with the use of HOQs.

To minimize the effects of HOQs the number of tests performed to the geometry during a simulation must be reduced. To achieve this goal one possible strategy is applying the principle of temporal coherence. Other types of coherence are identified in [5]. If the point of view to the scene is static or moving very slowly, calculations made for the current frame may still be valid for the following frames. This is the idea behind temporal coherence. In terms of visibility, an object identified as occluded in the current frame, will probably continue occluded in the following frames. However, it is difficult to predict how the point of view will move in the following frames or as a set of visible geometry will behave in dynamic scenes.

In this paper, initially, an occlusion culling method that takes advantage of temporal coherence is presented. This method is applied to a hierarchically organized 3D scene where the time interval between occlusion tests is set accordingly to the geometry past visibility test results. Simply put, if some geometry is found to be visible for several consecutive frames, the probability of continuing visible increases. Also in some cases, errors in the final image presented to the user are allowed. In this situation the algorithm could be set for aggressive culling reducing even more the number of queries deployed. Finally, some approaches are discussed on how to use this method in an adaptive way to determine the occlusion information in real-time based on scene and simulation properties. Several simulations with a very high geometry depth scene support that the use of this method allows the reduction of issued queries while keeping its conservative nature.

To take full advantage of the HOQ mechanism the scene must provide enough occlusion. In scenes with little geometry depth the performance could be worst then the use of the simple view-frustum culling [6]. There are also situations where the use of occlusion tests is not adequate even if the scene is complex. For instance, a flight simulator application is not a good candidate to use HOQs because of the lack of occluders from the user's point of view to the scene.

The remainder of this paper is structured as follows: in the next section is presented an overview of several visibility algorithms, in particular those that use HOQs. In Section 3 the proposed method is presented in detail. Section 4, presents the

simulation setup and the results obtained which are discussed in the following section. Finally, in Section 6 the conclusions and future developments proposals are presented.

2 Related Work

Over the past several years algorithms that address the fundamental problem of visibility determination have been presented. Covering the entire literature is beyond the scope of this paper. Detailed studies presented by Cohen-Or et al. [7] and Bittner and Wonka [8] provide an overview and taxonomy that allows the comparison of the different approaches to the visibility problem.

Regarding visibility determination the algorithms can be grouped in two sets: *from-region* and *from-point*. The *from-region* algorithms determine visibility for a region, and work in a preprocessing stage to determine the Potentially Visible Set (PVS). An example of this approach is the method of cells and portals for architectural models [9]. The *from-point* methods determine the visibility from a point (point of view) and work in runtime. The focus of this paper is in this last set of algorithms. These methods require more calculations each frame but have a greater flexibility, allowing dynamic scenes and naturally performing occluder fusion since they work in image space. An occlusion culling algorithm conceptually important, but that was never fully implemented in hardware was presented by Greene et al. [10], named Hierarchical Z-buffer. Despite the variety of proposals before the existence of dedicated hardware to support occlusion culling, the algorithms were considered too costly to be used in practice, with some exceptions such as the Hierarchical Occlusion Maps [11] and the dPVS [12].

With the introduction of hardware assisted visibility determination, several algorithms that use the HOQs have been presented. Hillesland et al. [13] presented one of the first algorithms to use the NVIDIA extension. This algorithm doesn't use temporal coherence and has limitations in scenes with variable geometry density. The Coherent Hierarchical Culling (CHC) algorithm presented by Bittner et al. [6] explores spatial and temporal coherence to avoid CPU stalls and GPU starvation. Regarding temporal coherence, this algorithm assumes that a visible scene node remains visible during a predefined number of frames. The queries time alignment resulting from this approach was addressed by Staneker et al. [14] who uses temporal coherence to distribute the queries throughout several frames. Presented in 2006 by Guthe et al. [15], the algorithm Near Optimal Hierarchical Culling (NOHC) applies a cost/benefit function to all geometry to decide on their rendering. This function is only applied to visible geometry if a fixed number of frames have passed since the last test. The outcome of the function is also heavily dependent on a calibration step performed in preprocessing. Recently, Mattausch et al. [16] presented a revised version of the CHC algorithm, the CHC++. It uses *multiqueries* (one query where several geometry occluded nodes are tested) and the distribution of visible geometry occlusion queries through a random range of frames. However, it is necessary to wait for a predefined number of nodes to become available before a *multiquery* could be performed, leading to a delay in the availability of results.

3 Temporal Coherence Applied to Occlusion Culling

The main goal of the following presented method is to reduce the total number of HOQs, thus minimizing the negative effects that the use of this OpenGL mechanism has on the software application. The strategy is to consider that geometry that is determined visible tends to maintain that state. In consecutive frames, if the geometry keeps the same visibility state, then the occlusion tests are performed with even greater interval. To make this algorithm conservative the same strategy can't be applied to previously occluded geometry, as explained in next section.

To develop the software application needed to test the method, the OpenSceneGraph API [17] was used. Since this API already provides means of applying HOQs to a 3D scene, the available basic occlusion culling algorithm was used as a work base.

3.1 Occlusion Culling in OpenSceneGraph (OSG)

With an object-oriented approach to scene graphs, the OSG API provides a range of nodes [18]. All nodes share a common class (`osg::Node`) with the specialized features of each one defined in derived classes. One of these specialized nodes is the `osg::OcclusionQueryNode` (OQN), which performs the occlusion test to the node's child tree structure using its bounding volume. Being a node, the OQN can be inserted permanently in the hierarchical 3D data structure or during runtime to test scene portions. The OQN class provides methods to: turn on/off the use of the node; set the minimum number of pixels to consider the tested volume visible; establish a frame interval between occlusion tests. These features allow a conservative or approximate approach to visibility determination. An inactive OQN in the scene graph behaves like a normal group node.

During the Cull traversal of the scene graph the OQNs are processed. In the OSG version used (v2.8.1) three situations may occur: if the OQN is outside the view-frustum, it's removed from further processing along with its child tree; if in the available occlusion query result the node was found occluded, the OQN child tree isn't rendered. This OQN only be tested again if the defined frame interval between tests is up; the third situation occurs when the node is determined to be visible. For that to happen, the query result (number of visible pixels) must be greater than the defined value. In this case the traversal continues to the child nodes. To update the visibility information a new occlusion test is performed, also if the defined frame interval between tests is up.

When compared with the algorithms presented in the previous section, the OSG's occlusion culling algorithm is somewhat simpler. The process of visibility information update doesn't distinguish previously occluded from previously visible nodes. This feature combined with the use of a frame interval between occlusion tests raises some issues. This use of temporal coherence means that the algorithm assumes that a node, regardless of its visibility state, tends to remain in that same state in the following frames. This approach generates a smaller number of tests, but also implies that the information for the correct visibility determination may be obsolete in the current frame. As a result, the variation of the frame interval between tests will have

an enormous influence in drawn geometry. For large intervals the negative aspects of HOQs have less impact on the application, which should lead to a better performance. However, the amount of wrongly drawn geometry increases. For instance, when an object which is in reality visible isn't rendered due to still being considered occluded as a result of the lack visibility information update. In this case some geometry will be drawn when in reality it could be occluded by the object that was not rendered by incorrectly being considered occluded. This situation shows that the decrease in the occlusion tests frequency not always leads to less geometry being drawn and therefore to a better performance. For small frame intervals, the inaccurately drawn geometry decreases. The worst case will be conducting occlusion tests in all frames. However, the increased test frequency can lead to the application's performance degradation due to the accumulation of each test cost.

3.2 OSG API Improvements

To achieve some of the proposed features of the occlusion culling algorithm it proved necessary to modify the API itself. The mainly amended classes were `osg::OcclusionQueryNode` and `osgUtil::CullVisitor`. The three most significant changes are presented and justified:

1. Definition of the interval between occlusion tests in time instead of frames. With the interval in frames, a simulation that can't reach the predefined frame rate will have less information to correctly determine the visibility. The same situation occurs in free-run simulations, where the oscillation in the number of generated images per second will also affect the number of occlusion tests performed.
2. The second change to the API is to detect when a node re-enters the view-frustum. In the current version, a node can leave and enter the view-frustum in different visibility situations without being tested. This is of particular relevance when temporal coherence is applied to occluded nodes and the frequency of view-frustum in/out nodes is high.
3. The third, and most important, amendment is to allow the definition of separate occlusion tests interval for visible and occluded OQNs. Currently the same interval of frames is used for both cases. With different intervals it is possible to tune the impact that the use of temporal coherence has in the application.

3.3 The Occlusion Culling Algorithm

The use of an interval between hardware assisted occlusion tests aims to reduce the negative influence they have in the application's performance. However, the absence of updated information to correctly determine visibility in each frame can have a more damaging effect, as discussed before. The OSG original method, in applying temporal coherence to previously occluded nodes, assumes that they tend to remain occluded for a period of time. If during this period the geometry becomes visible, is still not drawn. This has direct consequences in the correction of the image generated by not presenting the user objects that are actually visible. This loss of visible objects implies that this geometry is not used as occlusor.

This problem is addressed by applying in all frames the occlusion test to geometry determined as occluded in the previous frame. Thus, the total number of tests increases when compared to the original OSG algorithm. While the use of temporal coherence on occluded nodes impacts the amount of geometry wrongly drawn, when applied to visible nodes the consequence is the increase in unnecessarily drawn geometry. This strategy doesn't affect the image correction, only increases the amount of work sent to the graphics hardware. Thus, to improve the application's performance and generate a correct image, the number of tests to visible nodes must be optimized.

An approach to this problem was proposed by Kovalcik and Sochor in [19]. The authors propose that the interval between occlusion tests to visible nodes should be determined based on the number of times that a node was determined as visible. This idea was also later used by Mattausch et al. in the CHC++ algorithm [16] to find the occluded nodes candidates to be in a *multiquery*. Following this approach, the function (1) was defined to calculate the time interval between occlusion tests to visible nodes.

$$new\_time(i) = TB * (0{,}99 - 0{,}7e^{-i}) \quad \textbf{(1)}$$

The value *TB* defines the time range to use. This value should be adjusted according to the type of scene and the speed of the point of view. The parameter *i* represent the number of times that a visible node is tested and returns the same state of visibility. This parameter is reset whenever the node is determined occluded. The proposed function was adapted from the one described in CHC++ algorithm, which was obtained as an approximation to values collected during several simulations where the authors studied the visibility history of the nodes. Figure 1 illustrates the time intervals for *TB*=400 ms. In this example the initial interval is 116 ms which grows to 396 ms as the node continues to be visible. To find the adequate *TB* value several simulations where made and the results are presented in the next section.

The final proposed occlusion culling algorithm (Figure 2) incorporates the changes to the original OSG algoritm, the function that finds the time interval between tests to visible nodes and the test to occluded nodes in all frames.

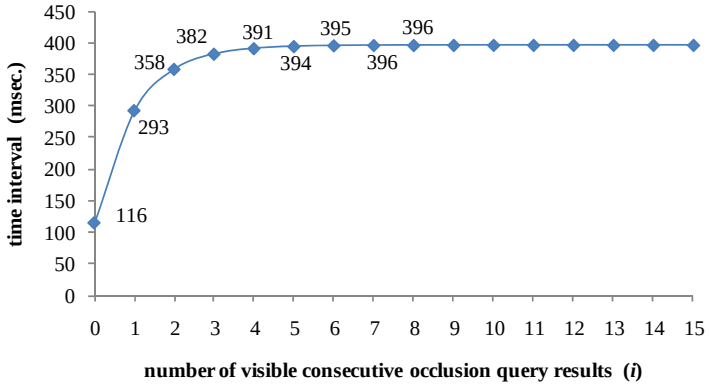


Fig. 1. Time intervals between occlusion queries applied to visible nodes for a *TB*=400ms.

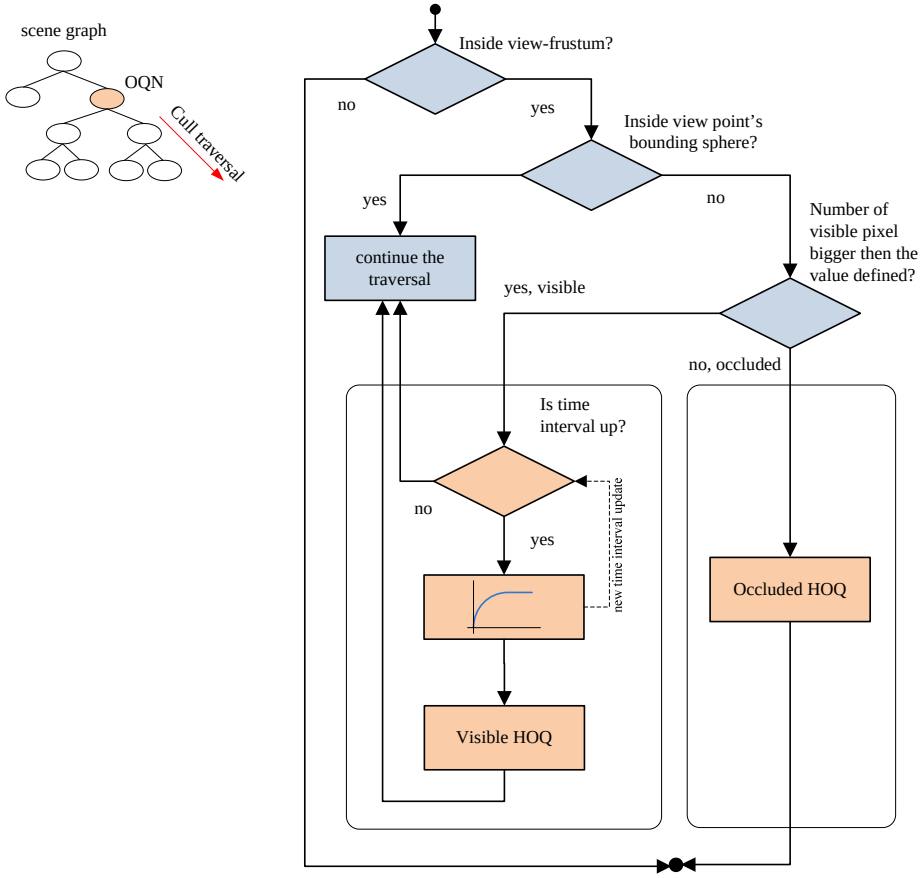


Fig. 2. Architecture of the proposed occlusion culling algorithm applied by the OQN during the Cull traversal of the scene graph.

4 Simulations and Results

The scene used for the simulations was set up like an urban environment where objects are distributed in an essentially flat surface but in this case, not aligned. This type of distribution provides more occluding diversity unlike a typical urban scene where the buildings are aligned with the streets. The hierarchical structure comprises four levels: root node, two group nodes levels and leaf nodes. To populate the scene two types of objects are present: buildings and small objects. The existence of buildings provides good occlusion but they are also the ones that should be removed when occluded since these are the most complex objects in the scene. Small objects are not good occluders. They are scattered and have low geometry complexity, so they are not directly occlusion tested because the cost of processing them is low. If they were tested individually, to each frame would be added the cost of the occlusion test without significant benefit when occluded.

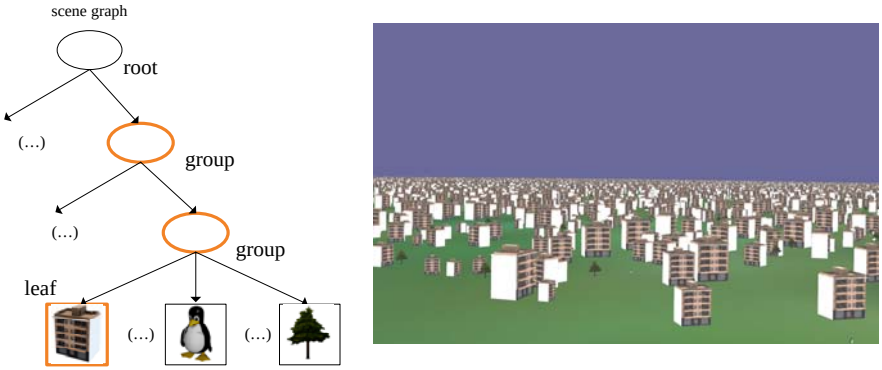


Fig. 3. The customized scene created to test the proposed method. The scene graph (left) where the orange contours represent the nodes tested in each graph level and a sample view (right) of the scene that demonstrates the depth complexity.

The scene (Figure 3) covers an area of 8,6 kms^2, has 25 426 buildings and 10 862 small objects. The total number of primitives is 45 527 127 to which correspond 160 950 781 vertexes. All simulation data was gathered during a walkthrough where the point of view moves through a previously recorded path. The route is traveled at constant speed and takes the point of view to areas with different depth and density geometry complexities. The simulations were made in a personal computer equipped with a Intel Core2 Duo 2.53 GHz CPU, with 3 GB of RAM, a NVidia GeForce 9650M GT graphics card (1 GB of memory) and the Windows Vista Home Premium (32 bits) installed operating system. To evaluate the performance of the software application the main criteria used was the frame calculation time measured for the duration of a constant (30 fps) frame rate simulation.

As described earlier, the OQN allows the setting of two parameters. One is the interval between occlusion tests. To find the best value for the *TB* parameter of function (1) thirteen simulations were made where different fixed intervals between tests to visible node were used. Figure 4 shows the results. The remaining OQN parameter is the number of pixels from which the tested bounding volume is considered visible. It was conservatively set to 1 pixel.

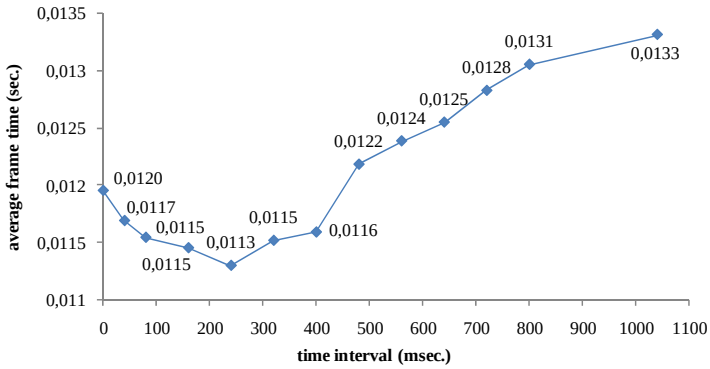


Fig. 4. Average frame length when used a fixed time interval between tests to visible nodes. Occluded nodes are tested every frame.

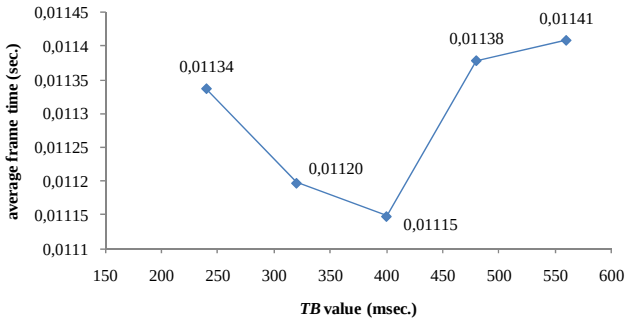


Fig. 5. Average frame duration for different values of *TB*. Also in this case the occlusion tests are applied to occluded nodes every frame.

For small time intervals between tests the performance degrades due to the increase in test frequency. By increasing the time interval the performance peak is reached (240 ms), which represents the ideal test period for this scene and point of view behavior. As the time interval was increased the performance goes down again due to the greater amount of geometry drawn. In this case, the decrease in the number of tests leads to less updated visibility information which affects the ability of the algorithm to remove geometry that is actually occluded.

To test the proposed algorithm that applies a variable time interval to visible nodes accordingly to the node's visibility history, a new set of simulations were made. Based on the results shown in Figure 4, the time interval *TB* should be selected around the peak performance time interval. Since the used scene has a lot of buildings moving between visibility states frequently, the *TB* values tested are those that quickly lead to an interval of around 240ms between HOQs. Figure 5 shows the results of these simulations, where the range of 400ms (recall Figure 1) is the one that allows the best performance.

To sum up, Table 1 compares some of the data obtained using different approaches to occlusion culling. The columns of the table refer to, from left to right: applying only the view-frustum culling; temporal coherence to visible and occluded nodes (with fixed interval between tests of 0,167 sec.); occlusion culling with fixed interval to visible nodes and in all the frames to occluded ones; the proposed algorithm that uses a variable time interval between tests to visible nodes.

Table 1. Some statistics of the several occlusion culling approaches used in the simulations.

Average:	View-frustum culling	OSG original (fixed interval to all nodes)	Fixed interval to visible nodes (240 ms)	Variable interval to visible nodes (400 ms)
Frame time (sec.)	0,1361	0,0096	0,0113	0,0111
Drawn vertexes per frame	44 679 143,7	449 119,5	435 496,5	435 644
Number of HOQs per frame	-	134,4	341,2	329,2

5 Discussion

While the original OSG approach emphasizes speed, applying hardware assisted occlusion tests to occluded nodes in all the frames allows a correct final image. The two methods presented (fixed and variable time interval), as an improvement to the original OSG algorithm, exhibit a very similar performance. Both achieve a reduction in the number of vertexes drawn at the expense of increasing the number of HOQs, as seen in the two rightmost columns of Table 1. However, the time taken per frame suffers when compared with the OSG method. The original method, despite drawing more vertexes, has a time per frame slightly lower due to fewer tests. The simulations with the original OSG were made using a time interval of 0,167 sec. that corresponds to a 5 frame interval in a constant 30 fps simulation.

In scenes with high geometry density the hierarchical use of HOQs provides significant gains in performance. The density of objects gives a good occlusion power, which allows that large bounding volumes can be determined as occluded. This means that with a single test large amounts of geometry could be culled in an early rendering stage. Also applying the occlusion test to geometry nodes enables that in the case of the above group node is visible, the algorithm can refine the visibility status of the child nodes.

The principle of temporal coherence has shown good results in the scene used. The application of this principle may be more problematic in dynamic scenes or scenes where the point of view moves faster. In these cases, the time that the result of an occlusion test can be considered valid varies with the nature of the scene in question, making it more difficult to generalize the results obtained. The proposed method of removing occluded geometry demonstrates a similar performance when compared to the OSG method, but avoids the loss of objects while minimizing the number of HOQs per frame.

Several strategies can be used to apply the proposed techniques in an adaptive way to perform occlusion culling in real-time graphical simulations:

- Preprocessing stage: in the case of driving simulations where the point of view often follows a predefined path, the 3D data structure could be analyzed in a preprocessing stage to identify the most often occluded portions of the scene. Along with the analysis of the geometry (complexity, size) and object density, the OQNs would only be inserted above the nodes containing those objects. Also the contribution of those objects to the final image could be taken into account. For instance, if an object is located far away from the road, the interval between occlusion tests could be widened while maintaining the level of detail when visible. In dynamic scenes the parameters of the OQN applied to moving objects also could be set to allow more occlusion tests.
- Runtime: in this case the main simulation parameter taken into account would be the speed of the view point. Adjusting the time interval between occlusion tests to static and moving objects accordingly with this parameter would allow a uniform experience during the simulation. The technique of adapting the level of detail based on the result of the occlusion test could also be used [20].
- Both: using the preprocessing and runtime approaches one unified method could be developed to achieve the best performance.

6 Conclusions and Future Work

This paper proposed an occlusion culling algorithm which exploits temporal coherence, in an alternative way of the original OSG algorithm. This algorithm manages the number of occlusion tests performed to visible nodes and performs tests in all frames to the occluded ones. By applying a function that increases the time interval between occlusion tests, as the node maintains the same visibility state, the number of occlusion tests decreases. As a consequence, also the negative effect that the OpenGL assisted occlusion determination mechanism has in the software application decreases.

To test the proposed algorithm an OSG API based software application was developed. To accomplish this goal several modification to API were necessary, as described in Section 3.2. Although simple, the three suggested modifications can be included in the OSG project in future releases so that the vast community that uses this API could benefit. Any of the suggested modifications can be included independently of one another.

Regarding future work, beyond the development of the unified adaptive occlusion culling algorithm, it's also intended to use the scenes used in the tests of some of the algorithms described in Section 2. This way, the comparison with the results published by their authors would be more feasible. It's also planned the implementation of the ideas described in this paper on the driving simulator Dris [21].

References

1. Hewlett-Packard (OpenGL extension specification), GL_HP_occlusion_test, 1997, http://www.opengl.org/registry/specs/HP/occlusion_test.txt
2. Scott, N., Olsen, D., Gannett, E.: An overview of the VISUALIZE fx graphics accelerator hardware. In: Hewlett-Packard Journal, vol. 49, nº 2, pp. 28-34 (1998)
3. Akenine-Möller, T., Haines, E.: Real-Time Rendering (2rd Edition), pp. 381-383. A K Peters Ltd, (2002)
4. NVIDIA (OpenGL extension specification), GL_NV_occlusion_query, 2002, http://www.opengl.org/registry/specs/NV/occlusion_query.txt
5. Gröller, E., Purgathofer, W.: Coherence in computer graphics. Technical report, Institute of Computer Graphics and Algorithms, Vienna University of Technology (1995)
6. Bittner, J., Wimmer, M., Piringer, H., Purgathofer, W.: Coherent Hierarchical Culling: Hardware Occlusion Queries Made Useful. In: Computer Graphics Forum (Eurographics 2004), vol. 23, nº 3, pp. 615-624 (2004)
7. Cohen-Or, D., Chrysanthou, Y., Silva, C., Durand, F.: A survey of visibility for walkthrough applications. In: IEEE Transactions on Visualization and Computer Graphics, vol. 9, nº 3, pp. 412-431 (2003)
8. Bittner, J., Wonka, P.: Visibility in computer graphics. In: Environment and Planning B: Planning and Design, vol.5, nº 30, pp. 729-756 (2003)
9. Teller, S., Séquin, C. H.: Visibility preprocessing for interactive walkthroughs. In: Computer Graphics (Proceedings of SIGGRAPH 91), vol. 25, nº 4, pp. 61-70 (1991)
10. Greene, N., Kass, M., Miller, G.: Hierarchical Z-buffer visibility. In: Computer Graphics (Proceedings of SIGGRAPH 93), pp. 231-238 (1993)

11.Zhang, H., Manocha, D., Hudson, T., Hoff III, K. E.: Visibility culling using hierarchical occlusion maps. In: Proceedings of SIGGRAPH 97, pp. 77-88 (1997)
12.Aila, T., Miettinen, V.: dPVS: An occlusion culling system for massive dynamic environments. In: IEEE Computer Graphics and Applications, vol. 24, nº 2, pp. 86-97 (2004)
13.Hillesland, K., Salomon, B., Lastra, A., Manocha, D.: Fast and Simple Occlusion Culling Using Hardware-Based Depth Queries. Technical report, Department of Computer Science, University of North Carolina at Chapel Hill (2002)
14.Staneker, D., Bartz, D., Straßer, W.: Occlusion Culling in OpenSG PLUS. In: Computers & Graphics, vol. 28, nº 1, pp. 87-92 (2004)
15.Guthe, M., Balázs, A., Klein, R.: Near optimal hierarchical culling: Performance driven use of hardware occlusion queries. In: Proceedings of Eurographics Symposium on Rendering 2006. The Eurographics Association (2006)
16.Mattausch, O., Bittner, J., Wimmer, M.: CHC++: Coherent Hierarchical Culling Revisited. In: Computer Graphics Forum (Proceedings Eurographics 2008), vol. 27, nº 2, pp. 221-230 (2008)
17.OpenSceneGraph, http://www.openscenegraph.org
18.Martz, P.: OpenSceneGraph Quick Start Guide. Skew Matrix Software LLC (2007)
19.Kovalcík, V., Sochor, J.: Occlusion culling with statistically optimized occlusion queries. In: Proceedings of WSCG (Short Papers), pp. 109-112 (2005)
20.El-Sana, J., Sokolovsky, N., Silva, C. T.: Integrating occlusion culling with view-dependent rendering. In: Proceedings of the conference on Visualization '01 (VIS '01), IEEE Computer Society, pp. 371-378 (2001)
21.Leitão, J. M.; Coelho, A.; Ferreira, F.N.: Dris - A virtual Driving Simulator. In: Proceedings of the Second International Seminar on Human Factors in Road Traffic (1997)

Design and Modeling of Road Environments

Carlos Campos[1], João Miguel Leitão[1,] Carlos M. Rodrigues[2],

[1]Instituto Superior de Engenharia do Porto
R. Dr. António Bernardino de Almeida, 471, 4200-072Porto Portugal
{crc, jml}@isep.ipp.pt
[2]Faculdade de Engenharia, Universidade do Porto
R. Dr. Roberto Frias, s/n, 4200-465 Porto Portugal
{cmr}@fe.up.pt

Resumo. A criação de ambientes rodoviários virtuais extensos, com o detalhe e o realismo adequados para simulação de condução, constituem uma das tarefas mais delicadas e consumidoras de recursos. Uma alternativa é possuir um conjunto de ferramentas que execute de forma automática essa tarefa. As metodologias apresentadas neste artigo permitem a utilização directa de especificações incluídas em projectos de estradas e de dados obtidos de Sistemas de Informação Geográfica (GIS). A partir da especificação do traçado, do perfil transversal e do perfil longitudinal, as ferramentas desenvolvidas produzem de forma automatizada um modelo 3D de uma via rodoviária ou de uma rede viária complexa. O modelo do terreno envolvente é também preparado de forma a permitir a sobreposição da estrada. Dos ensaios realizados, pode-se concluir que as metodologias apresentadas permitem a criação de ambientes rodoviários adequados à simulação de condução de forma rápida, eficiente e substancialmente menos dispendiosa de recursos.

Palavras-chave: *Driving simulator, 3D modeling, Road networks, Road environments.*

1 Introdução

Uma alternativa à realização de experiências de condução em ambiente real é a simulação de condução que, para além de um evidente aumento das condições de segurança, permite o controlo e monitorização de diferentes variáveis que seria inatingível em condições de tráfego real. Os simuladores de condução são cada vez mais uma ferramenta de estudo muito importante em áreas muito diversas, nomeadamente na psicologia, ergonomia e na engenharia rodoviária. Por outro lado, os simuladores de condução possibilitam a análise de situações de risco, impossíveis de estudar em ambiente real, e permitem o estudo isolado de um elevado número de variáveis que podem influenciar o comportamento dos condutores, garantindo-se sempre as mesmas condições de realização dos estudos experimentais. Independentemente do objectivo do estudo, a realização de experiências num simulador de condução exige a prévia preparação dos modelos dos ambientes rodoviários, podendo estes atingir facilmente várias dezenas de quilómetros. Estas grandes dimensões e os elevados requisitos de realismo, tornam a criação destes

ambientes rodoviários um processo muito complexo e dispendioso de recursos, caso não se possuam ferramentas adequadas [1] [2] [3].

O artigo está organizado de forma da seguinte forma: no capítulo 2 são apresentados os conceitos associados ao projecto de vias de comunicação. No capítulo 3 são apresentadas as metodologias automáticas de modelação visual de ambientes rodoviários. No capítulo 4 é feita uma análise crítica de resultados. Por último no capítulo 5 são apresentadas as conclusões e trabalho futuro.

2 Projecto de Vias Rodoviárias

Um projecto de uma estrada envolve várias especialidades de engenharia civil, tais como: traçado, terraplenagens, drenagem, pavimentação, sinalização, obras de arte e obras acessórias. O traçado geométrico da estrada é a primeira tarefa a ser realizada, estabelecendo-se a directriz, o perfil longitudinal e o perfil transversal tipo. Após este estudo desenvolvem-se as restantes especialidades, sendo geralmente os estudos conducentes à selecção dos dispositivos de sinalização a adoptar um dos últimos a ser realizados. A definição do traçado deverá obedecer aos critérios estabelecidos nas normas da Estradas de Portugal (EP), que é responsável em Portugal pela tutela da maior parte da rede rodoviária nacional.

2.1 Directriz

O traçado em planta de uma estrada tem como base o levantamento topográfico da área, sendo que a morfologia do terreno, a presença de outras estradas e as edificações existentes condicionam fortemente a directriz. Por outro lado, dever-se-á atender à velocidade de projecto, que ao fixar as características mínimas do traçado, influenciará igualmente a geometria da estrada. Com base em alinhamentos rectos implementam-se curvas compostas, constituídas por curvas de transição de raio variável e curvas circulares [4]. Deste modo, a directriz de uma estrada é constituída por um conjunto de troços rectos, curvas de transição e curvas circulares. Os arcos de transição vulgarmente utilizados no projecto são constituídos por curvas de raio variável designadas por clotóides.
Uma clotóide, ou espiral de cornú, é uma curva, definida matematicamente pela equação, em coordenadas polares:

$$S * \rho = A^2 \quad (1)$$

onde, S é a longitude do arco, ρ é o raio da curvatura e A o parâmetro da clotóide.

As curvas de cornú têm como função principal fazer a transição gradual e linear da variação da sobreaceleração centrífuga entre o troço recto e a curva circular.

2.2 Perfil Longitudinal

Após a definição em planta do traçado, define-se o traçado da estrada em altimetria. Este é condicionado geralmente pelo perfil longitudinal do terreno natural, tendo como objectivo que o traçado em perfil se aproxime o mais possível do terreno natural, de forma a minimizar as alturas de escavações e aterros e consequentemente os custos da obra. Tendo em conta o perfil longitudinal do terreno segundo a directriz, calcula-se o traçado em altimetria o qual é constituído por trainéis, segmentos rectos do perfil longitudinal, em rampa ou declive e concordâncias verticais. As concordâncias verticais podem ser do tipo côncava ou convexa e que permitem fazer uma transição gradual e suave entre trainéis de inclinação diferentes. As concordâncias verticais são definidas por equações do 2º grau, cujos parâmetros são geralmente o raio mínimo e as inclinações dos trainéis associados as concordâncias [4].

2.3 Perfil Transversal

A escolha do perfil transversal está associada quer à velocidade de projecto adoptada para a estrada quer ao tipo de estrada que se pretende construir. Após a definição do perfil longitudinal da estrada, e com base no perfil transversal adoptado, traçam-se os perfis transversais. No perfil transversal, encontra-se informação necessária para o traçado dos perfis, tal como as camadas do pavimento, a inclinação dos taludes de escavação e aterro, valetas, concordâncias de aterro, inclinações transversais, entre outras. Geralmente, o perfil transversal típico é composto por uma faixa de rodagem, com duas vias, e bermas. Em ambiente urbano, o perfil referido poderá ser complementado com passeios para peões. Estes últimos elementos de traçado, permitem definir com pormenor à escala 1/200, cortes transversais da directriz com equidistância, geralmente de 25 em 25 m ao longo da via. Por último, estes elementos são fundamentais para o cálculo e medição das terraplenagens, assim como, gerar vistas 3D, estáticas ou animadas da estrada [4].

2.3.1 Sobreelevação

A plataforma das estradas no sentido transversal não é plana, tendo duas águas em alinhamento recto e apenas uma inclinação em curva, a que se chama sobreelevação. A sobreelevação é utilizada para diminuir o risco de derrapagem e para melhorar a condições de drenagem de águas pluviais. A especificação da sobreelevação segue geralmente uma regra exposta na Norma de Traçado da EP, em que refere que no início da curva de transição a inclinação transversal da estrada deve ser a uma água e inclinada a 2,5% para o intradorso da curva, ou seja, para o lado interior da curva. No fim da curva de transição ou início da curva circular a inclinação deverá ser de valor Se (%) para o intradorso da curva. O valor de Se (%) é definido segundo o raio da curva circular e apresenta valor máximo de 7% [4]. A rotação para processar as variações ao longo da curva de transição é feita tendo como referência o eixo da estrada, no caso de estradas de duas vias.

2.3.2 Talude de escavação e aterro

Um talude é a superfície de terreno que se localiza junto a estrada e pode ser de origem natural ou artificial. Os taludes artificiais são criados durante a construção da estrada e identificam-se como resultantes de corte ou aterro no terreno. As inclinações dos taludes de escavação e de aterro são definidas em função de resultados dos estudos geológico-geotécnicos e têm como função garantir a estabilidade natural do terreno. Regra geral, os taludes de escavação e de aterro apresentam inclinações de 1 para 1,5 (V/H), de forma que a estabilidade do terreno esteja garantida [4]. Na figura seguinte, ilustra-se um perfil transversal to, composto por um aterro do lado esquerdo e uma escavação do lado direito, utilizado em projecto de engenharia de vias de comunicação.

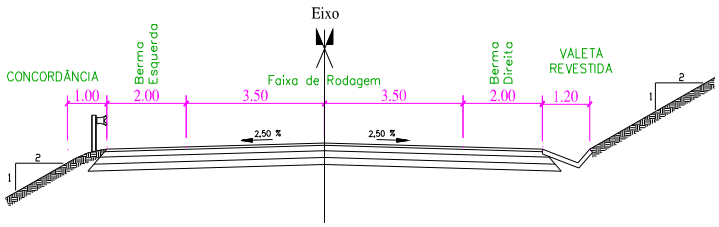


Fig. 1. Perfil Transversal

Se o terreno for menos estável, a relação do talude de escavação e do talude de aterro pode ser de 1:2, como representado na figura. Podemos ver do lado esquerdo a concordância de aterro, que é a superfície de terreno que faz a transição entre a berma e o talude, do lado direito a valeta, que serve para efectuar o escoamento das águas pluviais.

2.4 Sinalização Vertical e Horizontal

Uma das últimas tarefas a realizar num projecto de vias de comunicação é a execução do projecto de sinalização e segurança, onde se implementa os dispositivos respeitantes à sinalização horizontal e à sinalização vertical [5] [6] bem como são definidos os locais a dotar com guardas de segurança, atenuadores de impacto, ou outros dispositivos que concorram para uma circulação mais segura. Com a sinalização horizontal, ou seja, marcação no pavimento com pintura geralmente branca e reflectora, pretende-se guiar e orientar o tráfego, assim como alertar para situações de proibição de ultrapassagem. Fazem parte da sinalização horizontal a marcação longitudinal e transversal. Em termos sinalização longitudinal, faz-se referência geralmente às guias laterais e de separação de vias de sentidos. A linha axial pode ser do tipo contínua, descontínua ou mista. No que se refere às marcas transversais, estas podem ser por exemplo, barras de paragem aquando a presença de um sinal de STOP ou passadeiras para peões. A sinalização vertical é constituída por sinais de código (informativos, perigo, obrigatórios) e são geralmente colocados junto

às bermas das estradas de forma visível mas sem prejudicar o tráfego. O dimensionamento da sinalização vertical assim como a sua coloração está relacionado com a velocidade de projecto adoptada para a estrada. Na figura seguinte, ilustra-se uma colocação típica da sinalização vertical em meio urbano.

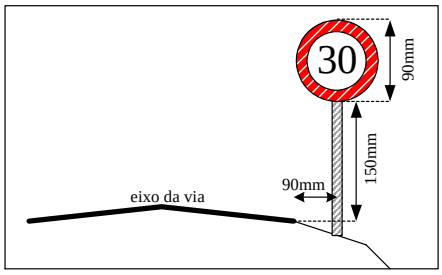


Fig. 2. Sinalização Vertical

Pode ver-se na figura a especificação da distância lateral à berma, a altura a que deve ficar o sinal relativamente ao pavimento e o dimensionamento para um sinal de proibição.

2.5 Zonas de transição

As zonas de transição, adoptadas sempre que é necessário considerar vias adicionais, consistem no alargamento ou diminuição da faixa de rodagem. Ao longo do percurso de uma via, pode-se encontrar troços em que o número de vias aumenta ou diminui. Sempre que numa via se introduz um separador, a transição do perfil transversal em plena via para o perfil transversal seguinte, deverá efectuar-se de forma suave, de modo a não obrigar os condutores a manobras bruscas. Esta transição deverá ocorrer através de uma curva e contracurva, de preferência com um alinhamento recto intermédio [7]. A extensão da zona de transição (ET) é determinada em função da velocidade de projecto e do alargamento do perfil transversal, podendo ser determinado pela seguinte expressão:

$$ET = \sqrt{a} * Vb \tag{2}$$

onde: Vb [km/h] - velocidade base; a [m] - metade do alargamento máximo do perfil transversal adicionado da berma interior.

A construção do alinhamento recto, ilustrada na figura seguinte, deverá será tal que o

alinhamento recto intermédio corresponda, no máximo, a 1⁄3 da extensão de transição.

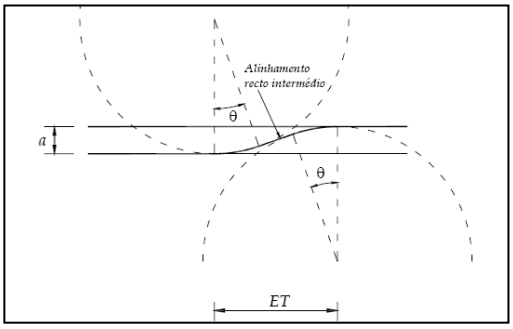


Fig. 3. Zona de Transição

3 Modelação de Ambientes Rodoviários

Para a realização de experiências de simulação de condução em tempo real, é neccessário criar os respectivos ambientes rodoviários. Esta tarefa pode ser muito morosa se não se possuir um conjunto de ferramentas que permitam construir de forma automática os ambientes rodoviários. Em engenharia de vias de comunicação utiliza-se uma directriz para definir o traçado longitudinal em planta (2D) da via. A directriz representa por onde vai passar a estrada ao longo do seu percurso longitudinal. A especificação do eixo da via no espaço tridimensional é obtida agrupando a definição de directriz que é definida no plano 2D e do traçado em altimetria que representa a cota Z. Neste trabalho os eixos das vias são tratados como polilinhas. Estas são representadas por listas ordenadas de segmentos. Em termos de concepção, um segmento de estrada pode ser recto ou curvo. No entanto, na implementação actual, admitem-se apenas segmentos rectos, existindo portanto a necessidade de decompor os arcos de circunferência e as clotóides em sequências de vários segmentos de recta. Desta forma, a polilinha pode ser representada apenas pelas coordenadas dos nós de ligação entre os segmentos adjacentes, como se pode ver na figura seguinte.

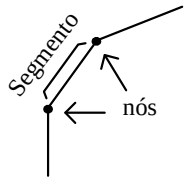


Fig. 4. Polilinha

Esta especificação permite obter facilmente uma base de dados que contém informação vectorial da via. Este tipo de informação é fundamental para consulta durante a simulação de condução. A determinação da distância de um ponto qualquer do espaço à polilinha é realizada sempre que é necessário conhecer, por exemplo, a posição de um veículo em relação ao eixo da via. A determinação do percurso sobre a polilinha, entre dois pontos da mesma, é útil para aferir a que distância longitudinal

que um veículo se encontra de um cruzamento ou de outro veículo. Na leitura da definição da via é mantida uma ordenação hierárquica dos segmentos que compõem a polilinha. Posteriormente foram implementadas funcionalidades optimizadas dos processos referidos anteriormente, que permitem uma aceleração acentuada dos cálculos a efectuar [8]. Na visualização, em caso de situações de recta com elevadas dimensões, esta pode ser definida apenas por dois pontos, nesse sentido implementou-se um filtro, que elimina pontos segundo um critério de erro, sem perda de qualidade do modelo [9].

3.2 Modelo da via

Tipicamente, a representação de superfícies em sistemas gráficos é aproximada por uma malha poligonal, sendo as mais comuns as triangle strip. Por analogia a uma estrada real em que esta é constituída por uma tira de pavimento ao longo do seu percurso, para representar graficamente um modelo de uma via utiliza-se uma tira de triângulos interligados entre si. A superfície da estrada será representada graficamente por uma tira de triângulos, como se ilustra na figura seguinte.

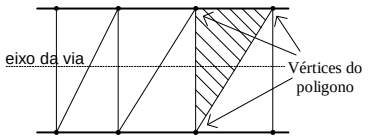


Fig. 5. Tira de triângulos

Em engenharia de vias de comunicação o construtor percorre o eixo da via, marcando pontos no terreno com intervalos de 25 metros em caso de recta e 25/3 em caso de curvas de raio reduzido. Para cada ponto, determina perpendiculares ao eixo da estrada e com a informação do perfil transversal constrói a estrada. Em construção real o problema da resolução de amostragem não é crítico, pois os pontos guia são calculados e a restante informação é interpolada. Em ambiente gráfico a tarefa de interpolação seria muito complexa, portanto parte-se de uma resolução dos pontos da polilinha superior ao utilizado em construção real, tipicamente de 1 metro. O procedimento para a geração do modelo poligonal da via é idêntico ao utilizado em terreno real. O modelo da estrada é construído percorrendo o eixo da via (polilinha) e em cada nó da polilinha, são traçadas perpendiculares ao eixo. Com a informação do perfil transversal e utilizando cálculo vectorial determinam-se os vértices dos polígonos que compõem a via. Para cada nó da polilinha, ao efectuar o cálculo vectorial é considerado o ângulo de sobreelevação da via. O ângulo de sobreelevação é especificado ao longo da via e para cada ponto da polilinha. Em situações de recta, não existe a necessidade de ter uma elevada concentração de polígonos para visualizar a via. Nesse sentido utiliza-se passo adaptativo para a visualização do modelo de estrada. Como acontece com outros sistemas [9], em recta, o número de polígonos gerados é inferior relativamente a situações de curva. Após obter o modelo poligonal da via, a etapa seguinte consiste em mapear uma textura sobre os polígonos, de modo a obter uma visualização da estrada mais realista, como se pode observar na figura seguinte.

Fig. 6. Modelo 3D da via

O processo de mapeamento consiste na especificação das coordenadas 2D de textura para cada vértice. A representação da sinalização horizontal segue a mesma metodologia descrita para o cálculo do modelo poligonal da via, sendo modelada apenas por polígonos com cor. Para a sinalização vertical, recorre-se a modelos 3D previamente modelados, que são colocados ao longo do ambiente rodoviário de acordo com a especificação. Para acelerar o processo de geração de imagem em tempo real, foi considerada a variação do nível de detalhe e a hierarquização espacial. Assim, o modelo da via é construído considerando a espacialização do modelo e a representação por variação de nível de detalhe. A hierarquia do modelo da estrada consiste em seccionar o modelo global da estrada em segmentos mais pequenos. À medida que os segmentos vão sendo construídos, estes vão sendo agrupados entre si, que por sua vez agrupa grupos de segmentos, construído a uma árvore hierárquica do modelo da via. Nos nós inferiores da árvore encontram-se os modelos de todos os segmentos de estrada considerados. Os nós superiores da árvore hierárquica são obtidos por agrupamentos sucessivos destes modelos. A cada nó é adicionada informação sobre o volume envolvente de todos os segmentos individuais. Outra optimização implementada consiste na definição de cada grupo dos nós inferiores, com representações de detalhe distintos. O nível de detalhe mais elementar consiste na visualização da via. O nível de detalhe seguinte considera a visualização da via e a visualização da sinalização horizontal. O último nível de detalhe considera-se todo o ambiente rodoviário. A função de avaliação implementada baseia-se na distância ao observador.

3.3 Zonas de transição

Na fase inicial da implementação deste módulo de tratamento de zonas de transição, por simplificação, estas são tratadas por transições lineares. Com a implementação de zonas de transição, o utilizador ao especificar o perfil transversal de uma via ao longo do seu percurso, tem a possibilidade de definir diferentes parametrizações. Para que o perfil transversal de uma zona ligue com o perfil transversal seguinte, o utilizador deverá especificar uma zona de transição. A extensão da zona de transição assim como toda a parametrização dos diferentes perfis podem ser totalmente definidos pelo utilizador. Nos testes realizados a extensão da zona de transição é de 100 metros. Este módulo, ao calcular o modelo geométrico da via, determina em função da posição longitudinal ao longo da via, qual a definição de perfil transversal. O cálculo das zonas de transição é realizado por uma transição linear, como se ilustra na figura seguinte.

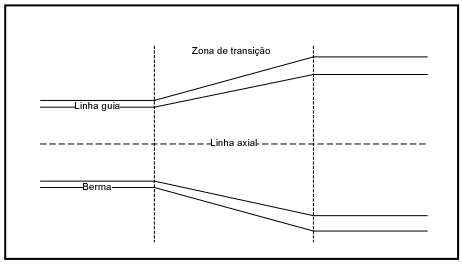


Fig. 7. Zona de transição

3.4 Alteração do Terreno

Na construção real de uma estrada, é comum percorrer o eixo da via, e traçar perpendiculares ao mesmo. Segundo essas perpendiculares determinam-se as alterações a efectuar nas cotas do terreno. A implementação segundo este método obriga a definir um passo ao percorrer a estrada. Em cada ponto da estrada teria que se efectuar uma pesquisa para identificar os pontos de terreno que deveriam ser alterados, esta tarefa poderia ser muito demorada ou até mesmo impraticável, caso o número de pontos de terreno fosse elevado. Este método, também iria causar inúmeros problemas de falhas e sobreposições, nomeadamente em situações de curva. Uma forma eficaz de efectuar a alteração do terreno mediante um traçado de uma Estrada consiste em para cada ponto do terreno, verificar se a sua cota é influenciada pela presença da estrada. Para isso, para cada ponto do terreno, calcula-se a distância mínima à estrada, e em função dessa distância, determina-se a necessidade de alterar a cota do ponto de terreno. As distâncias à estrada são calculadas utilizando as funcionalidades implementadas para geração do modelo da via. Se a distância do ponto ao eixo da via (Dp), for menor ou igual à largura da via (Dv), considerando a especificação de sobreelevação, então a cota do ponto é determinada pela seguinte equação:

$$Z_{ponto} = Z_{eixodavia} + D_p * Tan(\alpha) \tag{3}$$

Como se pode ver no exemplo da figura seguinte, o ângulo de sobreelevação (α) da via é positivo e a distância do ponto do terreno ao eixo da via (Dp) é menor que a largura da via (Dv). Neste caso, o ponto do terreno inicialmente com posição em Po, é deslocado para a posição Pf.

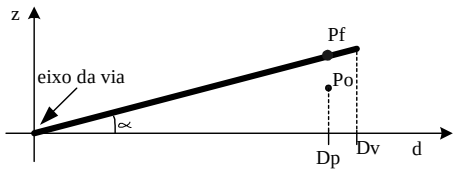


Fig. 8. Cota do terreno

Na prática, a via e colocada ligeiramente acima da cota do terreno para evitar problemas na detecção de visibilidade (ZBuffer). Quando a distância do ponto do

terreno ao eixo da via é superior à largura da via, então são consideradas as definições de talude de escavação e de talude de aterro. Na concepção é comum utilizar diferentes inclinações para a especificação destes tipos de taludes. Essas inclinações variam dependendo da morfologia do terreno, mais acentuada em terrenos rochosos e menos acentuados em terrenos macios. Os valores padrão de declive para este tipo de construções são de 1:1,5. A relação utilizada pela ferramenta para definir o talude de aterro e o talude de escavação é de 1:1,5, o que não impede a especificação de funções diferentes, e inclinações independentes para taludes de escavação e taludes de aterro. Para o caso de a distância do ponto do terreno ao eixo da via ser superior à largura da via, então tem que se avaliar se é necessário alterar a cota do ponto do terreno. A cota do ponto do terreno não é alterada se estiver posicionada entre as definições dos taludes, o que corresponde à área assinalada a vermelho da figura seguinte.

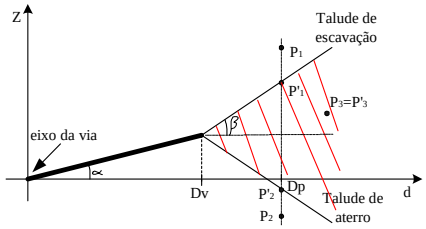


Fig. 9. Cota do terreno com talude

Para determinar se a cota deve ser alterada, são calculados dois valores possíveis para a cota do ponto do terreno, um para o caso do talude de escavação e outro para o caso do talude de aterro. A cota do ponto do terreno para o caso do talude de escavação é dada pela seguinte equação:

$$Z_{escavação} = Z_{eixodavia} + D_v * Tan(\alpha) + Tan(\beta) * (D_p - D_v) \quad (4)$$

Se a cota actual do ponto do terreno for superior ao valor calculado para o talude de escavação, então a cota é actualizada com o valor calculado para o talude de escavação. A cota do ponto do terreno para o caso do talude de aterro é dada pela seguinte equação:

$$Z_{aterro} = Z_{eixodavia} + D_v * Tan(\alpha) - Tan(\beta) * (D_p - D_v) \quad (5)$$

Se a cota actual do ponto do terreno for inferior ao valor calculado para o talude de aterro, então a cota é actualizada com o valor calculado para o talude de aterro. O processo de percorrer todos os pontos do terreno, garante que cada ponto só é alterado uma única vez. Mediante a definição de estrada e terreno utilizada, este método é o mais indicado, pois tem uma implementação simples e que não gera redundância na alteração da cota de pontos de terreno. Na figura seguinte pode-se observar talude escavação do lado esquerdo e talude de aterro do lado direito.

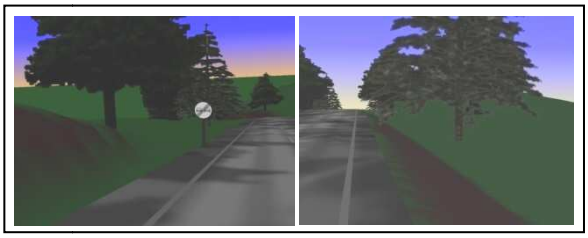


Fig. 10. Talude aterro e escavação

Num terreno em que o verde é a cor predominante, que correspondente à vegetação local, é natural que quando se efectue uma intervenção haja movimentações de terras e que seja alterada a cor natural do terreno. A cor pode transitar do verde de vegetação para castanho correspondente à terra movimentada. Para além de alterar a cota do modelo de terreno relativamente à passagem de uma estrada, é guarda informação associada ao modelo de terreno que possibilita alterar a cor das zonas onde existiu edição do valor da cota. À medida que os pontos vão sendo alterados devido à passagem da estrada, é associada uma etiqueta, que mais tarde serve para efectuar a alteração de cor [1] [2] [3].

4. Resultados

Para validar o desempenho das metodologias apresentadas foram realizados diferentes ensaios. Primeiro utilizou-se um modelador tradicional interactivo, *Multigene Creator*, recorrendo a um especialista da universidade de Leeds [10], para modelar uma estrada com cerca de 80 km. Em seguida utilizaram-se as metodologias apresentadas neste artigo, para modelar o mesmo ambiente rodoviário. Utilizandos as metodologias descritas, o tempo de criação do ambiente rodoviário é significativamente inferior, consumindo apenas 5% do tempo necessário utilizando o modelador tradicional.

5. Conclusões e trabalho Futuro

As metodologias descritas neste artigo permitem a criação rápida e eficiente de ambientes rodoviários de grandes dimensões. Os ambientes rodoviários gerados permitem a realização de experiências de condução simulada em diversas áreas científicas. Estas metodologias permitem obter modelos de excelente qualidade, reduzir significativamente o tempo e o custo de preparação de experiências. As técnicas de optimização utilizadas levaram à obtenção de resultados que satisfazem plenamente os requisitos da implementação. As metodologias de optimização adoptadas permitem eliminar uma percentagem significativa de polígonos a ser tratados no processo de síntese de imagem, não resultando em perda da qualidade de imagem. O ganho é significativo ao utilizar uma hierarquia na visualização do modelo de estrada, pois permite eliminar zonas extensas da estrada, que não são visíveis pelo observador. A utilização de diferentes níveis de detalhe permite obter melhores taxas de refrescamento, sem perda de qualidade da imagem [2] [3]. Apesar do actual estado de desenvolvimento das ferramentas apresentadas existem sempre funcionalidades

que podem ser melhoradas e novas funcionalidades que poderão ser implementadas, estando estas ferramentas sempre em constante desenvolvimento. Com a utilização frequente e continuada destas ferramentas é possível que sejam identificadas novas funcionalidades que ainda não foram equacionadas. O módulo de tratamento de zonas de transição poderá ser melhorado, passando a considerar não uma transição linear, mas também uma transição em curva e contra curva. Uma forma de optimizar o processo de cálculo para todos os pontos, é utilizar uma hierarquia do terreno, que possibilita eliminar zonas do terreno que não são afectadas pela passagem da estrada. O uso de hierarquia do modelo do terreno permite, no processo de alteração da cota dos pontos do terreno, eliminar o tratamento de zonas distantes que não são afectadas pela passagem da estrada. Uma nova aproximação as estas metodologias também pode ser equacionada, desenvolvendo uma nova abordagem para a modelação de ambientes rodoviários orientada para a modelação procedimental totalmente automatizada.

Agradecimentos

O trabalho desenvolvido no âmbito deste artigo teve o apoio do Laboratório de Analise de Tráfego do Departamento de Engenharia Civil da Faculdade de Engenharia da Universidade do Porto do qual é director o Professor Carlos Rodrigues.

Referências

1. C. Campos, V. Cunha e J. Leitão, Geração de Ambientes Rodoviários para Simulação de Condução, 12º Encontro Português de Computação Gráfica, p. 143-147, Outubro de 2003.
2. C. Campos, Geração de Ambientes Rodoviários para Simulação de Condução, tese para obtenção do grau de mestre, pelo Instituto Superior Técnico, em Novembro de 2006.
3. C. Campos, Modelação de Ambientes Rodoviários de Grandes Dimensões, 15.º Encontro Português de Computação Gráfica, Outubro de 2007.
4. JAE, Normas de traçado, Junta Autónoma de Estradas de Portugal, Actual Estradas de Portugal, Dezembro de 2010.
5. Manual de Planeamento das Acessibilidades e da Gestão Viária, Sinalização Rodoviária, C. Rodrigues, C. Roque, J. Macedo, Dezembro de 2008.
6. Manual de Marcação Rodoviária, C. Roque, Março de 2005.
7. Directrizes Geométricas Para Projectos de Estradas Nacionais - Intersecções, p.41 – 42, Dezembro de 2010.
8. J. Leitão, Agentes Autónomos Controláveis em Simuladores de Condução, tese para obtenção do grau de Doutor, pela Faculdade de Engenharia da Universidade do Porto, em Setembro de 2000.
9. S. Bayarri, I. Pareja, I. Coma e M. Fernández; Modelação de Carreteras para la Simulación de Conducción; 8º Encontro Português de Computação Gráfica, 1998.
10. University of Leeds. Leeds Advanced Driving Simulator, http://www.its.leeds.ac.uk/, Janeiro de 2011.

A Procedural Modeling Grammar for Virtual Urban Environment Creation

Pedro Brandão Silva[1], António Coelho[1],

[1] FEUP/DEI, INESC Porto, Rua Dr. Roberto Frias, s/n 4200-465, Porto, Portugal
{pedro.brandao.silva, acoelho}@fe.up.pt

Abstract. The creation of virtual urban environments, corresponding to real-world settings, has become a very frequent subject of content development for Virtual Reality Systems, serving multiple activities such as urban planning, virtual tourism and education, among others. Their extensive and complex nature, however, makes their conception very expensive and time-consuming, which has led many researchers for the look of semi-automatic methods for their creation. This paper describes the PG3D Grammar, a specification for modeling guidelines, which aims to provide a powerful control over the procedural modeling processes based on GIS data. This is achieved by allowing the integration of real-world data source queries and by providing management facilities to handle their large number, dimension, format and level of detail. The grammar can then be integrated into a procedural modeling tool for generating very large virtual real-world urban environments without further need for human intervention.

Keywords: Procedural Modeling, Shape Grammars, Spatial Engines, Virtual Urban Environments

1 Introduction

The creation of virtual real-world urban environments is a very demanding process, since it requires the design of large amounts of highly detailed contents, but unlike fictitious settings, they must be conceived according to real world information. Such tasks call for large modeling teams, long development times and, consequently, great production costs.

The employ use of procedural methods for generating three-dimensional content is becoming more frequent and has been delivering very interesting results at a lesser effort cost, by generating automatically (or at least, with much less human interaction) three-dimensional models.

Some guidelines and rules should be introduced by the user, capturing the knowledge about the modeling process, as well as the data sources containing existing real world information. In this sense, some methods have already been conceived, but, as far as it concerns the reproduction of real world environments, few of them address the challenge. These, however, offer only limited real data integration, especially when facing data sources with large dimensions, multiple formats or

incompatible spatial properties. On the other hand, since not all details of urban structures can be included in these sources, there is the need for additional stochastic and context-based inductions to fill the missing details. This motivates the development of more advanced methodologies which can not only manage multiple data sources, but fully enjoy the potential they may offer.

This paper aims to present the PG3D Grammar, a solution for defining rules and guidelines for procedural modeling tools. It enables the construction of large and detailed urban environments, based on various modeling procedures and on an iterative, conditional and characteristic growth, by relying on real information and the geographical relationship that exists between its elements.

This grammar has been conceived in the scope of a procedural modeling system named PG3D – *Procedural Generation 3D,* whose main characteristic lies on its implementation under spatial databases management systems. This type of platform serves firstly as a container for information sources, and secondly as a storage location for the created models, whereas these are operated by a set of stored procedures in the databases, thereby reducing the access time to data. This kind of approach is beneficial in the point that it allows spatial queries on the procedurally created data and therefore supports the contextual development of each urban element.

The paper is structured as follows: firstly, some related work will be presented, followed by a short background on the PG3D System, in order to understand the scope where the grammar was conceived and tested. Afterwards, the grammar features will be described, but not without explaining some important concepts first. Some information regarding the implementation will come next, followed by the results section, including the discussion. Lastly, the conclusions and some future work will be described.

2 Related Work

The work on procedural modeling requires always some type of rules and guidelines for modeling automatically large and complex objects and environments. Many interesting approaches for their design have already been proposed by some authors.

Parish and Müller used L-Systems, originally employed in the simulation of plant and organism growth [1], to generate extensive street networks [2]. The behavior of the development of an L-System can be parameterized and configured, allowing the control over the modeling processes.

Chen et al. suggested the manipulation of tensor fields to allow a more interactive control over road generation in cities [3]. In their CityGen system, Kelly and McCabe [4] introduced another interactive way to define primary and secondary roads, by allowing the change of their parameters while viewing their results in real time.

Regarding the modeling of buildings, Wonka et al. introduced the split grammars [5], a new type of parametric set grammar based on the concept of shape brought up by Stiny and Gips [6,7].He also presented an attribute matching system oriented by a control grammar, offering the flexibility required to model buildings with many

different styles and designs [5]. Based on this work, Müller et al. developed the CGA Shape [8], a shape grammar capable of producing extensive architectural models with high detail. The CGA Shape is a sequential grammar (such as the Chomsky Grammar[9]), therefore all the production rules are applied in sequence, in order to allow the characterization of structure [8]. The implementation of the CGA Shape is integrated in the CityEngine framework [10]. A simpler, yet interesting approach in this subject has been presented in Merrel's work regarding model synthesis [7], in which, by defining sets of valid combinations in example models of urban elements, more complex and diverse combinations can be created. This allowed a greater visual feedback as well as a more intuitive guideline construction.

To improve the process of creation of building façades, Müller presented a method which operates based on real world photographs [11], whose process can automatically create styles and ruling for the modeling processes. Finkenzeller, on the other hand, focused on more complex and less common façades that were not contemplated by the previous authors [12], increasing the necessary human interaction, but introducing greater control. The user provides the coarse building outline and style while the system generates the façade details.

Although these approaches are able to produce high quality fictitious urban environments, only a few are dedicated to the reproduction of real world urban environments, having therefore limited GIS data support. When correctly employed, these sources can serve as an extensive base and as a guideline for the procedural modeling processes. For this matter, Coelho presented in his work [13] the Geospatial L-Systems, an extension of parametric L-Systems which incorporates spatial awareness. This combines the ability of data amplification provided by the L-Systems (whose production rules are applied in parallel) with the geospatial systems, allowing spatial analysis of georeferenced data. This solution is integrated into a modeling tool called XL3D modeler, which provides interoperable access to various sources of information, allowing the reproduction of real urban environments.

3 The PG3D Concept

The recreation of large existing urban settings in virtual environments still presents several challenges to which many authors keep trying to answer. Procedural modeling based on GIS data has proven to be an effective approach; however their understanding is far from being trivial.

The name PG3D is an acronym for **P**rocedural **G**eneration **3D**[14,15]**.** Its fundamental design consists in performing procedural modeling directly on a spatial database management system where the geographic data sources (GIS) and the created models are saved. The modeling operations can be found in stored procedures, developed in programming languages of the database itself. A platform with PG3D modeling capabilities is called a PG3D modeler.

Since the objective lies on the modeling of real urban environments, the use of real world data is of utmost importance. However, due to the planet earth's ellipsoidal and irregular form, the use and interpretation of such data is difficult. Also, since multiple sources must be gathered, joined and organized, additional management problems

arise, especially if considering their usual large size. The use of spatial engines constitutes an important advantage, since it already includes advanced management features. Therefore the data, once loaded into the database, can be queried directly with just a few limitations in size, format or spatial manipulation.

By being implemented on a database, the PG3D modeler incorporates complex query abilities, allowing advanced spatial queries to be performed. This is especially valuable for the definition of modeling processes, which may act based on the relationship between the elements, instead of simply based on each individual's properties.

Being located in a database, the PG3D Modeler can be easily integrated into any platform, existing client or service that intends to enjoy from procedural modeling capabilities. Together, they do constitute a PG3D System [14,15].

4 The PG3D Grammar

The PG3D implementation in databases and their programming languages imposed the creation of adequate methods for procedural modeling. As a result, the PG3D Grammar was born, which intends to take most advantage of the PG3D approach.

PG3D grammars are an extension of shape grammars , more specifically, the CGA Shape, endowed with the capabilities of geospatial awareness and relational development, stemmed from Geospatial L-Systems[13,16]. Having been strongly influenced by both, it constitutes an attempt, where possible, to combine their greatest potential.

4.1 Important Concepts

Before discussing the structure of the grammars, some of its basic concepts must be defined:

PG3DShape. The PG3DShape (or "shape" for short) is the main data structure manipulated by all procedural modeling processes and therefore by the PG3D grammar. It is inspired by the concept of the shape grammar CGAShape [11] with the same name. Identified by a symbol, a shape can correspond to a surface street, building, or even just a portion, such as a wall or corner of a window (see Figure 1).

A shape contains a set of polygons, which in their turn contain vertices, each containing various properties (position, normal, color, texture mapping, etc.). The texture definition is held at the shape and is transitively applied to each of its constituent elements. Each shape is enclosed by a bounding box, called PG3DScope, which introduces quicker execution of spatial checks (intersections, unions) between shapes, and allows the user to set conditions on the shape's development based on its location and its dimensions. It also defines the shape's own axis and relative coordinate system, to facilitate the implementation of various modeling operations.

The PG3DShape is set in a hierarchy. As it shall be seen later on, shapes evolve by successive substitution. It is therefore possible for one to know its predecessor, i.e. the

shape from which it derives, allowing the query on shapes that descend from a particular predecessor.

PG3DLayer. A layer is a data structure that includes one or more shapes that share the same data source, therefore a common meaning, serving as a way to organize the shapes. A set of PG3DLayers define the starting point for a procedural modeling process, containing instructions on data load, to be used in further steps.

The geographic references contained in the layers' shapes make their relationship possible in such a way, as in the example below (see Figure 1), that can be combined information about the road networks and building footprints with surface information, fitting into each other in an appropriate manner.

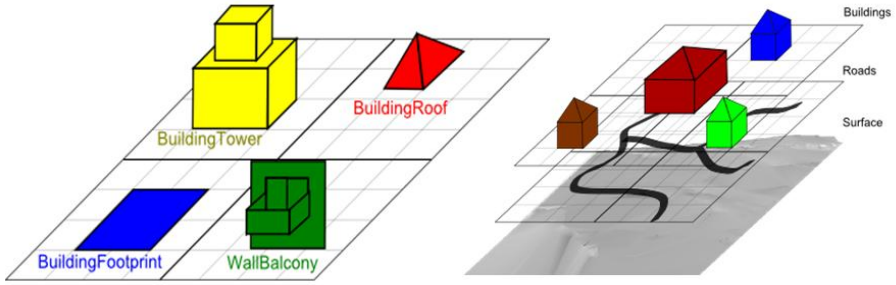


Fig. 1. Multiple kinds of PG3DShapes (left) may be created and related among multiple layers (right).

PG3DBoundary. In complex environments such as urban areas, different construction styles and layouts can be found throughout the city (greater variations can be seen between center and suburbs). For that reason, it becomes necessary to write distinct instructions for the development of each shape depending on their location. Since the precise definition of such variations is a laborious process, it is easier to achieve that by establishing a perimeter. This concept, called PG3DBoundary allows the specification of different modeling operations from one area to another.

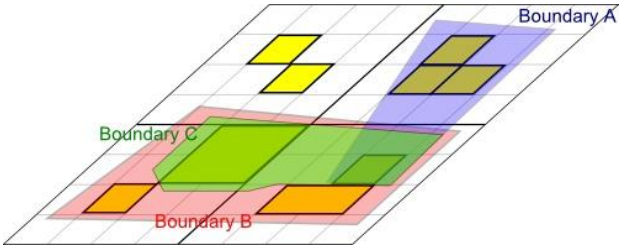


Fig. 2. Multiple intersecting boundary definition.

PG3DTag. A major difficulty that arises from handling such a large number of shapes, which may take various forms after the successive application of multiple transformations, is the management of their individual properties. It is therefore essential to create ways to select certain shape components based on their characteristics and perform operations only on these components. Although this can be partially achieved by accessing the property of each geometry, this is not always a simple task. In order to solve this problem, the concept of "tagging" components, after performing modeling operations on them, was introduced (see Figure 3, below)

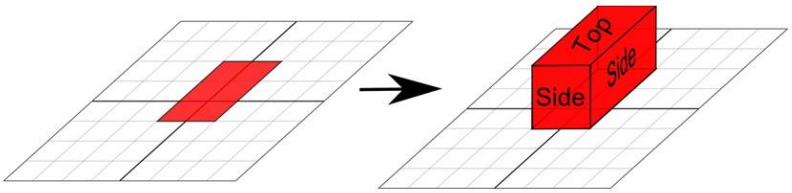


Fig. 3. Example of tagged shape parts according to their creation in an extrusion. Each side of the created volume will be tagged with labels such as "side" or "top", allowing further rule definition based on these tags.

4.2 Grammar Definition

The PG3D grammar can then be described by its constitution in the following elements:

- A set of PG3DShapes, which aggregates one or more elements with graphical representation. They may be considered terminals if no rule apply to them, or variables otherwise;
- The axiom, which defines the starting point of the modeling processes, defined by one or more PG3DLayers. Each layer, corresponding to a source of information, may also contain multiple PG3DShapes;
- A set of production rules, called PG3DRules, which define the modeling instructions, specifically the processes of replacement and development of PG3DShapes.

The procedural modeling process consists in successively replacing shapes in an iterative manner, starting with the axiom, and following the instructions contained in production rules. In the first iteration of the process, the layers that are part of the axiom (which indicate what sources of data should be loaded) are analyzed. Although all the information contained in each source can be loaded, it is often more convenient to work on smaller areas. The definition of such areas can be accomplished with PG3DBoundaries.

After this point, having already a set of shapes, the procedure is developed in both sequential and parallel way. At each iteration, for each shape created in the previous iteration, a production rule matching that shape is searched and applied. Once this is done for all shapes, it moves on to the next iteration which will handle the newly created shapes in the previous iteration. This process is repeated until no more rules are applicable to any shape or if an iteration limit, imposed by the user, has been

reached (see Figure 4). It is thus a process of **parallel** development (similar to L-Systems [1]), since all elements are replaced in each iteration. However, the production rules can be carried out in a composed manner, in the way that many operation calls can be joined in one production rule. This means that, inside the production rule definition, the process occurs **sequentially** (similar to Chomsky Grammars [9]).

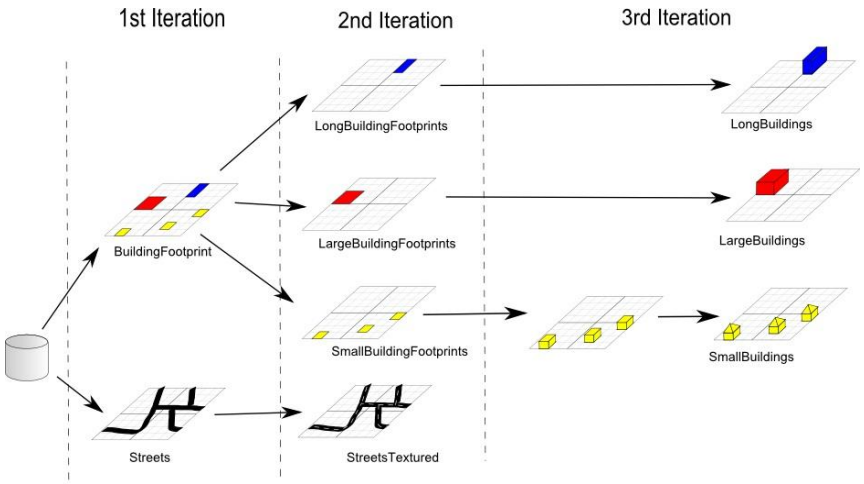


Fig. 4. Example of the iterative development of the modeling processes. At each iteration, the previous shapes are replaced or simply distributed among new shapes with new symbols, allowing specific rules to be applied.

The main idea behind the successive replacement of shapes is of progressive creation of detail, which means the models are incrementally improved. Thus, it is possible to generate simple, usable models in a first step, and evolve them according to available data sources and rules.

Each shape can be replaced by multiple ones, but it derives from one only, such as in a hierarchical structure. As a result, the modeling process creates a generation tree, whose nodes are shapes and its multiple levels correspond to the various iterations. This makes it possible to monitor the development after each iteration by simply querying all the leaves of the tree (nodes without children) at a certain tree level. Due to the progressive development of models, this can be extremely useful to generate various levels of detail of the same environment.

4.3 Production Rules

The production rules are the modeling instructions. Working in a similar way to the Stiny's shape grammars [7,6], the models evolve by consecutive shape substitution. They are structured in the following form:

$$\text{Predecessor} \rightarrow \text{Successors} \qquad (1)$$

In its most basic form, the definition of the predecessor consists of a symbol composed of alphanumeric characters and starting with a capital letter. This rule shall be applied to all shapes that have a matching symbol. The successor, in turn, can become quite complex, containing a sequence of modeling operations and symbols, which will mark each new shape creation. Each new shape will be the result of applying all previously called operations on the predecessor shape, and will act as its replacement. Consider the following example:

Shape1 → translate(vector3(0,10,20)) **Shape2** scale(vector3(5,5,5)) **Shape3** colorShape(rgba(255,0,0,255) **Shape4** extrude(20);

This rule will replace all shapes called "Shape1" by 3 different shapes. "Shape2" will correspond to a simple translation of "Shape1", "Shape3" will be 5-time bigger version of "Shape2", and "Shape4" will be equal to "Shape3", except red colored. The operations are therefore applied in sequence, while the shape naming "saves" these changes as new shapes, which will replace "Shape1". The result of the last operation, "extrude", will not be passed on to any shape, since there's no shape definition after it.

Special types of rules include:

- **Parametric Rules:** From one rule application to the next, it may be necessary to send certain information. For that reason, the production rules do accept parameter definition. These should be defined in the rule predecessor, after the symbol, and the parameter types should be declared. Rules can be overloaded (same as happens in programming languages like C++).
- **Conditional Rules:** The definition of rule conditions is essential for the development of shapes, especially when loaded with external information sources or when subjected to stochastic processes. The PG3D Grammar thus includes support for the IF…THEN…ELSE control structure. Conditions can contain also mathematical expressions and their boolean values can be negated or combined with AND and OR operators. It is also possible to nest conditions.
- **Stochastic Rules:** To compensate for the lack of information that some sources may contain, the use of randomness can become a form of data amplification, when used and controlled effectively. The parameterization can be achieved through specific expressions provided which operate based on probabilities, which can control, for instance, the frequency of appearance of a particular characteristic in a certain environment.
- **Parenthetic Rules:** By using a stack it is possible to save and load "states", i.e. results of transformations over shapes. By using PUSH, the current state is saved on a stack, and loaded from it when POP is used. The use of the name "parenthetic" derives from its concept from L-Systems, which use the "[" and "]" operators to achieve this goal, but to avoid conflicts with array definitions, it has been replaced in PG3D.

Other important features for rule creation are:

- **Attributes, Limits and Textures:** Using the same values in more than one production rule is a common practice. Therefore, it is possible to declare constants once only, and use them multiple times in rules.

4.4 Transforming, Reading and Accessing Real Data and Geometries

The first step in using geographical information sources is to understand their format and how to access them. PG3D considers GIS Data which follows the "Simple Features" specification[17,18] by the Open Geospatial Consortium, meaning that it stores primitive geometrical data types, such as points, lines and polygons. In the process of reading data, only two fields are loaded: the geometry and a primary key that identifies the record uniquely. Its existence allows, on one hand, to operate always on geometries (which must have spatial information) and, secondly, to own a reference to each record that was read. This allows the query of other fields of the record (text information, other numeric values, etc.).

4.5 Conditional and Characteristic Development

One of the major features of the PG3D Grammar is its ability to conditionally develop each shape (or any of its parts, namely vertices or polygons) based on its properties. In other words, the evolution of a shape may depend on its current status or on the corresponding data that may exist in the data sources.

The control structure IF...THEN...ELSE is one way of checking these properties. The other exists at function level, in the way the some support a boolean parameter, allowing the operations to be applied only to parts of a shape fitting a certain condition. To indicate the intent to act based on the state of the shape or its vertices or polygons, PG3D supports the three reference symbols starting by the percentage sign: *%s, %p, %v* for shape, polygons and vertices, respectively.

4.6 Geospatial Awareness

The concept of contextualization in geospatial PG3D grammars derived from its application in Geospatial L-Systems [13]. Its main idea consists in developing shapes not independently, but based on their surroundings, avoiding the creation of unrealistic structures. This is especially important, for instance, when performing occlusion tests.

Suppose a couple of buildings share common walls: these should not contain any windows, balconies or front doors. Since all the elements are spatially referenced, such check is easily achieved (see Figure 5).

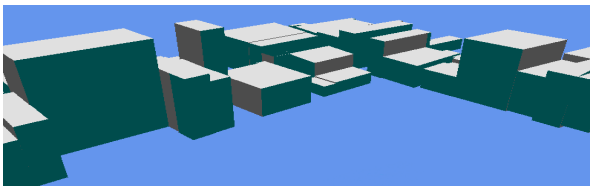


Fig. 5. Colored façades demonstrate shows building walls which do not have any other wall in front of them at a distance lesser than 3 meters away. This makes them potential good façades for buildings.

4.7 Selectivity

A simplified form of working on shapes is by decomposing them into smaller and less complex shapes, manipulating them in the next iteration. To allow this type of action, PG3D uses selection operators, also called "selectors". It is similar to a simple shape definition, but allows filtering by conditions on polygons (see Figure 6).

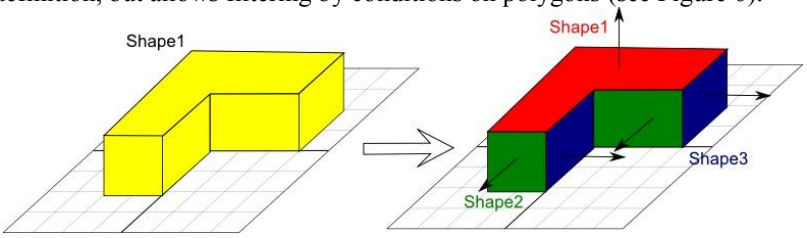


Fig. 6. A shape's polygons are selected and distributed among 3 different shapes, based on their normal direction along the x, y and z axis.

5 Implementation

This paper focuses on the PG3D Grammar, a more conceptual topic. Yet, its application is quite concrete in the scope of the PG3D Modeler. A prototype of this has been developed in order to validate and evaluate the concept, especially regarding the rule production definition through the PG3D Grammar.
As stated before, the main idea of the PG3D modeler relies on its implementation in spatial database systems, enjoying therefore their storage and query capabilities. This component is therefore responsible for:

- Storage of the geographical data sources;
- Storage of the many modeling parameters and created results;
- Execution of the modeling processes.

PG3D's full name is also related to the technology on which this first prototype was based. PG3D is therefore an acronym for the triple **P**ost**g**reSQL, **P**ost**GI**S, **P**rocedural **G**eneration **3D**. The database management system chosen was PostgreSQL, using the PostGIS spatial extension, which allows not only the load and storage of geographic information sources, but also the execution of spatial operations in the created data structures. The chosen programming language was PL/PgSQL, which takes advantage of its native implementation in the database to increase the power, flexibility and performance of the created functions.

6 Results

The PG3D Grammar was tested in the PG3D Modeler, serving as the main method of control over the modeling processes. It is therefore necessary to test its potential, simplicity and ease of use.

An important factor in modeling is the quality and level of detail of the created elements. To be able to achieve certain levels of precision, it is necessary that the modeling processes are capable of reproducing existing buildings with enough detail, and reproduce it over hundreds of thousands of buildings. The grammar is indeed powerful enough to respond to such requirements, as it can be seen in Figure 7.

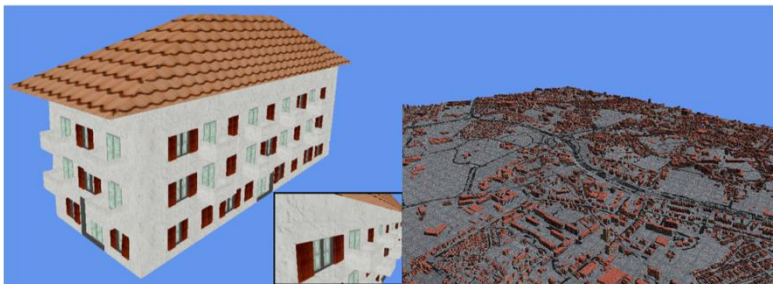


Fig. 7. (On the left) Close-up of a model of a house, featuring high detailed windows, shutters and balconies. This can be reproduced in large-scale environments (on the right).

The presented model shows that the various elements of the building façade of a house can be defined with great detail. The design of such structures, however, needs the definition of a larger number of production rules, which, in turn, take more time to process, especially when considering larger-scale environments.

In order for the urban environments to be recognized, is it important that its elements hold as many similarities to the real ones as possible. This is achieved mostly through the distribution of roads and buildings, but also through smaller details, such as façade outline and contained features. Whereas the PG3D Grammar does support intensive real GIS integration (see section 4.4), due to the limited amount of data sources used until this moment, only the first perception has been achieved (see Figure 8), so some work in this area is still being developed.

Fig. 8. Comparison of the Boavista Roundabout[1] in Porto, Portugal and its correspondent virtual environment.

[1]Source: http://www.architecture-page.com/assets/images/content/prj_oma_casa/1.jpg, 28-10-2010

A main problem that the PG3D Grammar presents is the fact it is declared through text, lacking an instant visual feedback of its effects. For this reason, it may become difficult to create models such as presented in Figure 7, which requires almost 20 rules. However, for simpler cases, such as the one presented in Figure 8, 5 rules were enough to describe such an extensive and complete environment.

7 Conclusion and Future Work

The PG3D Grammar, as well as the PG3D Modeler, constitutes work still under progress, yet it provides already very interesting and powerful features regarding the procedural modeling of virtual urban environments, corresponding to real-world settings, such as:

- Definition of layers for GIS Data and complex manipulation through shapes;
- Definition of various types of rules: parametric, conditional, stochastic, parenthetic as well as attribute, boundary and texture definition;
- Fine manipulation of polygons and vertices;
- Conditional and characteristic development, allowing not only the shape development based on their individual properties, but also on their surrounding elements (geospatial awareness);
- Possibility to create multiple levels of detail of the same model through the iterative development of the shapes;
- Easy implementation with any spatial database platform, since it does not rely on any specific technology.

By being implemented on a database, there are still many possibilities to be further researched, based on complex queries that such technologies allow. Many features are still to come, and the PG3D Grammar is yet to achieve its full potential with this approach.

One of the main issues to be addressed in a future work is the creation of more visual interfaces for their edition or, instead, to provide automatic rule creation based on example models, photographs or existing virtual environments. Other simpler approaches consist in allowing rule libraries to be created and shared among users, in order to provide easy style and architecture definition and reuse.

8 Acknowledgements

This work is partially supported by the Portuguese government, through the National Foundation for Science and Technology - FCT (Fundação para a Ciência e a Tecnologia) and the European Union (COMPETE, QREN and FEDER) through the project PTDC/EIA-EIA/108982/2008 entitled “3DWikiU – 3D Wiki for Urban Environments".

References

1. Prusinkiewicz P, Lindenmayer A (1996) The Algorithmic Beauty of Plants. Springer-Verlag,
2. Parish YIH, Müller P (2001) Procedural Modeling of Cities. (SIGGRAPH 2001):301–308
3. Chen G, Esch G, Wonka P, Müller P, Zhang E (2008) Interactive Procedural Street Modeling.
4. Kelly G, McCabe H (2007) Citygen: An Interactive System for Procedural City Generation.
5. Wonka P, Wimmer M, Sillion F, Ribarsky W (2003) Instant architecture. Paper presented at the ACM SIGGRAPH 2003 Papers, San Diego, California,
6. Stiny G (1980) Introduction to shape and shape grammars. Environment and Planning B 7 (3):343-351
7. Stiny G, Gips J Shape Grammars and the Generative Specification of Painting and Sculpture. In: Friedman CV (ed) Information Processing '71, 1972. pp 1460-1465. doi:citeulike-article-id:1526281
8. Müller P, Wonka P, Haegler S, Ulmer A, Gool LV (2006) Procedural Modeling of Buildings. Paper presented at the ACM SIGGRAPH 2006 Papers, Boston, Massachusetts,
9. Chomsky N (1956) Three Models for the Description of Language. (IRE Trans. Information Theory (2),):113–124
10. Procedural Inc. (2009) 3D Modelling Software for Urban Environments. http://www.procedural.com/.
11. Müller P, Wonka P, Zeng G, Gool LV (2007) Image-based Procedural Modeling of Facades.
12. Finkenzeller D (2008) Detailed Building Facades.58-66
13. Coelho A, Bessa M, Sousa AA, Ferreira FN (2007) Expeditious Modelling of Virtual Urban Environments with Geospatial L-systems. Computer Graphics Forum 26 (4):769-782
14. Silva PB, Coelho A Procedural Modeling of Urban Environments for Digital Games Development. In: ACE'2010: 7th International Conference on Advances in Computer Entertainment tecnology, Taipei, Taiwan, 2010. ACM,
15. Silva PB, Coelho A Procedural Modeling for Realistic Virtual Worlds Development. In: SLACTIONS, 2010.
16. Coelho A (2009) Procedural Modeling of Buildings - Automating Virtual World Creation with Procedural Techniques.
17. Open Geospatial Consortium (2010) Simple Features. http://www.opengeospatial.org/standards/sfa.
18. Open Geospatial Consortium Inc. (2006) OpenGIS® Implementation Specification for Geographic Information - Simple feature access - Part 1: Common Architecture. In: Herring JR (ed)

SESSION 5

COMPUTER GRAPHICS / MOBILE COMPUTING

Chairman: Vítor Manuel Rodrigues da Cunha

Hybrid methodology to segment skin lesions based on active contour and region growing techniques

Alex F. de Araujo[1], Aledir Silveira Pereira[2], Norian Marranghello[2], Ricardo Baccaro Rossetti[3], João M. R. S. Tavares[1]

[1] Faculty of Engineering, University of Porto (FEUP) / Institute of Machanical Engineering and Industrial Management (INEGI), R. Dr Roberto Frias s/n, 4200-465 - PORTO, Portugal
fa.alex@gmail.com, tavares@fe.up.pt
[2] Departamento de Ciências da Computação e Estatísticas (DCCE) / Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, R. Cristóvão Colombo, 2265, 15054-000 – São José do Rio Preto, Brazil
{aledir, norian}@ibilce.unesp.br
[3] Clinica DERM, Av. José Munia 5950. 15054-000, São José do Rio Preto – SP – Brazil
rbrossetti2010@hotmail.com

Abstract. Computer aided methods are ever more widespread in medical applications. For instance, skin lesion diagnosis benefits from such methods for automatic contour extraction. In this work, we propose a hybrid medical image segmentation method to detecting and extracting skin lesion contours. To obtain accurate segmentation results, the region growing technique, based on a Quadtree implementation, is used to extract the initial contour that is then refined by using a traditional active contour model. Experimental results indicate that the proposed computational approach is promising, being able to detect skin lesion areas and to extract their contours from images, maintaining the original irregularity that is a characteristic usually used in the medical diagnosis of these lesions. Additionally, from the experimental findings, it is possible to conclude that the extracted contours adequately represent the lesions, being well adjusted to their borders and retaining their main features.

Keywords: Medical Image Segmentation, Region Growing, Active Contour, Skin Lesions.

1 Introduction

Skin lesions are frequent and can indicate serious diseases, such as skin cancer. Usually, the initial diagnosis of those lesions is obtained from image analysis. From the visual analysis of the lesion regions, the dermatologist is able to give an initial diagnosis about the lesions that is very important to define an efficient treatment plan. However, some factors, like the fatigue of the dermatologist, the small dimensions of the lesion and the interferences caused by noise and reflexes, can make medical diagnosis a difficult process eventually resulting in uncertain outcomes. Computational methods for medical image processing have been extensively investigated, and several solutions have been accomplished to assist medical

professionals in the diagnosis and follow up of diseases from images in a fast and accurate way. Some characteristics commonly considered by dermatologists in the analysis of skin lesions from images include: irregularity of their borders, the asymmetry of their shapes, the color variation along their internal areas and their dimensions [1]. Hence, image processing and analysis techniques can be used to help dermatologists on their diagnosis by extracting lesion contours from images and characterizing the lesions undervaluation.

For a successful computational diagnosis of skin lesions from images, the extracted contours should preserve the original irregularity. In an attempt to extract the contours of skin lesions from images maintaining their original irregularity, we propose the application of the region growing method, followed by the application of the active contour method proposed by Kass, Witkin and Terzopoulos [1, 2]. Thus, the region growing method was applied to pre-segment the input image and extract an approximate contour of the skin lesion presented. Then, this rough contour is used as the initial curve for the method of active contours.

The region growing was implemented using a Quadtree algorithm adopting as the growing control parameter the average value between the maximum and minimum of the intensity component of the Hue-Intensity-Saturation (HIS) color space. After the execution of this algorithm, a merging algorithm is applied, based on the color intensity of the lesion areas, to merge the illness regions, making possible the extraction of the initial lesion contour. The topology of this initial and rough contour is similar to the one of the lesion border. It then needs to be further refined to conveniently represent the desired border. For this refinement, the traditional method of active contours is used, considering as the initial curve the resulting contour from the merge algorithm.

This paper is organized as follows: In the next section, works related to image segmentation are presented. In section 3, the methodology considered and the method developed are described. A discussion on the experimental tests as well as on the corresponding results is included in section 4, followed by the conclusions and suggestions for future works.

2 Related Works

Generally, medical image segmentation techniques aim to detect structures, such as organs, lesions, tumors and tissues, represented in images as well as to extract their contours in an efficient, robust and automated way. Researches in this area have been looking for methodologies able to successfully segment images that are corrupted by noise and other interferences, maintaining the main characteristics of the original borders. Another feature commonly searched is the automation of the methodologies in order to avoid external interventions and subjectivity. There are segmentation methods based on many different concepts and techniques, such as, image thresholding, region growing, Fuzzy logic, genetic algorithms, artificial neural networks and models of active contours.

Image segmentation methodologies based on thresholding separate the regions of the desired structures from the image background, using values (thresholds) as

classifiers of a particular feature. For example, in a simple thresholding of a grayscale image, one gray level (threshold) is chosen, then all pixels with intensity higher than this level are classified as belonging to the structure and the other pixels are classified as image background [3].

In order to merge the higher number of pixels into regions, methods of region growing have been proposed. One approach frequently used is to divide the input image in sets of disjoint regions, such as the Quadtree method performs. This method, known as split and merging method, consists in dividing recursively the input image in quadrants, until a parameter *P* is true. Usually this parameter is based on the level of color intensity of pixels in each quadrant, the median of intensity for example. Thus, a tree is built, where each non-leaf node possesses 4 (four) child nodes, which are merged according their similarity [3]. Lee and collaborators [4] presented a methodology to extract features from ultrasound images, using a region growing method (k-means) to join the pixels into interesting regions.

Unlike the traditional segmentation methodologies, the ones based on fuzzy logic try to define a level of pertinence to establish if the image regions belong to the background or to some desired structure. This level of pertinence is not binary, in other words, is not "yes" or "no", as it can assume more states [5]. The Fuzzy C-Means algorithm (FCM) and its variations are the most used methods for image segmentation based on fuzzy logic [6, 7].

Genetic algorithms (GA) have been used in the segmentation of images with diverse characteristics [8-10]. GA use some functions, known as genetic operators, to generate new populations from an initial one, in order to produce the fittest individuals. The most common operators are the crossover and mutation. The former recombines the parent features during the reproduction process, resulting in the inheritance of features over the generations. This operator can be implemented in different ways: such as one-point-crossover, where a point is selected to divide each parent chromosome in two parts, say a lower portion and a higher one, then the genetic information (genes) of the parents is exchanged in order to match the lower genes of one parent to the higher genes of the other parent; and multipoint-crossover, and multipoint, where the genes are changed considering more than one cut point. The mutation operator acts in the maintenance of genetic diversity of the population, avoiding the local minima either, these result in false-positives results.

Artificial neural networks (ANNs) try to simulate the human brain operation to interpret and solve computational problems. Techniques based on ANNs have been extensively used in medical image processing and analysis. Stanescu and collaborators proposed a methodology to medical image segmentation based on ANN [11]. In this approach, the color system of original image is transformed from RGB (Red, Green, Blue) to HSV (Hue, Saturation, Value), and the back-propagation algorithm is performed to decompose the image in regions. Iscan and colaborators presented in [12] the use of an incremental neural network for image segmentation. The methodology provides the updates for weight to nodes during training of the network, improving the performance of the regions classification.

With the goal of developing segmentation methods more accurate and able to realize acceptable detection of irregular object borders, several techniques of image segmentation based on the active contour model proposed by Kass, Within and Terzoupoulos [2] have been proposed. Usually, these methods start with an initial

closed contour, defined within the image domain, and deform it into the desired border by the action of internal and external forces applied on the contour. This deformation is obtained by the minimization of the contour energy, and the contour energy is minimal when the deformed contour is over the feature to be segmented. The use of segmentation methods based on active contours has been extensively explored in medical image processing [13-15].

The recent developments regarding image segmentation indicate a tendency to merge different techniques [16, 17]. These hybrid methodologies have gained special attention because of their ability to produce more accurate results, as well as to process more complex images. By combining the best characteristics of two or more techniques, these methodologies try to overcome some difficulties such as nonhomogeneity, noise interferences, pose variations or occlusions.

3 Proposed Method

Skin lesions present different shapes and usually suffer from interferences caused by noise and artifacts, such as hairs, bubbles and light interferences, making difficult their visual characterization. The efficient extraction of lesion borders is crucial to facilitate their diagnosis by dermatologists. In this work, a hybrid method is proposed to segment skin lesions from images, which uses the methods of region growing and active contours. In Fig. 1, the flowchart of the proposed method is presented.

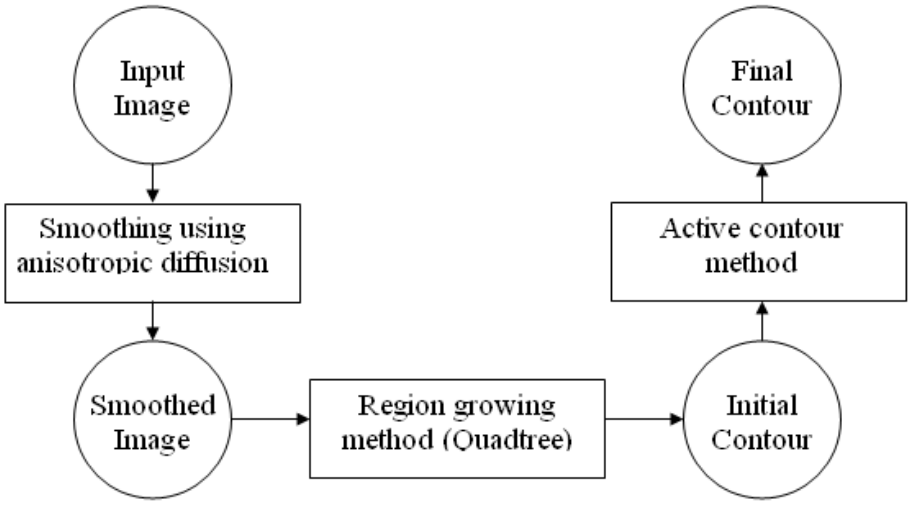


Fig 1. Flowchart of the proposed segmentation method.

Due to the usual noise and artifacts that normally corrupt dermatologic images, the input images are pre-processed using the non-linear diffusion filter [18]. Then, the images are segmented by using the method of region growing; in particular, a Quadtree region growing method is used to obtain the initial lesions contours. In this

initial segmentation, the images are transformed from the RGB (Red, Green, Blue) space into the HIS (Hue, Saturation, Intensity) space [3]. The growing parameter adopted is the average between the maximal and minimal intensities (*I* component in the HSI space) of the pixels belonging to a quadrant. This step allows for the input smoothed images to be represented by a set of homogeneous regions, facilitating the detection of skin regions that are potential ill. To merge the several segmented regions and isolate the image backgrounds, a merging method is applied considering the distance between the regions intensity and grouping the regions with similar intensities. The Quadtree method was adopted because it does not require the division of all quadrants until the lowest level is reached; in other words, the quadrants that represent one region are not further divided, and so the associated computational cost is reduced. The result of this step is a binarized image, which turns expeditious the extraction of approximated contours for the lesions by evaluating the pixels intensity. These contours are then better adjusted to the lesion borders by using a method of active contours. Two methods of active contours were studied: the traditional snake method [2] and the Gradient Vector Flow (GVF) method [19]. The used parameters related to elasticity and rigidity of the contour and the weight of the external forces applied are indicated in Table 1 for the two methods.

Table 1. Parameters of elasticity, rigidity and weight of the external forces used in the two methods of active contours tested.

Model	Elasticity	Rigidity	Weight of external force
Traditional	0.05	0.05	0.1
GVF	0.05	0.05	0.1

In the test images, the initial contours obtained by the region growing method were almost coincident with the lesion borders and the GVF method generated higher deformations on the initial contours than the traditional snake method. Thus, the traditional snake method model was chosen to adjust the initial contours obtained by the region growing method to the skin lesions.

4 Experimental Results

20 color images, with 256x256 pixels, from the image database Dermatlas [20] were used for the tests. This image set presented atypical, malignant and no-malignant lesions, and include images with good contrast, low contrast, and corrupted with different quantities of noise and artifacts, such as hairs and light reflexes. The obtained segmentation results were validated by an expert on dermatologic lesions from the Clinica DERM, in São José do Rio Preto – São Paulo – Brazil.

In Fig. 2 some segmented images are presented: The original images (a and c) were processed by the proposed method and the final contours presented in the images (b and d, respectively) were obtained. All final contours were visually evaluated by the expert whom confirms that all ill regions were successfully detected and that approximately 90% of the segmented contours were acceptable.

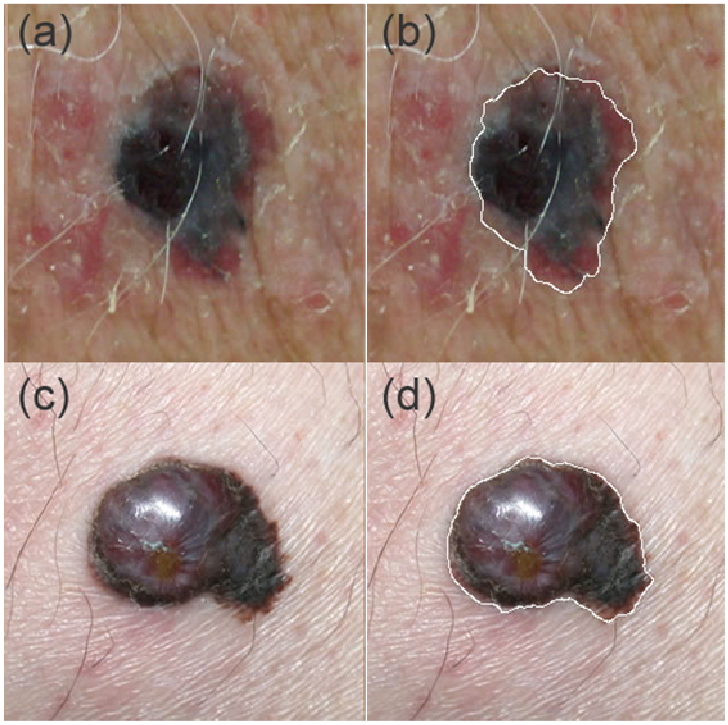


Fig 2. Two results obtained by the proposed method: original images (a) and (c); original images with the segmented contours overlapped (b) and (c), respectively.

In the images that had very smooth transitions between ill and healthy regions, the segmented contours were of inferior quality, as can be seen in Fig. 3. In this figure, we can verify that the upper region of the lesion (indicated by a yellow circle) was not correctly involved by the segmented contour. However, it is important to highlight that even in the cases with segmented contours not well adjusted to the lesion borders, the proposed method was able to successfully detect all existing skin lesions.

Fig 3. Example of segmented contour that is not correctly adjusted to the lesion border.

5 Conclusions and Future Work

A color image segmentation method to detect and extract skin lesion borders from images was presented to assist the medical diagnosis.

The method developed uses the region growing and active contour methods to extract and refine the lesion contours while preserving their main features, such as roughness and irregularity. Hence, the region growing method is used to pre-segment the input images by defining the initial segmentation contours, which are then deformed into the final segmentation contour by the active contour method.

The evaluation of the experimental results by a dermatologist allows to conclude that the proposed method is promising, being able to successfully detect the skin lesion regions in the input image. However, it still has some limitations, such as when it is used to detect lesions that have very smooth transitions between the ill and healthy areas. To overcome this problem, the used active contour method will be replaced by another method able to successfully deform the initial contour given by the region growing method even those cases.

Acknowledgments.
The authors are thankful to CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) and FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) in Brazil by the financial support.

This work was partially done in the scope of projects “Methodologies to Analyze Organs from Complex Medical Images – Applications to Female Pelvic Cavity”, “Aberrant Crypt Foci and Human Colorectal Polyps: mathematical modelling and endoscopic image processing” and “Cardiovascular Imaging Modeling and Simulation - SIMCARD”, with references PTDC/EEA-CRO/103320/2008, UTAustin/MAT/0009/2008 and UTAustin/CA/0047/2008, respectively, financially supported by FCT - Fundação para a Ciência e a Tecnologia in Portugal.

The first author would like to thank his PhD grant from FCT with reference SFRH/BD/61983/2009.

References

1. Ma, Z., Tavares, J. M. R., Jorge, R. N., Mascarenas, T. , *A review of algorithms for medical image segmentation and their applications to the female pelvic cavity.* Computer Methods in Biomechanics and Biomedical Engineering, 2010. **13**(2): p. 235-246.
2. Kass, M., A. Witkin, D. Terzopoulos, *Snakes: Active contour models.* International Journal of Computer Vision, 1988. **1**(4): p. 321-331.
3. Gonzalez, R.C., R.E. Woods, and S.L. Eddins, *Digital Image Processing Using MATLAB.* 2003: Prentice-Hall, Inc.
4. Lee, W.-L., Chen, Y.-C., Chen, Y.-C., Hsieh, K.-S., *Unsupervised segmentation of ultrasonic liver images by multiresolution fractal feature vector.* Information Sciences: an International Journal, 2005. **175**(3): p. 177-199.
5. Carvalho, B.M., G.T. Herman, T.Y. Kong, *Simultaneous fuzzy segmentation of multiple objects.* Discrete Appl. Math., 2005. **151**(1-3): p. 55-77.

6. Liu, H., Xie, C., Chen, Z., Lei, Y.., *Segmentation of Ultrasound Image Based on Morphological Operation and Fuzzy Clustering*, in *Proceedings of the Third IEEE International Workshop on Electronic Design, Test and Applications*. 2006, IEEE Computer Society. p. 397-400.
7. Lung, H.V. and J.-M., Kim, *A generalized spatial fuzzy C-means algorithm for medical image segmentation*, in *Proceedings of the 18th international conference on Fuzzy Systems*. 2009, IEEE Press: Jeju Island, Korea. p. 409-414.
8. Hashemi, S., Kiani, S., Noroozi, N., Moghaddam, M. E., *An image contrast enhancement method based on genetic algorithm.* Pattern Recogn. Lett., 2010. **31**(13): p. 1816-1824.
9. Lai, C.-C. and C.-Y. Chang, *A hierarchical evolutionary algorithm for automatic medical image segmentation.* Expert Syst. Appl., 2009. **36**(1): p. 248-259.
10. Mukhopadhyay, A., U. Maulik, *A multiobjective approach to MR brain image segmentation.* Appl. Soft Comput., 2011. **11**(1): p. 872-880.
11. Stanescu, L., D.D. Burdescu, and C. Stoica, *Color image segmentation applied to medical domain*, in *Proceedings of the 8th international conference on Intelligent data engineering and automated learning*. 2007, Springer-Verlag: Birmingham, UK. p. 457-466.
12. Iscan, Z., Yüksel, A., Dokur, Z., Korürek, M.. Ölmez, T., *Medical image segmentation with transform and moment based features and incremental supervised neural network.* Digit. Signal Process., 2009. **19**(5): p. 890-901.
13. Yoon, S.W., Lee, C., Kim, J. K., Lee, M., *Wavelet-based Multi-resolution Deformation for Medical Endoscopic Image Segmentation.* Journal of Medical Systems, 2008. **32**(3): p. 207-214.
14. Lu, R. and Y. Shen, *Automatic Ultrasound Image Segmentation by Active Contour Model Based on Texture*, in *Proceedings of the First International Conference on Innovative Computing, Information and Control - Volume 2*. 2006, IEEE Computer Society. p. 689-692.
15. Hodge, A.C., Fenster, A., Downey, D. B., Ladak, H. M., *Prostate boundary segmentation from ultrasound images using 2D active shape models: Optimisation and extension to 3D.* Comput. Methods Prog. Biomed., 2006. **84**(2-3): p. 99-113.
16. Dokur, Z., T. Ölmez, *Segmentation of ultrasound images by using a hybrid neural network.* Pattern Recogn. Lett., 2002. **23**(14): p. 1825-1836.
17. Kolár, R., J. Kozumplík, *Fuzzy Approach in Ultrasound Image Segmentation*, in *Proceedings of the International Conference, 7th Fuzzy Days on Computational Intelligence, Theory and Applications*. 2001, Springer-Verlag. p. 924-929.
18. Barcelos, C.A.Z., M. Boaventura, E.C. Silva, *A well-balanced flow equation for noise removal and edge detection.* IEEE Transactions on Image Processing, 2003. **12**(1): p. 751-763.
19. Xu, C., J.L. Prince, *Snakes, shapes, and gradient vector flow.* IEEE Transactions on Image Processing, 1998. **7**(3): p. 359-369.
20. Dermatlas. *Dermatology image atlas*. Available from: http://dermatlas.med.jhmi.edu/derm (accessed in 2010).

A Decomposition Approach for the Complete Coverage Path Planning Problem*

Pedro Rocha, A.Miguel Gomes,

INESCPorto, Faculdade de Engenharia, Universidade do Porto,
Rua Dr. Roberto Frias s/n, 4200 – 465 Porto, Portugal
{pro10015, agomes}@fe.up.pt,

Abstract. In this paper, a complete coverage path planning problem is discussed, for application in real situations in the areas of agriculture, autonomous robotic house cleaning, and hydrographic survey. Since the approach to this problem is currently being developed, only the work developed so far is presented. In this problem one aims to find the most efficient path or circuit that completely covers a closed region. For the time being, no algorithms with universal real application capable of solving every variant of this problem exist. The proposed decomposition approach is based on dividing the initial region into convex sub-regions (which are easier to manipulate), followed by computing the full sequence of sub-regions to be traversed, and, finally, finding the full path inside each sub-region. This approach does not return an optimal solution to the complete coverage path problems, but its implementation define the basic steps for creating a software library that can be used to solve many geometrical based problems, in which these are included.

Keywords: Complete Coverage Path Planning, Autonomous Vehicle Routing, Algorithm Design, Computational Geometry.

1. Introduction

The Complete Coverage Path Planning problem (CCPPP) is a difficult problem to solve, where the main objective is to find the shortest path or circuit that covers an entire region. The degree of optimality of the solution has a direct effect on the energy consumption needed to completely cover a certain region, and also in the required amount of available time to do so. The resolution of these kind of problems have a broad range of real application situations, and can be widely used in tillage, planting, cultivating, spraying, among others. In many cases, specially in the agriculture applications [1], [2], the regions have forms bounded by natural obstacles like large stones, rivers, forests, and others. This reason causes the geometrical representation of the natural form to be an approximate of a complex geometric form, which can be considered irregular. Some other fields with some similarities with agriculture are autonomous robotic house cleaning [3], [4], [5], [6], in which a robot has to cover all the region of a surface with the goal of maximizing the cover but with the secondary objective of minimizing the overlapping of its path or circuit; and in the field of

* Partially supported by Fundação para a Ciência e Tecnologia (FCT) Project PTDC/EME-GIN/105163/2008 - EaGLeNest

hydrographic surveying [12], in which the vehicle makes the mapping of the underwater terrain and searching irregularities in the ocean. The particularity of the hydrographic survey problem is that, although it is limited by the same set of constraints, the area that is covered along its path is variable, since the visible area in the ocean floor depends on the field of view and distance to the same ocean floor, being greater with greater depths, smaller with lower depth, and the area monitored at one side of the path may be different from the area monitored on the other side when moving horizontally regarding a slope.
The proposed decomposition approach has also the objective of testing the capabilities, in performance as also in adaptability, of a geometric library focused in solving nesting problems, like two-dimensional irregular cutting-stock problems[13]. This library is being reconstructed to expand its functionality, performance and ease of use. Its development is progressed in a modular structure where each module has a specific application. With this kind of experiment, we hope to verify the current capability of the geometric library, to increase the diversification of its application areas and to further expand its functionality.
This paper is organized in 5 main sections. Starting with the Introduction, in Section 1. After it, in Section 2, it is presented an overview of the problem to solve, a general description of how it is usually solved, and also what tools are used to solve it. In Section 3, a detailed description of our proposed solution approach to the problem is given, with the methods and algorithms used, the requirements and choices made based on the constraints of the problem. Section 4 has the overview of the geometric library that is used as a tool to implement these algorithms, presenting some main functions used to build these algorithms, and its current state of development. In Section 5 the final comments and future work are presented. No results are given since this is a preliminary work, currently under development.

2. The Complete Coverage Path Planning Problem

The CCPPP includes many distinct cases of real application, in several distinct fields, one of them being the field of agriculture. It is usually represented by irregular geometrical forms, with a varied degree of complexity, which can include holes and curved segments. This complexity in its geometrical form causes some limitations in the possible ways that can be used to solve the CCPPP, so some approaches divide the initial geometrical form into more simpler forms to ease the problem resolution [1], [7]. Ideally, one would get the full path or circuit that would cover the entire region, without any kind of overlapping, with the minimum number of turns, if they represent and additional cost, in time or distance. Some methods divide the initial form in triangular, trapezoidal, rectangular, and other polygonal forms, to reduce a single complex problem into several simpler problems. After the division, the objective is to try to find the best path or circuit that covers all of the polygons, regarding the restrictions that increase or decrease the total cost.
Another methods proposed to deal with the CCPPP are based in rectangular cell decomposition [7], triangular cell decomposition [8], neural networks [9], genetic algorithms [10], but some of these methods have specific limitations. They need to have previous information about the area that they will have to go through, with the

size and location of the obstacles. That information may not be initially available. One example of an application with an initially unknown environment is [7], which makes the mapping of the terrain in real time, and adjusts its defined path when needed.
In the chosen solution, the approximate representation of the real form, for which is determined the maximum possible cover path or circuit with the minimum cost, is described by a set of regions which can have holes and disjoint regions. All the representations will initially be made with linear segments, although curved segments, circles, ellipses and splines are supported in the library. One example of an region divided into individual regions, and with the global path already determined is presented in the Fig.1.

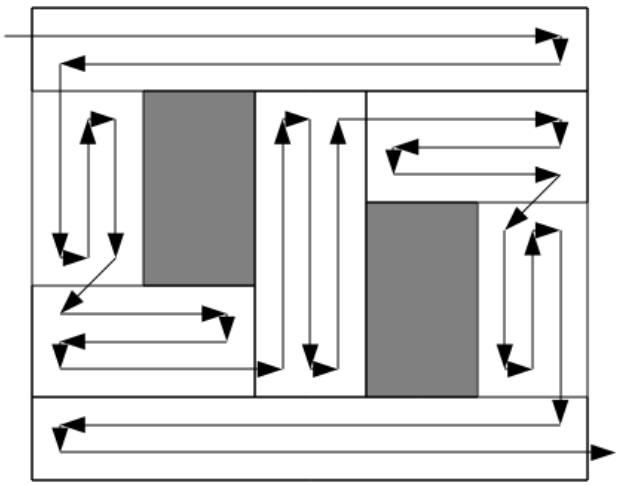


Fig.1. Global path example.

3. Solution Approach

In this work, the chosen method of resolution uses a decomposition approach to tackle the CCPPP, which uses some functions developed in the library. This decomposition approach has the advantage of being easier to implement, and to return results faster than other approaches. Other necessary functions that do not yet exist will be developed and inserted into an independent module on the EaGLeNest library when necessary.
Two solutions to solve two variants of the CCPPP are presented. The first one has the main objective of minimizing the number of turns in the complete circuit, while the second one, the main objective is to minimize the maximum length of the circuit, also achievable by minimizing the overlapping of the circuit path with itself.
Both solutions are divided in the same three phases. These three phases consist in decomposing the complex form into convex polygons, then finding the Hamiltonian circuit in the graph generated by connecting the centers of the convex polygons, and finally the computing of the complete cover path for all convex polygons. The two

solutions only differ in the third phase.

3.1. Convex Decomposition

In the first phase, the first step is to make the transformation of the real area into an approximate representation by connected linear segments, with the obstacles being defined as holes, but with a sequence of linear segments connected in inverse order compared to the outer layer. When a polygon is separated from the others (without any edge that connects it to the other polygons) it is treated independently. If some connection between invalid areas is allowed, we need to specify the extra cost to traverse it from one valid area to another. As such, as a first step, the holes are removed, by dividing the polygon in several parts, sectioned by an horizontal line positioned at the half of each hole.

In the end, we will get several parts of the initial form, without holes, with a maximum number equal to the number of holes plus one unit. Each of these elements will be further divided into triangular polygons. Each triangle is made from three sequential vertexes in a polygon, if they create a convex polygon in the inside of the same polygon. The outer vertex is ignored on the next step, and the process repeats itself until only 3 vertexes remain. The triangularization algorithm does not guarantee that the minimum number of triangles will be generated for any single piece. We can see that in Fig.2.

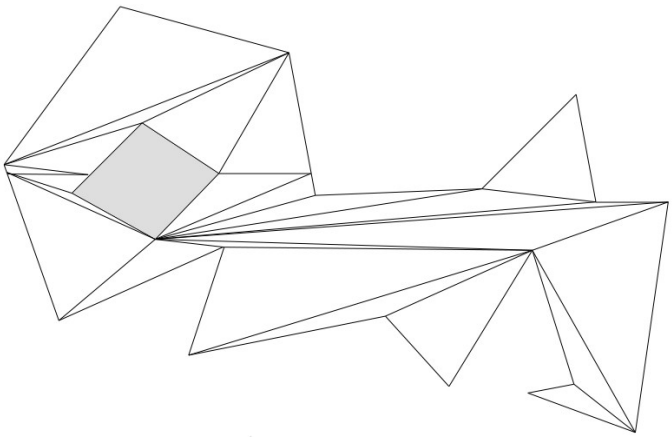


Fig.2. Complete triangularization of a complete form.

After the triangularization, all triangles are represented in a single mesh. This mesh allows that the transformation algorithm of triangular polygons into convex polygons tries to maximize the size of the convex polygons that may be build from a vertex that belongs to the outer layer, minimizing the total number of convex polygons generated. This transformation algorithm works by choosing an outer layer segment that has not yet been selected, and proceeding to the next segment that does not breaks the convexity criteria. In the event of not finding a next segment that complies with that restriction, it returns to the previously selected segment, marks the current segment as

invalid, and continues the search from that point. In the worst case, the convex polygon generated is the same triangle from where the algorithm started. The effect of this can be seen in Fig.3.

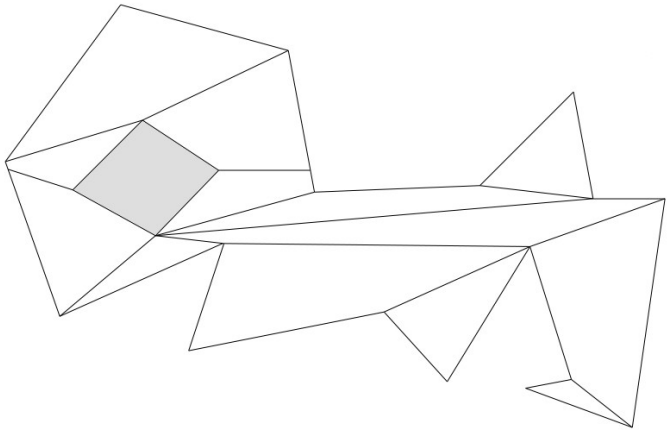


Fig.3. Convex polygons generated by the transformation algorithm.

3.2. Hamiltonian Circuit

The second phase starts by calculating the positions of the center of the convex polygons generated, and making the connections between them accordingly to shared segments between polygons. This step creates a graph with the connections between the center of each polygon connected to the other centers of polygons that share a segment. If they share just a vertex, that is not enough to make a connection.

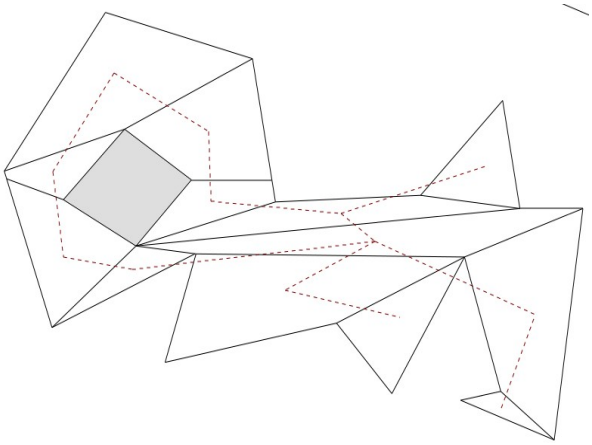


Fig.4. Equivalent graph generated from the center of the convex polygons and their contiguous polygons.

One needs to generate an unique complete circuit that reaches all the vertexes of the graph. This type of problem can be described as an equivalent to the Traveling

Salesman Problem, assuming that all the vertexes have a valid path to any other vertex. Since we cannot guarantee that a viable path exists between a given pair of vertexes, we compute the connections between any directly unconnected pair, with the minor cost between them, that goes through other intermediate vertexes. These virtual connections are added to the main graph, making the full Hamiltonian circuit[11] easy to compute, but however, it does not guarantee that the triangular inequality condition is satisfied.

To compute the full Hamiltonian circuit, we use a simple Traveling Salesman Problem heuristic, through a greedy algorithm that starts in any vertex, and proceeds to the next closest vertex, until it completes the path. At the end, it just connects to the starting vertex, and the full circuit is complete. Since all vertex have a connection to any other vertex, there is always a free connections to another vertex until all vertexes are included in the path.

The determination of the Hamiltonian circuit has the advantage of allowing to know which will be the edges of each convex polygons that will act as entry and exit points. Through these edges we can optimize the way the internal cover is made in each convex polygon, and also determine where the entry and exit point will be located in each entry and exit edge.

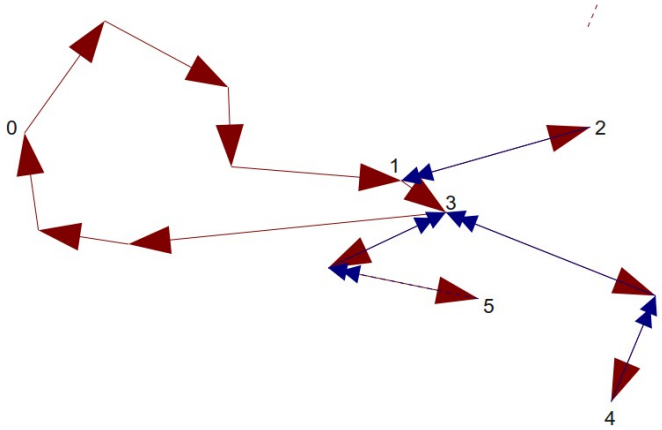


Fig.5. Hamiltonian Virtual Circuit in the graph.

The Fig.5. presents the virtual full Hamiltonian circuit that was computed in the graph. The connections represented by a single end arrow are normal connections. The double head arrows represent part of a virtual connection that is overlapping with the previously traversed path. The direction of navigation in the circuit of the graph is determined by the arrows. Since this circuit travels through the same vertexes more than one time, we cannot say that this is an Hamiltonian circuit, but it can be considered one if we convert the returning paths into alternative connections that connect only the end vertex of a real path, and a virtual free vertex that it returns to. The real circuit in the presented graphs starts in 0, proceeds to 1, goes to 2, returns to 1, advances to 3 and then 4, returns to 3, goes to 5, turns back to 3 and then finally ends with 0. However, the virtual Hamiltonian circuit does not return to previously traversed vertexes. The Hamiltonian circuit is 0, 1, 2, 3, 4, 5, 0.

3.3. Polygon Coverage

In the third phase, with the complete circuit already defined, and the entry and exit edges already selected, the complete circuit that covers the whole area can be built, with the goal of minimizing its cost, in two distinct ways, depending on the imposed conditions. If each turn implies an extended cost, a solution with the goal of minimizing the total number of turns might be preferred, even considering the increase of overlapping in the final circuit. In the opposite approach, if the turns do not add a significant cost, or if the cost of overlapping is minimal compared to it, the solution that minimizes the total overlapping is the favored resolution approach.

If we consider the option of minimizing the total number of turns, the proposed heuristic is described in the following paragraphs.

Ignoring the positioning of the entry and exit segments in every convex polygon, the coverage path of each polygon is achieved in a perpendicular pattern to the segment that connects the most distant pair of vertexes for every convex polygon. This minimizes the number of turns in the convex polygon coverage path [1].

Unfortunately, with this type of coverage path, the entry and exit points for each convex polygons are going to be modified. The added cost is the cost of the path overlapping necessary to connect one exit point from one convex polygon to the entry point of the next one. Even so, this type of solution might have an advantage in some instances of this kind of problem. The time that the path takes to be traversed is not taken into account, and so it is not considered as an added cost, but usually traveling a certain distance in a straight line is faster than traveling a same distance in a path with many turns.

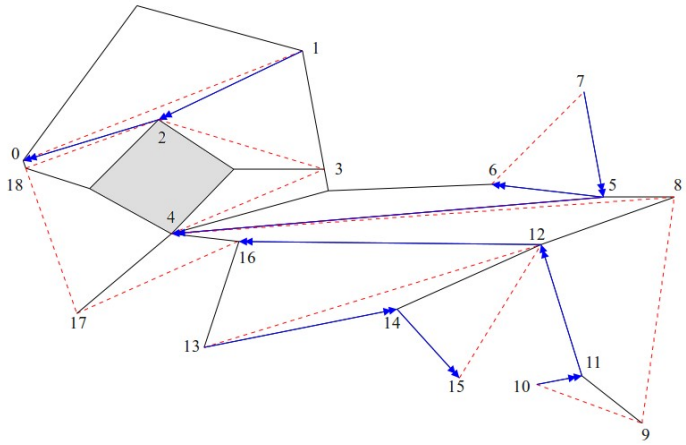


Fig.6. Complete path generated by the turns minimization heuristic.

In the previous image, the dashed lines indicate that the convex polygon which they are contained will be covered in a perpendicular pattern to that same line. The arrows with double end show the additional overlapping cost of connecting the entry and exit points between convex polygons. The full circuit, based on the solution from the previous step, follows the sequence defined in the image:

0, 1, 2, 3, 4, 5, 6, 7, 5, 8, 9, 10, 11, 12, 13, 14, 15, 12, 16, 17, 18, 2, 0.

When we consider the option of minimizing the overlapping of the circuit, we use an heuristic like the one proposed in the following paragraphs.

Taking into consideration the entry and exit edges of each convex polygon, a few particular cases might need some adjustments, so that the algorithm can generate a path without any overlapping.
When a convex polygon has several entry and exit points, to make a correct coverage the polygon needs to be divided accordingly to the number of the matched pairs of entry and exit points. This type of division divides a convex polygon with several entry and exit points, into several convex polygons with only a pair of entry and exit points, just like the example in Fig.7.

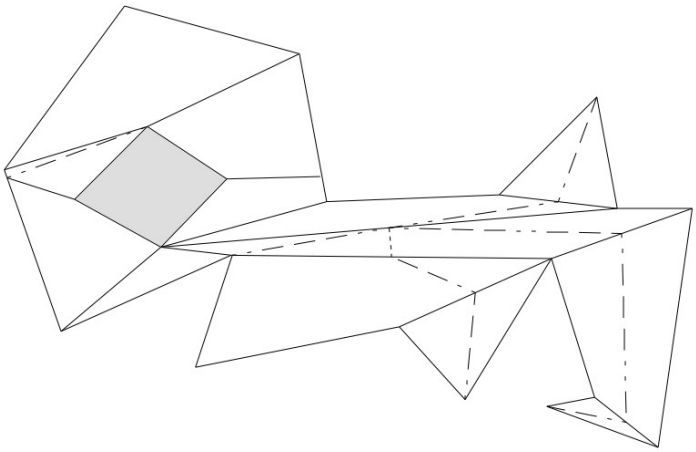


Fig.7. Convex polygon division into polygons with only one pair of entry and exit vertexes.

Still, another transformation might need to be done. We cannot always cover a convex polygon without any overlapping if we follow an exclusively perpendicular pattern to the line that connects the entry and exit vertexes, even when the vertexes are in contiguous edges.

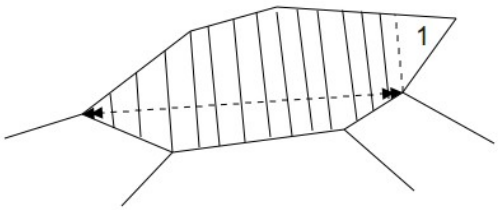


Fig.8. Invalid path coverage in a convex polygon.

As it is shown in Fig.8, the area marked by 1 cannot be covered if we follow the perpendicular pattern to the dashed line that connect the two most distant vertexes of the entry and exit segments.
As such, to solve this problem, the convex polygon is further divided, with a first division made from the common vertex, in the case of contiguous segments, or from

the last of the in-between outer layer segments that connect the vertex in the exiting edge up to the most distant vertex from it, that preferentially does not belong to the entry edge. One exception to this case, in which no division is made, is when the convex polygon is a triangle. In this exception, we cover the polygon with a parallel pattern relative to the entry edge. If the edges are not contiguous, we need to make an alternative kind of division, to make sure that the cover ends at the exiting edge. When an edge is simultaneously an entry and exit edge, that edge will be divided, connecting to an opposite free vertex or edge, also dividing the polygon in which it is contained. We can see an example in Fig.9.

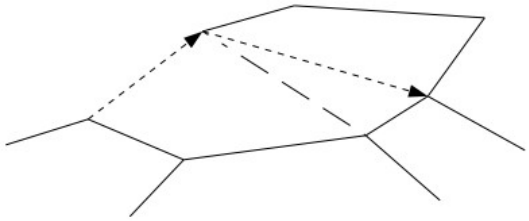


Fig.9. Main coverage lines after division.

This method allows to have a single path, without any overlaps, that minimizes the total distance traveled when comparing to the alternate solution. As a consequence, we now have a high number of turns, that can severely worsen the cost if they are taken into account, due to the further division of convex polygons, increasing their number, and creating the need to the individual coverage of every independent component.

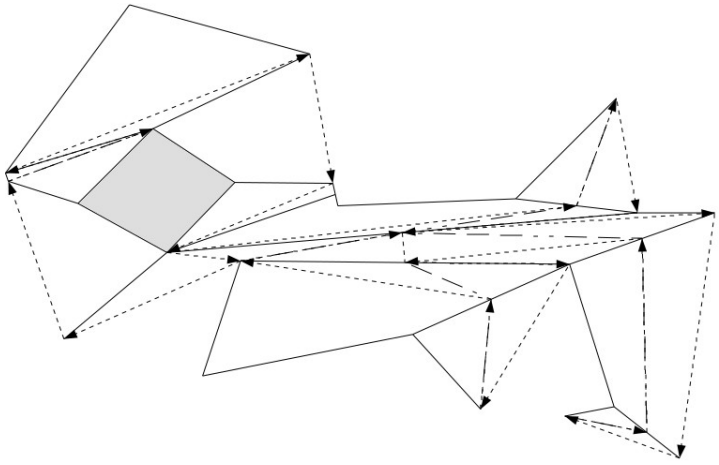


Fig.10. Coverage path created with the overlapping minimization heuristic.

The arrows in Fig.10 present the complete path without any overlapping, generated by

the described heuristic. The coverage pattern is perpendicular to the correspondent line contained in each convex polygon.

4. Implementation

The implementation of these algorithms will rely in some existing functions from the EaGLeNest geometric library. This reassures the capabilities of this library to diversify the range of practical application, having the simultaneous advantage of testing and develop this tool to be applied in this kind of problems.
Some functions might need to be slightly modified, depending on the problem, but no heavy modification of functions will occur, since it is preferred to develop an independent function from scratch with a particular use in mind.
The data to any problem, be it vertexes, edges, polygons, and groups of polygons, and also algorithms to use in the resolution of the problem with their execution sequence, are contained in a XML file, which is the default file to use with the library. This format ensures backwards compatibility with the current version of the library, while also allowing the ease of reading, construction and manual modification of the contents of the file by any person. The main data structures are based on simpler one's, with identifiers. The most elementary unit is the vertex, or dot (with identification). With two vertexes we build a linear segment and its identification, and with an array of segments we build a polygon, also with identification. The higher hierarchical structure available is designed by geometrical shape, which can contain several polygons that might not be contiguous with any other, not be defined as holes, nor contained in another polygon's hole. These structures are also designed to support curves, with Bezier representation, including also ellipses and circles (and arc if an angle is specified), but the functions used to manipulate these forms are not yet implemented in the geometric library.
The most basic functions of the geometric library that were implemented to this moment are:
Reading data in XML file format, and loading the data structures accordingly; intersection detection between segments, that return the intersection points (be it linear segments or curved segments); computation of the minor angle between the intersection point generated with the intersection of two segments; the lesser distance of a vertex to a straight line; position of a vertex relative to a straight line segment (to the left, right or co-linear); the linear segment most to the left or to the right of a current selected segment; the nearest linear segment to the left or to the right of a current selected segment; determine if two vectors are oriented to the same quadrant; determination of the representation of a polygon (if it is a hole or an outer layer, according to the settings specified in the system); inversion of a polygon (transform a hole into an outer layer and vice-versa).
The functions composed by these basic functions of the geometrical library are:
Transformation of a polygon with holes into several polygons without holes; triangulation of irregular geometrical forms, construction of lists that contain every reachable vertex from a specified vertex (useful to determine which vertexes belong to a polygon); decomposition of a mesh into individual polygons; function that returns only the outer layer of a mesh and another that subtracts only the outer layer from a

mesh; computation of biggest convex polygons in a mesh (with exclusion for partition, and inclusion for coverage of polygons); construction of an adjacency matrix from a mesh; merging of geometrical forms, and finally the No-Fit Polygon construction from a pair of convex polygons.
The functions used in the geometric library, that can be used in the resolution of the problem discussed in this paper, are mostly the elementary functions, including a few of the more complex functions like the transformations of convex polygons in polygons without holes, the triangularization, and merging of polygons. The functions destined to the computation of the coverage path of the convex polygons are currently being implemented. The functions destined to the computation of Hamiltonian paths, and to the resolution of variants of Traveling Salesman Problems will be implemented later.

5. Final Comments

Since this is still a preliminary work, we cannot solve for every best solution possible with these heuristics, but we can use them to attempt to get solutions that achieve the minimum amount of turns or the minimum overlapping.
The geometric library has only some basic functionality, but already shows some good flexibility and support for generic applications, when considering usage of some of the implemented functions described in the previous section (4). It cannot be used extensively in any area, but can be used to start solving these problems in the fields of agriculture, hydrographic surveys, autonomous robotic house cleaning, and a few others.

As future work we expect to fully implement the modules presented in this work. We then proceed to collect the results from some variants of this proposed problem, and make some adjustments to improve performance. We also plan to improve the algorithms used in this paper, and continuously improve the geometric library to support more features. Most of the features have the possibility of being used in other kinds of problems that can have a representation based on geometrical forms.

References

1. G. Zuo, P. Zhang, and J. Qiao, “Path planning algorithm based on sub-region for agricultural robot,” in Proceedings of the 2nd international Asia conference on Informatics in control, automation and robotics - Volume 2, CAR’10, pp. 197–200, IEEE Press, 2010.
2. T. Oksanen and A. Visala, “Coverage path planning algorithms for agricultural field machines,” J. Field Robot., vol. 26, pp. 651–668, August 2009.
3. H. Choset, “Coverage for robotics – a survey of recent results,” Annals of Mathematics and Artificial Intelligence, vol. 31, pp. 113–126, 2001.
4. De Carvalho, R.N.; Vidal, H.A.; Vieira, P.; Ribeiro, M.I.; , "Complete coverage path planning and guidance for cleaning robots ," Industrial Electronics, 1997. ISIE '97., Proceedings of the IEEE International Symposium on , vol.2, no., pp.677-682 vol.2, 7-11 Jul 1997.
5. X. Wang and V. L. Syrmos, “Coverage path planning for multiple robotic agent-based inspection of an unknown 2d environment,” Mediterranean Conference on Control

and Automation, pp. 1295– 1300, 2009.
6. R. Mannadiar and I. Rekleitis, “Optimal coverage of a known arbitrary environment,” in Robotics and Automation (ICRA), 2010 IEEE International Conference on, pp. 5525 –5530, May 2010.
7. H. Choset and P. Pignon, “Coverage path planning: The boustrophedon cellular decomposition,” in In International Conference on Field and Service Robotics, 1997.
8. J. S. Oh, Y. H. Choi, J. B. Park, and Y. Zheng, “Complete coverage navigation of cleaning robots using triangular-cell-based map,” Industrial Electronics, IEEE Transactions on, vol. 51, pp. 718 – 726, June 2004.
9. S. Yang and C. Luo, “A neural network approach to complete coverage path planning,” Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, vol. 34, pp. 718 – 724, February 2004.
10. A. Ryerson and Q. Zhang, “Vehicle path planning for complete field coverage using genetic algorithms,” Agricultural Engineering International: the CIGR Ejournal, vol. IX, July 2007.
11. Reinhard Diestel – “Graph Theory” - Springer - Verlag New York 1997,2000 – Page 214, theorem 10.1.1.
12. Cheng Fang; Anstee, S.; , "Coverage path planning for harbour seabed surveys using an autonomous underwater vehicle," *OCEANS 2010 IEEE - Sydney* , vol., no., pp.1-8, 24-27 May 2010.
13. Julia A. Bennell, Jose F. Oliveira, “The geometry of nesting problems: A tutorial”, European Journal of Operational Research, Volume 184, Issue 2, 16 January 2008, Pages 397-415.

A Mobile Location-Based Game Framework

João Tiago Pinheiro Neto Jacob[1],

[1] Faculdade de Engenharia da Universidade do Porto
joao.jacob@fe.up.pt

Abstract. As mobile platforms evolve, so does the concept of video games for these platforms. Location-based games are a recent type of games that explore the unique capabilities of the GPS-enabled mobile devices. However, the development of these games isn't easy as they often encounter several issues regarding the usage/availability of location-related content. This paper presents an in-depth analysis of these games and their issues and presents a solution to these problems through the creation of a framework. This framework also offers many features useful for location-based games, such as remote-map access, weather location and easy access to the device's GPS module. A proof-of-concept location-based game based upon said framework is also presented, successfully validating the model.

Keywords: Location-based games, mobile computing, GPS, location-based services, mobile games

1 Introduction

Location-based games are relatively new in the entertainment industry, as the first location-based games only came to be in 2002 [13]. These games are known for their unpredictability as they rely upon the user's real location (or other location data), as a means of input, or as a means of generating/accessing game specific content. As such, these games provide players with distinct gaming experiences, not only from player to player but also from location to location, effectively increasing the game's longevity and its possibilities. As these games use location services to function (usually a GPS - Global Positioning System - module for user positioning), they can almost exclusively be found on mobile platforms, as these are more prone to be the most adequate.

However, these games often become unplayable in situations where the location services are inoperative, such as inside buildings when playing a game that requires the usage of the GPS module, as the GPS needs line-of-sight with the sky in order to pinpoint the user's position. Still, some of these game's issues aren't so closely related with the hardware limitation, be it performance or communication wise. In fact, most of the issues found on these games are related to their location-based characteristics. So, while a game may be played adequately in the location A, it may be too difficult (or unplayable altogether) in the location B. Additionally, some locations may be more easily and naturally incorporated in the game's mechanics,

while some won't, usually due to the lack or poor quality of information that location has associated or generated within the game.

During the analysis of some location-based games and location-based game's platforms, some generic issues were identified. With the goal of solving most of these issues, a location-based game framework was designed and implemented. Finally, a location-based game was developed atop said framework [18] in order to serve as a proof-of-concept and determine if the designed framework was able to cope with the issues. The testing of the game (Geo-Wars) is being made by several anonymous users, but the feedback gathered for now helped to fine tune the framework, and only some minor issues are amiss.

This paper accommodates a small section describing some location-based games and their technology as well as a "location-based games' issues" section with an in-depth analysis of the found issues. A section detailing both the framework and its implementation and another section regarding the concept of the location-based game Geo Wars is also present. Finally, the discussion and the conclusions and future work sections, portraying a critical and objective analysis of the developed work and its result, follows.

2 Location Based Games

A Location-based Game is a game that uses the player's physical location (or any other location) as a means of input or to generate or access location-based information. These games are almost exclusively available on mobile platforms [16]

Location-based games haven't been around for too long. In fact, they would only see any commercial use in 2002 with the arrival of Botfighters [2], the first pay-per-locate GPS game. The concept was simple: each player was a robot with the mission of destroying the enemy robots. In order to play the game, the player would move around the city, scanning for enemy robots, which would be other players with the same goal.

With the birth of this genre only roughly eight years ago, location-based games have now gained a considerable popularity. For an instance, Geocaching [6], probably the most popular of geocaching games, claims to have more than a million registered users. The idea behind it is even simpler: a player stashes a "cache" (with contents in accordance with the geocaching's rules) anywhere in the world and shares its approximate location with other geocachers around the globe. Other players will attempt to discover this cache by solving some riddles and exploring the general area the cache is known to be.

Of course other games, although using the physical location of the player, need to be played with a mobile location-capable device. Pac-Manhattan [12] is such a game. In it, players will have to enact a classic Pac-Man game. In order to do so, ten players are required: one to be Pac-Man, four other to be ghosts, and the remaining five to control and coordinate the actions of the others via mobile phone. As the Pac-Man player wanders around the streets of a real city, the player responsible for controlling him will guide him through the streets, avoiding the ghosts and, of course, eating pellets.

As location-based games gained popularity, several platforms of mobile games emerged, with two having a particular success. One of them is Groundspeak [5], the platform that contains not only the Geocaching game, among others, but also the Whereigo tool, a tool for creating and playing GPS-enabled games [11]. The other, Locomatrix [4], currently in quick expansion, offers two games: Treasure hunt, a geocaching like location based game, with a more virtual component, as the player has to use pictures to solve riddles and find places, and fruit farmer, a game where the player must collect "fruits" around him by moving around in the real world, avoiding also having these fruits stolen by other farmers. As it can be seen, location-based games have come to be quite a success from their origins in the early 2002.

2.1 Location-Based Games' Issues

The issues that were found in most location-based games usually fall into one of these following categories:

- Game design issues
- Hardware limitations
- Location-related information availability and suitability
- Player's fitness and pace

There are several possible solutions for these issues.

2.1.1 Game design

Regarding game design issues it is important to keep in mind that location-based games involve player exposure to physical interaction with the real world. Meaning that it is important to consider where and how the player will attempt to play the game and try to keep the player safe. Being a location-based game it always involves some degree of unpredictability. However, that unpredictability can be reduced if the player's behavior is limited thanks to the gameplay. For an instance, setting a racing location-based game that involves the player to go from A to B may be tempting for some player to attempt to break the record using a motorized vehicle and thus risking his safety and that of those that surround him. However, if a restraint is set, by determining another objective like, considering the previous example, stating that the player's top speed must not exceed 30Km/h, the risk of a player speeding his way through the game is reduced. Additionally, and in order to ensure that the game is played on foot, using the devices' sensors such as the accelerometer to work as a pedometer may help. These solutions don't alter the game altogether but help ensure that the gaming experience remains constant and that the game's unpredictability is lowered. Since game design is such a vast area, it is difficult to analyze location-based game design issues without going through case by case analysis. Even so, the important thing is to keep in mind what possible issues may arise from the conceptualized gameplay.

2.1.2 Hardware limitations

Hardware limitations are difficult to overcome. In regular mobile games, the developer might limit the user's actions or the game's requirements. However, in the

case of location-based games, since other hardware is often used, it is mandatory to have them taken into account.

Since these games often rely on GPS or data connections in order to be playable it is often the case where the player, unable to have one or neither of these services available, is incapable of playing the game. However, if the game's mechanics allowed playing the game without using GPS, but still using some location as an input (such as from logs or via direct input by the user) this situation could be avoided, sacrificing gameplay experience. On the other hand, the game could be played indoors or without any preparation whatsoever, which would be sure to please the most casua gamers. Regarding data connections, more specifically via mobile carier, since these are usually paid-for connection and have a limited coverage, the most adventurous of the players will probably find himself in a situation where the access to this service is precarious or too expensive. Under these circumstances, allowing the player to still play the game, either using cached location-related data to be used or, in case that location-specific data is lacking, using other location data. Such approach would allow the player to overcome this issue and still play a location-based game. In the case of the game using location data that isn't specific for the location of where the game is being played, the game can still use the GPS signal to move the player's avatar in the game (if applicable). Obviously, although mapping his movement in the game, the game itself won't be considering the players true surroundings. If the game requires some data to be transferred, such as scores, statistics or saved games, this data transfer can often be postponed with no harm to the game.

2.1.3 Location-related information availability and suitability

Since location-based games often use maps, weather information, or any other kind of location-related information in order to make the game truly unique and location-based, it is not uncommon to find location-based games that are rendered unplayable in many parts of the globe, due to requiring information that is either not available or that isn't relevant and cannot be used in the game. Creating such a game, like Pac-Manhattan [12], means that the game is unplayable outside of the area it was designed to be played, and thus, has a narrowed-down target audience. There are three possible solutions for this issue: not using location-related information, generating the needed information randomly, by using the player's GPS position as an input, so that, even though the location's true novelty is lost, a new virtual novelty is created for that location, guaranteeing that the game is playable worldwide. Alternatively, and since mobile devices storage capabilities are now into the tens of gigabytes, it is possible to store data for all/several of the needed locations. However, and while this may ensure that the novelty of the location is used (even if still not available for some people), that information is prone to become out-dated, particularly if that information is very ephemeral, such as that of weather or traffic. Ultimately, the usage of remotely stored location-related content that is often updated via web-services will guarantee that the needed information is up-to-date and available to everyone without sacrificing the device's limited storage. Unfortunately, this implies the usage of data connections, limited location coverage (as the information may not still be available for every playable location), and possibly remotely inaccessible servers. Often, even if the

location-related content is available, it may still not be of any use if the game's mechanics doesn't take it into account.

2.1.4 Player's fitness and pace

As it happens with most location-based games, the player's movement around the real world is used as an input, often to move his avatar around the virtual world. And like some games, such as Fruit Farmer [4], the player's speed and stamina is determinant for the outcome of the game. Unfortunately, this means that many of such location-based games are too difficult or altogether unplayable by the unfit player, providing an unbalanced gaming experience. If, however, the game was able to take the players pace into account, balancing its difficulty in real-time, the player would feel that the game was suited to him, and was still challenging without being too easy or overwhelmingly difficult. In the case of the "Zombies, Run" [14] game, (a game where the player must go from point A to point B in a real map by physically moving from the real point A to real point B while avoiding the hordes of Zombies that will pursue him) if it automatically adjusted the number of zombies and their speed to the pace of the player (taking into account the distance travelled, the average speed and the current speed), the game would provide a custom-tailored experience. Of course the game would still value the fastest of the players via metagaming (such as scores or boards), but the game would be playable for everyone. Something often overlooked by location based games is that the player will probably need to stop to catch his breath. If possible (with the exception of real-time multiplayer location-based games) the game should be paused automatically whenever the player's stopped, or , if not pausing, slowing the game down significantly, giving the player the chance to gather his strength, as opposed to him losing the game due to pausing for a few seconds.

3 Location-Based Game Framework

In order to solve the aforementioned found issues (using some of the possible solutions for them), a generic solution for a location-based game, a location-based game framework, was designed. This solution, aims to be as general as possible, so as to be usable by any location-based game that requires support for the previously numbered issues.

The implementation of both the game's concept and the core components of the framework followed a "waterfall" methodology variant called the Sashimi model [20], allowing to jump back and forth between the several phases of development, such as the game's implementation and the game's design as well as the distinct phases of the framework development, so as to use the feedback each of them provided to alter and complete other phases. After the creation of a first fully working prototype based on this framework, this product was tested on the field both by the developer, some other volunteers and also by the community at xda-developers.com. This evaluation allowed the adoption of the "spiral" model, a model that uses user feedback to reevaluate product features and the whole design of the solution, improving it with each iteration. Since both (Sashimi and spiral models) are agile

methodologies they were sure to provide flexibility in order to spot and solve problems on the fly, even if spotted during distinct phases of the development [1]. The design here presented for both the framework and the location based game prototype (Geo Wars) is the final one, considering already most of the provided feedback. The design of the solution shows the relevant key elements needed for it to function.

Each step of the methodology and the design of the components of the framework will be covered in further detail during the sub-sections that follow.

3.1 Location-Based Game Framework Architecture

The created framework is capable of accessing either local (on the mobile device) or remote (stored in a server) location-specific content. Meaning that the player can play a cached game, one that he has already played or has already downloaded the needed content, or the player can easily play a new, never before played game. This allows the player to:

- Play a game even without data connections, simply playing a saved game,
- Playing a new game, without worrying about having the necessary data to play it.

Furthermore the framework supports the option of letting the player decide to use his GPS to play the game, or not. Obviously, for some games, this may or may not make sense; e.g.: playing a Manhattan Pac-Man with no GPS at all means that the player will be playing a regular pac-man game. Still, it is usually better than no game, or an unplayable game, at all.

If the player decides to not use his GPS (or the service is unavailable), the location information can be either inserted into the system via prompting the user (e.g.: selecting the country or city he is playing at), or allowing the input to be done another way; e.g.: if the player's avatar is moved by mapping the player's real position, obtained via GPS, with the avatar's virtual position, then the avatar may also be played with on-screen controls, even though the game may lose its location-based characteristics. So, regarding the GPS a player can decide to:

- Turn the GPS off, while giving the needed data for the game to remain playable,
- Keep the GPS on, in order to take full advantage of the location-based characteristics of the game.

Additionally, the player may opt to create and share a customized location-based game's settings online. This allows for the games to gain a social and competitive aspect, as players may play against other players (through a game specific scoreboard), and can also share and check ideas for location-based game's settings. This is important, as some players will be unable to play a desired location-based game. For instance, a player may wish to play a game as if he was in Paris. Even though the player has GPS available, he isn't there. So, the option of playing a game with content that isn't available for the player's location while using the GPS enabled device to sense the player's movement around is possible.

The framework's architecture can be summarized as follows:

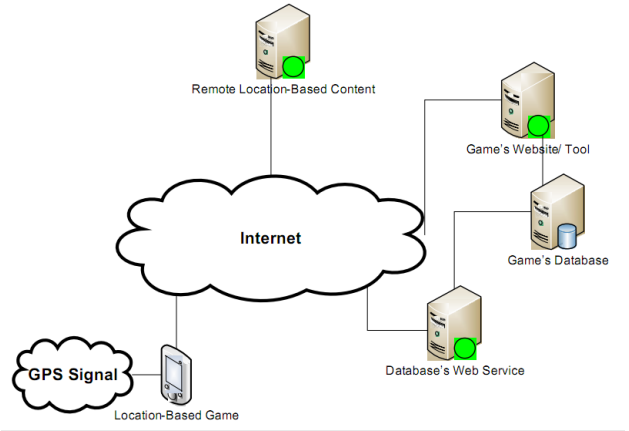


Fig 1: The solution's design

The Figure 1 summarizes some of these capabilities of the framework, in a physical architecture point of view of the solution. The framework supports the usage of a database's web service that allows for accessing games' scores and loading previously created games (using the game's website/tool and the game's database for storing this content). It is also explicit the access of remote location-based content via the internet (when needed), while the location-based game client itself runs on the device.

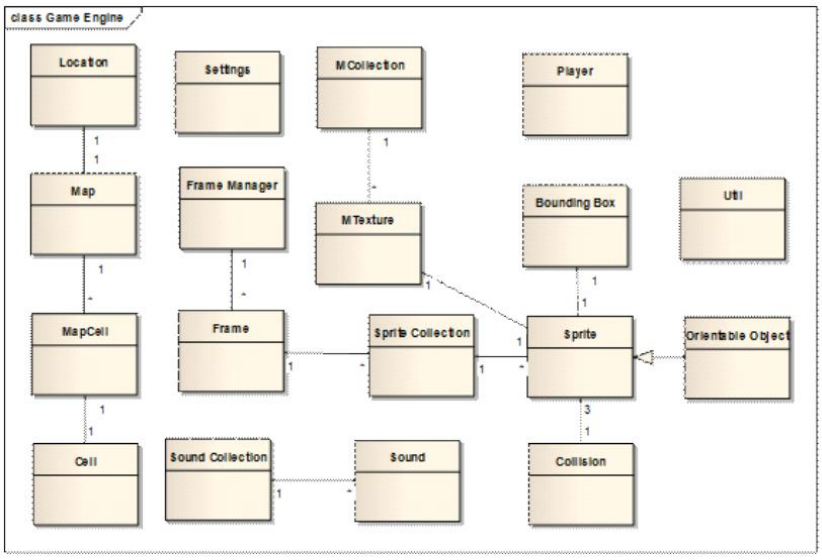


Fig. 2: Diagram depicting the core components of the framework

While some of the components presented in the Figure 2 provide features present in most game development specific frameworks, such as the Sprite class, Bounding

Box based collisions, Frames and Frame Managers, as well as sound playback, and some generic functions, present in the Util class, such as cache for sine/cosine operations, it also offers some distinct ones that will be here described. Other functionalities such as texture caching, sprite animation and transformation won't be explained, as they are common in most game development frameworks and aren't both the scope of this paper nor the most valuable assets of this project.

Cell: This structure holds information regarding the type of cell, the considered atomic structure type of a map. The type of a Cell is an enumeration and can be Building, Grass, Water or Road. It can be expanded to include other types of unitary cell blocks or points of interest as needed.

MapCell: A MapCell is the smallest, atomic element that constitutes a Map. It holds information of a particular part of a map, such as the type of that part (the Cell) and the coordinates of it.

Map: A Map is formed by several MapCells, detailing the maps content. Here is an image depicting the information present in a Google Maps Static API image versus the respective generated Map structure containing the location-based information found on the original map:

Fig. 3: Original Google map and visual representation of the in-game generated map

As it can be seen in Figure 3 a Map will hold information regarding streets among other elements (such as green areas, buildings, etc). Its resolution depends on the number of MapCells available for storing that information. Additionally, a Map will also contain a simple implementation of the A* algorithm with 8-conectivity and Manhattan distance heuristic in order to calculate shortest paths along roads or any other type/types of MapCell [21]. If the start and/or finishing nodes given to the A* algorithm are not accessible nodes, the algorithm will replace them with the closest node possible in order to function fully.

Location: The Location class is responsible for accessing all location-based services and other location related content. It will access a Google Maps Static API image with the given address or, in case an address wasn't given, it will use the built in GPS module of the device in order to determine the player's position and so, get the respective image. Afterwards, it will filter the image, storing a Map representing the information found (such as roads, buildings, etc). It is also responsible for

constantly updating the Player's position in accord with the Map (so that the virtual positioning of the Player matches the real positioning of the player) and also accesses Google's Geo Referencing Web Service to determine the city and country the player is at in order to access the weather web service and so, download the local weather and store it. Other features can be extended, or created in order to fulfill game-specific needs.

As expected from a framework, each of these features may or may not be implemented or used in the location-based game. Furthermore, some classes may be extended or implemented providing the necessary flexibility that any game requires from such a tool. The framework currently provides features that are most common in location-based games, but even a game that needs little to none of the features provided will at least find the usage of the GPS module, accelerometer, web services, simplified.

4 Location-Based Game: Geo Wars

In order to both test and prove the framework's capabilities, a location-based game was created on top of it.

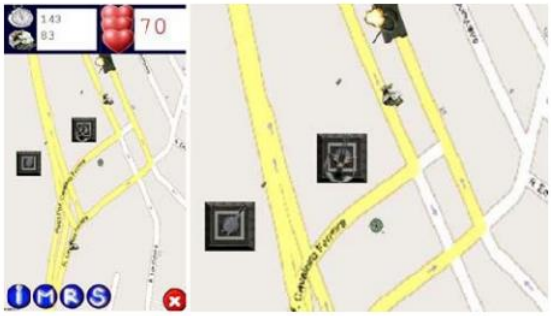


Fig. 4: Geo Wars look and feel

The game, Geo Wars, is a free to play location-based tower-defense game. In it, the player takes the role of a general with the goal of defending his sector (a portion of a map) from enemy forces, depicted in Figure 4.

The player must resist several waves of enemies, either by physically moving around, evading enemy fire or luring them into friendly crossfire situations, or by building defensive towers, each with its unique strengths and weaknesses. The enemy attacks by land, air and sea, using tanks, soldiers, airplanes and cruisers. While tanks and soldiers can only move around streets or parks, aircrafts can move anywhere in the map. Each unit has its own firepower, primary and secondary objectives (some units may prefer targeting specific towers rather than the general) and A.I.. Player uses money to build towers, money that can be gained over time, by destroying enemy units, or by physically moving to the locations of virtual bags of money displayed on screen. The game loads saved games settings via a web-service that

accesses a remote database that contains player-related saved games created using the online portal.

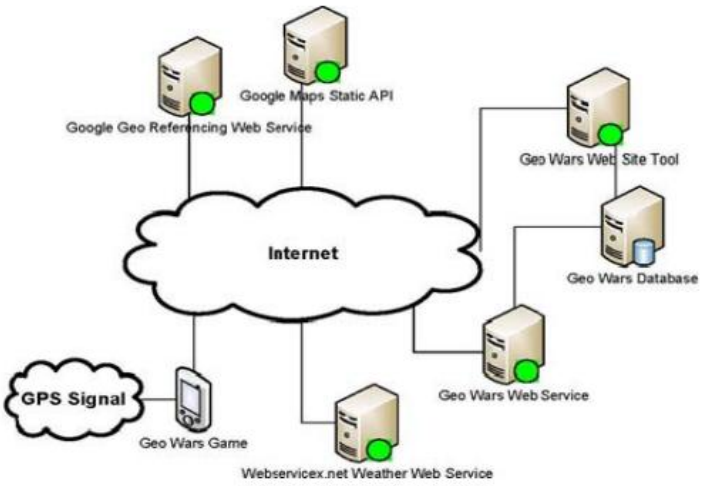


Fig. 5: Geo-Wars Physical Architecture

As far as location-based content, the game uses Google's Geo-Referencing webservice to determine the current city the player is at (in order to download weather data for it), Google map's static API to access the map of the game and Webservicex.net Weather webservice to determine the weather for any given city, as shown in Figure 5.

The game also takes into account weather (increasing the enemies speed if the weather is sunny and slowing it if it's cloudy or rainy), and allows the game to be played indoors by specifying an address for the game to be played or a cached map and by disabling the general's movement (which can make the game a bit harder to complete).

The full description of the game's architecture, features and a more thorough analysis of the users' feedback can be found here [18].

5 Discussion

Regarding the used features of the location-based framework by the Geo Wars location-based game, and comparing them with the code that would be needed to write if the game wasn't based atop said framework, for an instance, using only the directx mobile framework, the gps intermediate driver among other available solutions, it is reasonable to conclude that the framework, for this proof-of-concept game, allowed a faster implementation of the game's concept. In some situations where the framework was too generic for the game, e.g.: when writing the enemies' AI, it still provided some needed features (such as picking/collision detection) with a simple extension of the class Sprite or OrientableObject. Other features that are

probably needed in many location-based games, such as downloading maps, looking up addresses or accessing the player's last coordinates are easy enough to use and extra effort was put into making these features easily available.

Of course, judging from the fact that the only game developed so far using this framework was Geo Wars, it is difficult to determine what other features might this solution be lacking. Still, Geo Wars seemed to be well received by the gaming community [19], and although it would be incorrect to assume that any good game comes necessarily from a good framework, it is important to note that Geo Wars itself was developed in a very short time, thanks to the framework.

However, initially, the game was somewhat simpler, as the player could not choose where to play the game. Fortunately, the framework already allowed the download of a map given a coordinate, so it was only a matter of adding the ability of converting an address to a coordinate, and using said coordinate to play the game.

Performance wise, the framework is capable of a very modest performance, mainly due to the fact of relying on directx mobile for its graphics. Since the directx mobile framework isn't as used as its desktop counterpart, some OEM (Original Equipment Manufacturer) don't optimize their video drivers for directx mobile. However, Geo Wars is still capable of being run at about 20 frames per second in high-end mobile devices.

6 Conclusions and Future Work

During the course of this project, it became notorious that some limitations in location-based games were difficult if not impossible to overcome completely due to their natural unpredictability. As such, some of the mentioned issues were not solved as much as they were avoided. An example of this is the lack of GPS coverage in buildings that was overcome with the feature of allowing the player to play a game that is not based on his location. This doesn't constitute a solution for the problem but an alternative to not being able to play a location-based game at all. However, and thanks to the methodology based on the agile Sashimi model, that allowed to adjust requirements, features and the design of the solution even as other phases of the project went on being developed, many solutions for issues of location-based games that were only spotted during the implementation phase of the project were included in the game's design. Additionally thanks to the creation of a general location-based game solution (the previously mentioned framework, the base upon Geo Wars was implemented), the game's final concept was only elaborated and completed long after a simpler and different location based game was implemented using the very same general design, proving that the generic approach to the creation of location-based games was possible through the said framework.

The game Geo Wars has had a good acceptance in the mobile gaming community as it can be seen in this xda-developers thread [19] used for testing and gathering of feedback to be used in subsequent iterations. So much that the framework is being ported to 3D so that Geo Wars may also be a 3D game. Prototype versions for both 2D and 3D are also available for download in said thread as well as a small video depicting a Geo Wars match. Many of the testers of the game asked for a multiplayer

version of the game. Alas, the framework, as of now, is incapable of allowing such a feature, although it will most likely be altered in order to accommodate that possibility. Additionally, due to the fact of newer, more powerful devices and mobile OS are now surfacing, extra effort is being put in recreating the current framework in order to support Android OS.

References

1. Hunicke ,Robin, LeBlanc ,Marc, Zubek ,Robert , "MDA: A Formal Approach to Game Design and Game Research "
2. Novak, Jeannie (2008), "Game Development Essentials", Delmar Cengage Learning
3. Steiniger, Stefan, Neun, Moritz Edwardes, Alistair (2006) "Foundations of Location Based Services" in "Lecture Notes on LBS", 2006.
4. Locomatrix, 2010. http://www.locomatrix.com/about.html , 10/2009.
5. Groundspeak 2010 http://www.groundspeak.com/ , 10/2009.
6. GroundSpeak.Geocaching, 2010.http://www.geocaching.com/ , 10/2009.
7. GPSgames.org.GPSGames.org, 2009 . http://www.gpsgames.org/ , 10/2009.
8. GPS Games. Minute War, 2006. http://ultimategps.com/gps_fun.html , 10/2009.
9. GPS Games. Shutterspot http://www.gpsgames.org/index.php?option=com_wrapper&wrap=Shutterspot , 10/2009.
10. Navigadget. "Take down your targets! A gps phone tag game." http://www.navigadget.com/index.php/2006/04/04/take-down-your-targets-a-gps-phone-tag-game , 10/2009
11. GroundSpeak, WhereIGo, 2008. http://www.wherigo.com/, 10/2009
12. Pac Manhattan http://www.pacmanhattan.com/about.php, 10/2009
13. Dodson, Sean (2002) "Ready, aim, text", The Guardian, http://www.guardian.co.uk/technology/2002/aug/15/electronicgoods.games, 5/2010
14. Fikkert, Erik (2008), "Zombies, Run" http://www.androidapps.com/t/zombies-run, 9/2010
15. Araújo, Manuel, Roque, Licínio (2009) "Uma proposta metodológica para organizar o desenvolvimento de jogos originais", VideoJogos 2009
16. Joselli, Mark, Clua, Esteban (2009) "Mobile Game Development: A Survey on the Technology and Platforms for Mobile Game Development", VideoJogos 2009
17. Rising, Jim (2009) "Sashimi Waterfall Software Development Proccess", Managed Mayhem, http://www.managedmayhem.com/2009/05/06/sashimi-waterfall-software-development-process/ . (last accessed on June 14th, 2009)
18. Jacob, João,Coelho, António (2010) "Geo Wars – the development of a location-based game", Prisma 2010
19. João Jacob (2010) "[APP] Geo Wars - A location based game for the HD2" http://forum.xda-developers.com/showthread.php?t=694276 (last accessed on January 22th, 2011)
20. Rising, Jim (2009) "Sashimi Waterfall Software Development Proccess", Managed Mayhem, http://www.managedmayhem.com/2009/05/06/sashimi-waterfall-software-development-process/ . (last accessed on June 14th, 2009)
21. Borges, Ricardo, Augusto, Cícero, Volkweis, Mauricio, Konzen Andrea (2002) "Jogos em Inteligência Artificial", Universidade Luterana Do Brasil

Indoor Localization Using Bluetooth

Tiago Fernandes,

Faculdade de Engenharia da Universidade do Porto,
Rua Dr. Roberto Frias, s/n 4200-465 Porto, Portugal
pro10018@fe.up.pt*

Abstract. This article proposes a solution for user localization in indoor areas using the compass, accelerometer and Bluetooth to calculate the user's position within this environment. This application is viewed through a 3D virtual environment representing a simple room. The basis of this solution is the utilization of a mobile Bluetooth-enabled device, such as a PDA, where the application is deployed. The application will then use the mobile phone's Bluetooth to determine the Received Signal Strength Indicator (RSSI) of beacons located within the area. This information is then used to determine the virtual position of the user by triangulation. Additional sensors, such as the accelerometer or the compass provide extra precision and compensate the latency that the Bluetooth positioning solution provides. This solution has proved to be accurate, inexpensive, and very usable, as it uses virtually no input from the user (since the input the user provides is actually passive). Also, it is tolerant to errors and does not conflict with any other Bluetooth devices, such as other mobile phones.

Keywords: Mobile, Indoor, Localization, Bluetooth

1 Introduction

Having accurate information about people's location in indoor environments is crucial for some applications such as e-commerce and e-museums. Solutions such as GPS and GSM location systems, for example, are very inefficient when used inside buildings. Acquiring relaying hardware that could enhance the signal inside these areas, therefore granting a better performance for the location, could solve this problem. However, the required components are usually very expensive, especially since the necessary number of relays would vary according to the area being covered.

This article proposes a low-cost solution for creating a system that is able of providing indoor-location information using Bluetooth and modern handheld devices. The idea behind this project is based on the utilization of small, inexpensive Bluetooth devices that are placed on the area to be covered. These devices are registered on the handheld during a calibration phase and are used for triangulation. However, to enhance system accuracy, the mobile device used for location may also contribute by providing information from other sensors such as accelerometers and digital compass. These extra bits of information are weighted and mixed together to

produce a final calculation over the user position and orientation. Also, this system is targeted to achieve several goals besides the indoor localization of users, such as maintaining user privacy and provide near real-time information, since all the calculations are decentralized from the infrastructure and performed directly on the mobile device. This guarantees that no personal or private data is stored in external servers, since the external Bluetooth devices are only used to broadcast their own address.

This article starts by presenting an overview of the existing technologies in section 1.1. The methodology used for this work is presented in section 2, which is divided in three parts, namely for calibration, localization and alternative data sources in sections 2.1, 2.2 and 2.3, respectively. The obtained results are presented in section 3. Finally, section 4 presents the work conclusions and perspectives of future work.

1.1 Previous work

Some work has been developed for indoor-location, using several distinct technologies. Veljo Otsasson et al. [1] conceived a system that was able of providing user position in inside environments using GSM triangulation. The idea behind this project is to use wide fingerprinting that uses GSM cells that are strong enough to be detected but too weak to be used in communication, in addition to the six cells defined in the GSM standards. This system has many advantages such as the range of signal coverage, the fact that any mobile phone could be used for positioning and that the system would be highly tolerant to power shortages. In order to be able of detecting the user position accurately, this system requires some calibration that was performed by measuring both the 802.11 and the GSM signals in each division of the tested areas. By using the proposed algorithms that held the best results, this solution was able of reporting the user location with a median localization error between 2.5 and 5.4 meters.

In a different perspective, the Cricket project [2] uses both Radio-Frequency (RF) and ultrasound signals to identify a user's position. The utilization of both sensors is based on the fact that RF propagates in non-linear and possibly unpredictable ways inside buildings. Therefore, it was necessary to consider alternative ways of providing increased precision to the position calculation. So, to perform the calculations, the beacons send concurrently RF and ultrasonic signals. As the speed of sound is smaller than the transmission speed of RF signals, the later will arrive sooner to the listeners. When a listener receives a RF, it uses the first bits as training information, enables the ultrasonic receiver and waits for the ultrasound emitted by the beacon. The calculation is then performed by using both the strength of the RF signal and the time difference between the arrivals of each signal. One of the great advantages of this project is the low cost that is required to buy all of the components. The error rate reported for mobile devices is however, somehow big, being around 20-25%.

HP also developed a solution that uses infrared beacons instead of ultrasound and typical RF emitters, called HP Cooltown [3]. To find its location, the user must point its infrared-enabled handheld device to the infrared beacons. This has the clear problem of requiring user interaction to work, but on the other hand, this method also

protects the user's privacy, since he only interacts with the system when he really wants to.

Finally, F. J. González-Castaño and J. Garcia-Reinoso [4] developed a system that attempts to provide user location in indoor environments using only Bluetooth devices. This proposal uses a network of bluetooth devices, organized in hexagonal grids. Each node is either a slave or a master node. The user is equipped with a Bluetooth enabled or Bluetooth badges and broadcasts its address to the nodes. Every slave node receives the RSSI value from the user and sends it to the master node. The master node performs every calculation to triangulate the user position based on the RSSI values that were received as well as the slave nodes positions and sends the computed data to some servers that will use that information for some service. This approach is very expansible since the system is able of auto-configuring itself automatically. Also, there are no collisions with other existing devices, because the work is centralized on the slave and master nodes which conform to a specific protocol. However, given that all the calculations are done by the master nodes the system may become quickly overloaded which brings performance problems in terms of response times, depending on the number of position calculations and Bluetooth devices in the network.

2 Methodology

The first step towards the resolution of this problem was to create a solution that was able of receiving any type of sensor data and return a position. For that, it was necessary to have some calibration results from the sensors, so that the range of values was known and the distances that those values correspond to. By doing so, it is possible to compute a linear calculation based only upon these two points. However, if possible, some intermediate values could be used for a more precise interpolation, if needed, increasing the overall accuracy of the solution. Therefore, prior to developing the solution itself, a calibration tool was needed.

2.1 Calibration

A simple calibration tool was developed with the single purpose of selecting the sensors that will be considered by the application, so that it does not conflict with other Bluetooth devices. This simple application finds every Bluetooth device that is detectable in the area that is being tested and lists it on the device. To perform calibration, the user must select each of the relevant devices and save the RSSI value for each distance that is to be used. The values saved in the calibration tool will then be exported in an XML file that specifies which sensors are to be used and also some other information that is to be used within the application, such as the position of points of interest (POI) and the 3D model filenames.

It is important to notice that the calibration distances are not fixed. The granularity of the measurements and the distance values that are to be used in calibration may vary according to the environment and the Bluetooth devices being used. This

calibration step only needs to be performed once and the application is ready to be used with the same XML file in every other run, unless maintenance is required.

2.2 Localization

To find the user's location, the application starts by an initialization phase in which the data that was saved in the XML file is read. As mentioned above, the XML file contains a set of RSSI-distance value pairs for each sensor. These values are read onto a hash map during the initialization. When the application loads, the Bluetooth receiver is enabled and the device starts to look for nearby known devices that have addresses that were registered within the XML file during the calibration. If at least three values are found, the software uses a triangulation algorithm that uses the obtained values and calculates the user's position. For each of the sensor values found, the application searches the hash map for the calibration points read from the XML file and finds all the intervals in which the obtained sensor value is contained. Since the RSSI variation is much greater when the device is near the Bluetooth beacon [5][6], the algorithm weights the possible found intervals by giving slightly more relevance (~10% more) to the calibration values that are more distant.

Each of the RSSI values obtained represents the radius from a circle, whose center is the position of the sensor that originated that value. In a first step to perform triangulation, two of those RSSI values are used, as well as the positions of the respective sensors. If the circles intersect then three points are saved, namely the two intersection points and the point where the line defined by the circle centers crosses the line defined by the two intersection points. With these three points saved, the algorithm searches for a third sensor that defines a circle that intersects the previous two. To do so, it simply checks if any of the intersection points calculated previously is within range from the sensor's radius. If that is the case, the intersections between the first and third circle and between the second and third circle are calculated. Finally, the midpoint of each of these intersections is used to calculate an average for the user's position.

There are two clear problems with this algorithm. The first problem appears if one of the circles is too small, there will be no intersection with any other two circles. If this happens, it means that the user is actually very close to the sensor itself and that is the reason for such a small radius. To solve this case, the user's position is snapped to the sensor position as long as it is within the map-defined bounds. The precision of the value that triggers this snapping behavior is controlled by a parameter that can be set in the application configuration. However, there is a second problem. Even if the circle is not sufficiently small to cause the user's position to be snapped, there might be times where one circle is contained inside another circle. These cases happen when the user is relatively close to one sensor, and far away from other sensor that has a very wide range. Also, this might also happen when the sensors are too close to each other and the user is close to one of the sensors. This causes a logical problem since the intersection points can't be calculated. But, since this means that the user is close to one of the sensors, the solution is to ignore the intersection calculation (since it

wouldn't be possible, anyway) and consider that the central intersection point is the center of that circle.

The final position is calculated by using the values obtained from the hash map inside the triangulation algorithm. This algorithm produces a final set of coordinates based on the interpolation of the three points that were found, if any, using the method described above. This process guarantees that the coordinates conform to the specific environment, in a somehow pessimistic form, since the farthest values have more weight than the closest.

2.3 Alternative Data Sources

Since the Bluetooth takes some time to obtain the updated RSSI values, it was necessary to compensate the utilization of these values in the meantime with values from other sources. The modern PDAs and mobile phones are usually equipped with accelerometers and digital compass. By using the accelerometer data, it was possible to develop a rather sensible pedometer that indicates if the user has walked. Additionally, the accelerometer data also allowed for the creation of a module, capable of determining the 3D orientation of the device. Mapping this orientation of the device with the OpenGL camera, allowed for a pseudo augmented reality interaction with the virtual environment. Note that the user is only providing passive input to the device, meaning that virtually no experience is needed with the handling of the device, or the application. The walking direction is given by the digital compass and introduced into the application. So, in case the Bluetooth fails or the handheld takes a long time to receive RSSI values, the user position is updated with the information provided by both the pedometer and the digital compass.

This raises a significant problem that can't be overlooked. With the pedometer it is possible to know if the user walked and the compass provides the direction he was facing. However, it is impossible to know if the user walked backwards. Theoretically, this could be read from the accelerometer values, but due to the variations that are usually read, it would be very difficult to know accurately if the user actually moved backwards. In these cases, the only solution is to wait for the Bluetooth triangulation to reposition the user to the correct position, or try to use other sensors such as Wi-Fi or even the camera, although this is out of the scope of this project.

3 Results

To test this project, an application was built to reproduce the user movements inside a 3D environment, by using an HTC HD2, running a custom Android build and OpenGL ES [8]. The virtual camera placed on the scene turns automatically its direction to face the correct direction when the user moves the mobile phone, by using values read from the accelerometer and compass values. These values are submitted to a simple filter to avoid unwanted noise and to increase precision. The device was

calibrated for each of the sensors, registering its Bluetooth address and the values at several distances that were previously marked. These results are presented on Figure 1, for one of the sensors, in terms of dBm and distance. The distance is measured in centimeters and, for convenience the Bluetooth dBm values are represented with its absolute values.

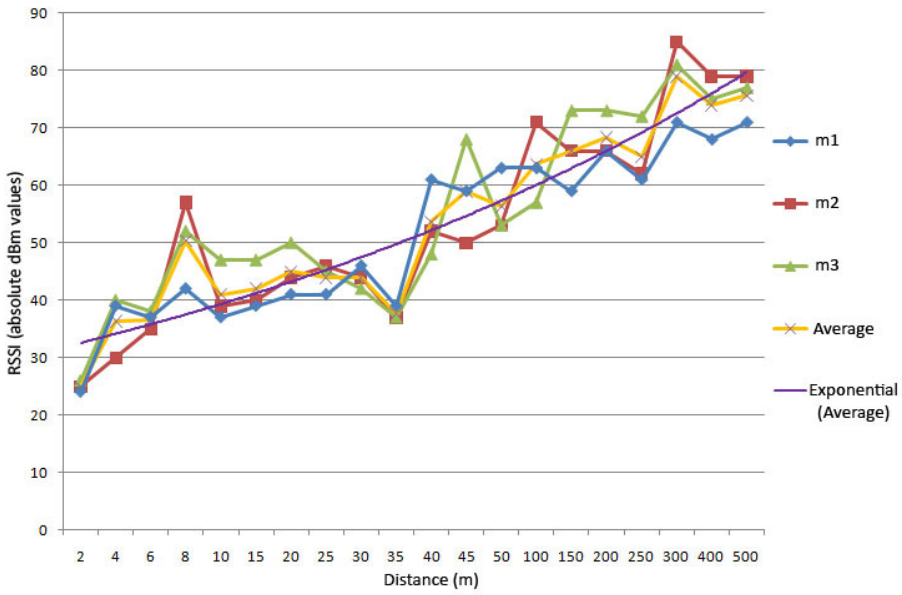


Fig. 1. RSSI values read during the calibration phase, according to several different distances in three distinct measurements. The graphic also presents an average calculation for the values and an exponential curve calculated from the average points.

The tests took place in a room with 71 m^2 with a near-square shape. This room was modeled in 3D and exported to the .obj format in order to be easily imported by the application. In the virtual environment, the whole scene transforms correctly according to the angle in which the user is pointing the device with minor differences of ~2 degrees. Also, the pedometer works very well if the model has the correct dimensions or the step size and the sensitivity are correctly configured for an average person. Even without using Bluetooth localization, if the starting points for both the virtual scenario and the user are aligned, the system is able of tracking the user with great precision (< 1 meter, for a correct step and sensitivity calibration). The room scheme and the approximate Bluetooth beacons positioning is presented in Fig. 2.

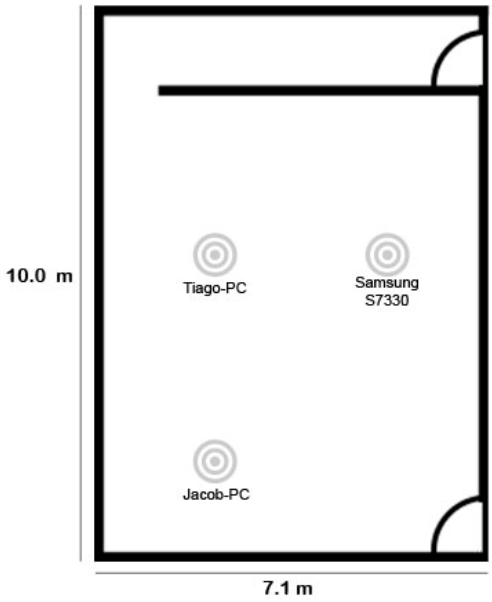


Fig. 2. Room scheme and Bluetooth beacon approximate positioning.

The Bluetooth location system is able of positioning the user correctly in a scene. However, during the testing sessions, some noise was registered making the virtual camera jump from one location to another at some times. This effect was reduced by using noise filters during signal capturing and also by using the position history to infer the probable position of the user. This avoids all of the signal peaks that are sometimes registered due to noise and signal reflection, but are still inefficient when the signal varies more noticeably for longer periods. Such problems happened mostly when using a Samsung S7330 as a Bluetooth beacon, since the signal emitted by this device is very unstable and has very large fluctuations. However, although the test space was relatively small (and the number of Bluetooth devices was also small, since there were only three testing units) the positioning with general Bluetooth devices does not have any type of collision problem. This happens because the addresses are registered during a calibration phase, and provide indoor-localization with high accuracy (~1.5 meters) when combined with another sensors. The final results are shown on Table 1. The accuracy for each method in each case is presented on Table 2.

Table 1. Obtained distance measurements using different methods and devices

Reading	Linear (m)	Inv. Quad. (m)	Realistic(m)	Device	Real Dist. (m)
1	0.28	0.83	0.306	Jacob-PC	0.40
1	3.73	4.34	3.5	Tiago-PC	3.20
1	4.68	4.41	3.6	Samsung	4.0
2	0.32	1.26	0.3167	Jacob-PC	0.40
2	3.73	4.34	3.5	Tiago-PC	3.20
2	4.89	3.89	3.0	Samsung	4.0
3	0.32	1.26	0.3167	Jacob-PC	0.40
3	5.41	5.57	5.0	Tiago-PC	3.20
3	5.32	6.33	4.25	Samsung	4.0
4	0.32	1.26	0.3167	Jacob-PC	0.40
4	3.73	4.34	3.5	Tiago-PC	3.20
4	4.89	3.89	3.0	Samsung	4.0

Table 2. Accuracy percentage for each of the different methods and devices

Reading	Linear	Inv. Quad. (m)	Realistic(m)	Device	Real Dist. (m)
1	70.0%	48.2%	76.5%	Jacob-PC	0.40
1	85.7%	73.7%	91.4%	Tiago-PC	3.20
1	85.5%	90.7%	90.0%	Samsung	4.0
2	80.0%	31.7%	79.2%	Jacob-PC	0.40
2	85.7%	73.7%	91.4%	Tiago-PC	3.20
2	81.8%	97.3%	75.0%	Samsung	4.0
3	80.0%	31.7%	79.2%	Jacob-PC	0.40
3	59.1%	57.5%	64.0%	Tiago-PC	3.20
3	75.2%	63.2%	94.1%	Samsung	4.0
4	80.0%	31.7%	79.2%	Jacob-PC	0.40
4	85.7%	73.7%	91.4%	Tiago-PC	3.20
4	81.8%	97.3%	75.0%	Samsung	4.0
Average	79.2%	64.2%	82.2%	---	---

The above tables represent a usual situation where the mobile device is able to pinpoint its real location thanks to some Bluetooth devices nearby, recognizable thanks to the XML file. The four readings were made without moving either the mobile phone or the devices. However there was a considerable discrepancy in the third reading. This was due to a person being between the mobile device and two of the beacons, effectively altering the read values. Still, these values were actually very accurate. The latency between readings would vary between 5 and 10 seconds, varying from 2 to 1 Hz refresh rate respectively, which is acceptable, for indoor navigation. It is important to notice that the measurements took place in a room with 5 to 10 persons, moving around without any pattern. During the tests, some other Bluetooth devices entered the room or walked nearby. These facts created some noise in the positioning and beacon detection but not sufficient to invalidate the calculation. This was mostly due to the fact that a medium number of points (20 RSSI-distance pairs) were used in the calibration.

4 Conclusions and Future Work

The proposed methodology described above achieves the goal of providing a simple, inexpensive and ubiquitous indoor localization solution. It is capable of pinpointing the user's location with reasonably high accuracy. Also it provides an alternative form to bypass the Bluetooth location technique's high latency by using the accelerometer as a pedometer.

The solution described above heavily depends on a calibration phase that could require too much time and effort to perform without a tool designed for that purpose. However, the utilization of the small calibration software greatly reduces the required time and expertise to configure the system, making it accessible to users without great computer knowledge and proficiency. The application is not designed to be auto-configurable, since it needs to know at least which Bluetooth sensors shall be used and two RSSI values for two given distance points to perform the triangulation. After the calibration has been done, there was no need to perform it again, even when other Bluetooth devices entered the area. Since the application knows which addresses shall be considered, there are no possible collisions between other Bluetooth devices. There exists, however, some noise due to the influence of other Bluetooth emitters, creating some fluctuations in the readings. These fluctuations are also noticeable when there are more persons in the room, increasing when they are moving, since this affects the reflection of the signals. Nevertheless, this problem proved to be irrelevant due to the fact that in most situations, the PDA is in range of more than three well-known Bluetooth devices, therefore using all of the values to compensate for eventual errors. Even when the application is run with only three sensors, the effect of the noise induced by people and from other Bluetooth devices greatly depends on the calibration. Finding more RSSI-distance pairs during the calibration makes the application more reliable and error-resistant.

According to [7], the RSSI values are more inconstant and vary greatly and in possibly unpredictable ways depending on the environment and devices being used. Instead, the Bit Error Rate (BER) or Link Quality (LQ) metrics should be used for greater precision. Yet, this was not possible, since the underlying software did not provide any access to these indicators and therefore, RSSI had to be used instead. Attempts were also made to use the iOS Bluetooth features to develop to an iPhone, but it also wasn't possible to obtain access to the BER and LQ information. From the methods depicted in Tables 1 and 2, the linear method also has a reasonable performance, but is much more susceptible to the signal fluctuations induced by noise and possible reflections. The inverse of quadratic function presents the worst results, although in some cases the values are much better than the other methods and this method performs better when the user is farther from the beacons. The realistic method, which calculates the position based on several points obtained during the calibration held the best results, with an average accuracy of 82.2%.

Furthermore, the user's interaction with the application is greatly simplified, as the user needs only to aim the phone at the areas from which he wishes to receive more information from. This has been proven to be quite a successful feature, although it was only tested with few individuals. Some of these individuals didn't possess any kind of technological background. However, all of them were able to grasp the applications concept and the ways to use it with no problem. Further testing of the

application will be required in order to fully validate the interaction paradigm. Additionally, an increase in performance and precision is also considered needed, in order to make the user's experience as seamless as possible.

This solution could be improved by using computer-learning algorithms to increase user localization precision, especially when the Bluetooth signals are weak and/or combined with accelerometer information. The software could use this method to learn how much the virtual position should be changed according to the accelerometer values, when the application is waiting for new Bluetooth RSSI values. Also, this information could be complemented with a better filter for the accelerometer and prediction algorithms, typically used in network games, to compensate for possible loss of signal.

References

1. Otsason, V., Varshavsky, A., Lamarca, A., Lara, E.D.: Accurate GSM Indoor Localization. UbiComp 2005, 141-158 (2005)
2. Priyantha, N.B., Chakraborty, A., Balakrishnan, H.: The Cricket Location-Support System. ACM MobiCom (2000)
3. Kindenberg, T., Barton, J.: A Web-Based Nomadic Computing System. Hewlett-Packard (2000)
4. González-Castaño, F.J., Garcia-Reinoso, J.: Bluetooth Location Networks. IEEE GlobeCom'02 (2002)
5. Hightower, J., Borriello, G.: Location Systems for Ubiquitous Computing. IEEE, 57-66 (2001)
6. Dishongh, T.J., McGrath, M.: Wireless Sensor Networks for Healthcare Applications. Artech House, Massachusetts (2010)
7. Ørbæk, P.: Positioning and Location Technologies, http://www.daimi.au.dk/DIS/materialer/positioning-tech.pdf (2005). (last viewed on 12/12/2010)
8. Android: The Developer's Guide, http://developer.android.com/guide/index.html (last viewed on 21/12/2010)

SESSION 6

MOBILE COMPUTING / NETWORKS

Chairman: Pedro Filipe de Monteiro Rocha

Rui Chilro, Ana Ferreira, Bruno Oliveira and Ricardo Morla
Intermittent connection effect in the Message Ferry Delay Tolerant Network

Carlos M. D. Viegas and Francisco Vasques
Real-Time Communication in IEEE 802.11 Wireless Mesh Networks: A Prospective Study

Paulo Neto
Demystifying Cloud Computing

Pedro Moreira da Silva, Jaime Dias and Manuel Ricardo
Survey on Privacy Solutions at the Network Layer: Terminology, Fundamentals and Classification

Intermittent connection effect in the Message Ferry Delay Tolerant Network

Rui Chilro[1], Ana Ferreira[2], Bruno Oliveira[1,3], Ricardo Morla[4]

[1] Faculdade de Ciências da Nutrição e Alimentação da Universidade do Porto, Portugal
[rchilro, bmpmo]@fcna.up.pt
[2] Center of Informatics and CINTESIS – Center for research in health information Systems and technologies, Faculdade de Medicina da Universidade do Porto, Portugal
amlaf@med.up.pt
[3] LIAAD, INESC Porto L.A., Portugal
[4] INESC Porto, Faculdade de Engenharia, Universidade do Porto, Portugal
ricardo.morla@fe.up.pt

Abstract. Delay Tolerant Networks (DTN) give a base of communication that permits the delivery of data within harsh environments or even between networks without interconnection. The connection between networks happens when some nodes cross between them and carry information. Our model consists of the Message Ferry Mobility Model (MFMM) where all nodes in the network are confined to their village except one mobile node (the message ferry). In order to reduce the energy consumption the wireless interface is not always on. We aim to study how much does the on/off state of the wireless interface reduces the total connection time available in the MFMM. To answer this question, we created a simulator that uses the MFMM. The simulator generated mobility traces of all nodes and measured the contact time under different patterns of the on/off cycle. As expected the contact time are longer when the interface is active more often. However we found an unanticipated reduction in contact time vs active wireless interface ratio. Hence, we concluded that in the MFMM, the on/off duty-cycle of the wireless interface influences the total contact times.

Keywords: Delay Tolerant Network, DTN, mobility model, Message Ferry, metrics, wireless interface

1 Introduction

Places difficult to access in developing countries, harsh environments, space, wildlife watch, natural disaster zones are cases where full linked networks may not be feasible. In particular, connection between mobile equipment, or even internet access may not be possible all the time. Delay Tolerant Networks (DTN) approach this problem by giving the bases of communication to deliver messages or data even when there isn't a direct communication between source and destination. This is done through moving middle nodes that facilitate this contact, thus allowing the inter-exchange of data.

1.1 DTN

Wireless networks are getting more and more popular, allowing the carrier of a mobile device the capacity to connect from multiple places to the Internet. However, hot spots are not available everywhere. Even in if a hotspot is present, access to the information required may be delayed due to, e.g. the volume of users, short fails, small delay or low error rate. These problems belong to the "well behaved" cases when compared to the scenarios approached by Delay/Disruption Tolerant Networks (DTN). In DTN there isn't always end-to-end connection in real time, that can be caused by nodes mobility or short radio range, physical obstacles, scarce resources, etc [1]. There are limitations that the standard TCP/IP protocol won't be able to work with caused by special requirements such as: harsh environment, remote locations, frequent connection fail, long delay [2]. On the other hand, there could be limitations to the resources like: size, radio capacity, small memory, low bandwidth, low processing, millions of nodes, small batteries [2]. They can use exotic methods of communication, be military networks on hostile environment or war, environmental limitations, intentional interference, security [2]. These networks can still be influenced by delay, communication errors, security and confidence, reliability, etc [2].

DTN solve this kind of problems using a message exchange technique [3]: each node stores-and-forwards the messages he collects. In this way there is no need to have an end-to-end connection, since consecutive point-to-point connections will suffice. The creation of the bundle concept, the unit of transmission, and the inherent protocol were initially defined in 2 RFCs from 2007, RFC 4838 [4] and RFC 5050 [5]. The protocol describes a way to transfer custody of a message, priority, state report and security, among other information. There was also initially created a standard for the representation of numbers: the Self-Delimiting Numeric Values (SDNV) [6] that are able to represent huge numbers.

The radio wireless interface is the main mean of transmission in DTN, although other means of communication are supported. This resource spends energy essential to the mobile device, so all processes that can reduce the consumption of battery are important. The deactivation of this module can save much energy at the cost of making communication impossible for that period of time. Our motivation is the study of the implications of this temporary deactivation in the data transmission. The simulation is an easy way to explore scenarios and patterns of mobility for wildlife, people or vehicles, gathering the needed data and metrics.

1.2 Mobility models and metrics

Mobility models are simplified representations of reality that describe the pattern of movement of a known group. These models can always be empirical models of mobility.

There are innumerous mobility models ([7], [8]) so we choose to use one model that represents the principle of DTN, the Message Ferry Mobility Model (MFMM) also known as Mule Mobility Model [9]. There are many variations of this model but the basic one has some areas (villages) were the units ("villagers") move around

based on the Random Waypoint Model [10]. The "villagers" never enter in contact with other villages. There is one other unit that moves between villages and enters in contact with the locals exchanging with them the messages received earlier. Then he sets of to the next village carrying the messages received. This is why he is also called the "Mule".

This scenario is frequently used in sensor networks to recover data and remote villages like in the projects RuralKiosk[11] and Wizzy Digital Courier [12].

1.3 Objectives:

The aim of this study is to discover how much does the on/off state of the wireless interface reduces the total connection time available in the Message Ferry Mobility Model. To this end it was created a simulator, patterns of duty-cycle to the wireless interface and a wireless interface component to evaluate the contact time metric and total contact time with the generated traces of units.

In the next section we describe the methodology used, followed by the results obtained. The discussion and the conclusions end this document.

2 Methodology

With the reduction of active wireless interface time it is expected to reduce the time of contact between the units. From the many metrics available to analyze the impact of the patterns created we chose the metric total contact time. This metric is the sum of all the contact times between units and should only sum the contact time between A and B and not accumulate both times. This is the time the units have to transfer data so it is an important measure.

The patterns' usage are based in an automatic on/off state of a wireless interface so they were thought for the 1/3, 2/3 and 1/6 spare of battery time. The first pattern (duty-cycle) is the base of comparison ("always on") (a), 180s off followed by 120s on (b), 180s off followed by 60s (c) on, 360s off followed by 120s on (d), 360s off followed by 60s on (e), and random (f).

For these duty-cycles, all units are on or off at the same time and for the last one (f) each unit uses a uniform distribution function to calculate the on/off times with a maximum of 10% of the simulation time for each phase.

Before initiating the collection of data it was necessary to determine the simulation area and the total simulation time. Usually researchers choose areas of simulation close to squares of 1000m [13-16] and some hours for the simulation time. In this case we choose 3600s of simulation time [17].

The number of units to use in each simulation is suggested to be in the order of 40 units [16,18]. This value permits almost the coverage of all simulation area making most units interconnected with each other all the time. Since this is not one of the

characteristics of the DTN networks we defined a 10 unit limit so 1/3 of all simulation space could be covered at max.

Since the number of possibilities increase at each parameter to control, we set the maximum speed of the village units to be 10% of the message ferry unit. The speed function is calculated using the speed and direction defined by the model.

The coverage area of each antenna was set to 100m and after this range there is no contact. It isn't considered the possibility of better signal and better transfer rates at closer range.

The units are initially placed randomly inside their village area and the ferry starts from the center of one village.

Random numbers are generated using the uniform distribution.

The mobility model chosen (Message Ferry Mobility Model) was initialized with 3 villages with 3 inhabitants each.

Two scenarios were created: one scenario with the message ferry unit circulating at a maximum speed of 20m/s; and a second scenario where the unit is set to a maximum speed of 30m/s (close to a car speed limit). This way the units in the villages will be at a speed of 2m/s or 3m/s which is closely the walking speed.

3 Results

The platform developed was based in the client-server architecture and presents a page for the user to choose the parameters for the simulation and to receive the generated results. There are parameters for: speed; size of simulation area; mobility model to process; number of units; time of simulation; etc. (see Fig. 1). the browser should support the Adobe SVG plugin for the movement to be generated.

The platform was created in PHP and all the processes are executed in parallel. It supports batch executions for long simulations with different parameters.

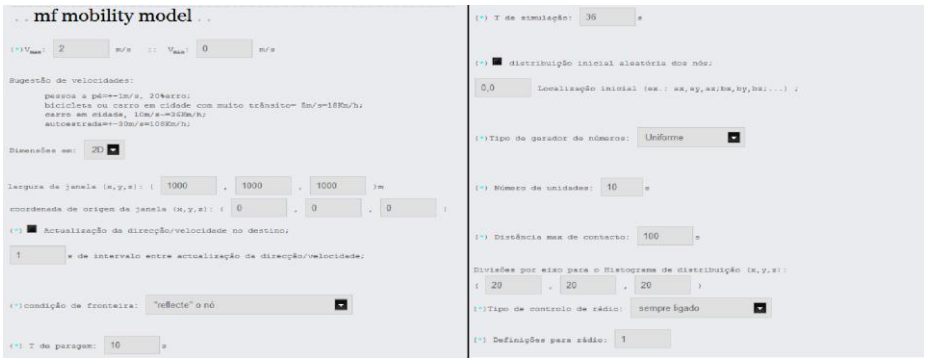


Fig. 1. Window of parameters from the simulator (snapshot).

Figure 2 presents the class model of the created platform. The main class creates an area of simulation with the defined parameters, sets the clock of simulation and creates one or more Mobility Units (MU), its attributes and associated to this MU the

chosen mobility model and radio pattern. Then it starts by calling the method *proxposicao* from class *modelo* in each MU to create the movement for each unit until the simulator time expires. Each MU can be a process executing in parallel.

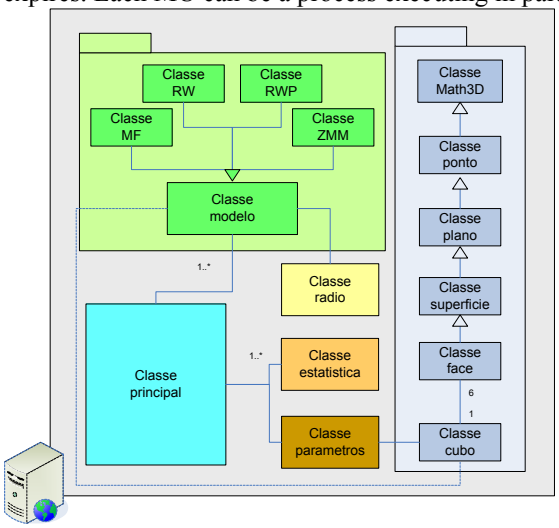


Fig. 2. Class Diagram for the simulator prototype.

After the creation of the traces, the main class launches processes to get the statistics needed. This task may take a long time. In the following figure (Fig. 3) we present 2 histograms of positions taken from the 100 simulations of data generated.

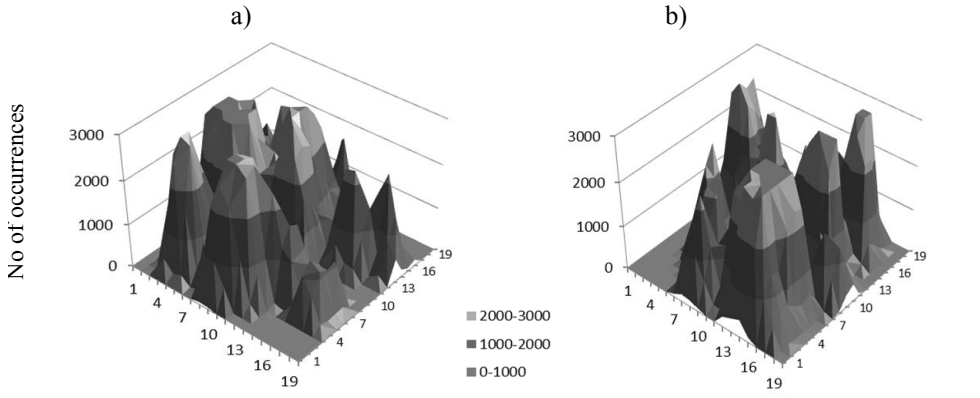


Fig. 3. Histogram showing the 100 simulations for the message Ferry model with the MF moving at 20m/s (a) and at 30m/s (b) in an 1s interval (scenario 1).

In the following picture (Fig. 4) there is a snapshot of the application showing a 360s of one of the simulations. We can see clearly the 3 villages and some occupants moving around. This is a typical movement in the message ferry model.

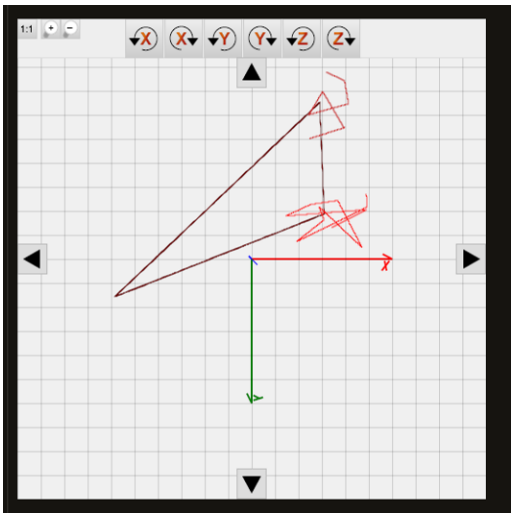


Fig. 4. Simulator showing a typical 360s movement of the Message Ferry Model.

The following table (Table 1) contains the contact time and intercontact time that is, the time between contacts for the simulations. Each column subdivides in the 2 scenarios created. It indicates the minimum and maximum for the 100 simulations, the sum, average and standard deviation. This table indicates that in average, the contact time between units in MF 20/2 is approximately 2min. this is the time available to transmit data between mobile units. If we change to the MF 30/3 there is a slight reduction in the average time.

Table 1. Contact time and intercontact time for the 2 scenarios with some statistical values.

	Contact time (seconds)		**Time without contact (seconds)**	
Vmax (m/s) of MF and village units	**20 and 2**	**30 and 3**	**20 and 2**	**30 and 2**
Minimum	1	1	1	1
Maximum:	1691	1758	3581	3083
Sum	293398	333746	421265	507667
Average	128,01	104,04	210,32	163,98
Standard deviation	175,00	144,40	322,66	287,03

The next two tables (Table 2 and 3) indicate the deviation of the contact time if we apply the duty-cycles on the wireless interface. The parameters from the previous table were maintained.

Table 2. The contact time (in seconds) for various automatic patterns applied to the wireless interface (scenario 1).

	Always on (a)	180s off 120s on (b)	180s off 60s on (c)	360s off 120s on (d)	360 off 60 on (e)	random (f)
Minimum	1	1	1	1	1	1
Maximum	1691,0	121	61,0	121,0	61,0	281
Average	133,8	59,73	38,5	55,8	36,8	45,98
Sum	293398	151354	62329	62266	29154	45581
No of contacts	2292	2567	1636	1135	805	998
Standard deviation	170,6	41,51	20,7	39,5	20,3	46,77

Table 3. The contact time (in seconds) for various automatic patterns applied to the wireless interface (scenario 2).

	Always on (a)	180s off 120s on (b)	180s off 60s on (c)	360s off 120s on (d)	360 off 60 on (e)	random (f)
Minimum	1	1	1	1	1	1
Maximum	1758,0	121	61,0	121,0	61,0	284
Average	109,9	53,53	35,7	50,8	35,9	41,87
Sum	333746	171412	70866	71085	33553	47214
No of contacts	3208	3292	2015	1448	950	1165
Standard deviation	140,0	39,24	20,5	37,6	20,0	40,42

The next table (Table 4) resumes the last ones indicating the loss of total contact time in percentage for each scenario. The pattern is indicated on top.

Table 4. Percentage of lost contact time.

	180s off 120s on. (b)	180s off 60s on (c)	360s off 120s on (d)	360s off 60s on (e)	Random (f)
Ratio	**2/3**	**1/3**	**1/3**	**1/6**	~1/2
MF 30m/s and villagers at 3m/s	**48,64%**	**78,77%**	**78,70%**	**89,95%**	**85,85%**
MF 20m/s and villagers at 2m/s	**48,41%**	**78,76%**	**78,78%**	**90,06%**	**84,46%**

4 Discussion

The mobility model chosen has a big concentration of units in specific zones, called villages. We can see on Fig. 4 the typical example of this situation and in Fig. 3 the over-position of all simulations.

The histogram shows both zones with high density with abrupt reductions and zones clearly empty near the borders. In this case there is a great possibility of having long life contacts.

This model doesn't have border issues (units don't try to leave the simulation area) since the units are limited to their zone of activity and the Random Waypoint (RWP) and message ferry don't permit this action.

4.1 Patterns (duty-cycle) on/off

To the 100 simulations in the Message Ferry Mobility Model (MF) it is clear that there is a reduction of contact time when we reduce the active time of the wireless interface. For 2/3 of reduction of active time the loss is close to 50% and to a reduction of 1/3 the loss is close to 80%.

The size of the contact time block is close to the average which suggests that the speeds don't have interference in the size of these blocks.

The slow movement of the units in the village may justify why the maximum time connected is the same as the active state. This slow movement may justify why the average is always bigger in scenario 1.

Looking to the number of contacts these are always higher in scenario 2. The speed makes the units move faster, making them break contact more often but increases the likelihood of having a contact. This is a very important factor to the DTN networks because it increases the possibility of delivering the messages since it is more likely to contact with everyone in the network.

The total contact time is always higher in scenario 2 so this may have some relation with the number of contacts. The speed may reduce the average contact time but the possibility of another contact adds more time to the total time and optimizes the usage of the active time.

The patterns (c) and (d) in Table 2 and 3 are both 1/3 of the time active, but the active time is the double from the first to the second. The objective was to see if there was a significant difference between active times with the same ratio. Clearly all the values from Tables 2 and 3 are closely the same with the exception of the number of contacts. This metric is always inferior in the case of (d) but almost half in the scenario 2. It seems that the speed greatly affects this pattern so it would be better to use short active times than long ones.

For the random pattern only the maximum contact window is greater than any other statistical value. All the patterns make the interface active at the same time so

all the units in the vicinity can be contacted during the active phase. In the case of the random pattern, one unit may have their interface active but the other unit passing by may not. This behavior seems to clearly reduce the performance of the pattern.

In Table 4 the values are very close so there is no clear best result from the data collected. From the results obtained it seems that the random pattern isn't the right choice for this mobile model. The best values were collected by the 2/3 pattern although they presented a 48% loss in contact time.

4.2 Simulation platform

The created platform permits the easy inclusion of new mobility models by just extending the class *modelo* and the implementation of the function *proxposicao*. This function is called at each simulation unit of time.

Each simulation has only one parameter changed so it was a good way to obtain the greatest number of statistics in the same computational time. This has the drawback of not being possible to analyze the interactions between parameters. The choice of 2 speeds allowed to reducing the number of simulations to 100 in a short period of time.

The wireless interface was created independently of the mobility model class to permit the abstraction of all models. This has 2 advantages: the movement was created to each simulation and afterwards all the patterns were applied to the same movement without the need to repeat the simulation; and in the generation of the statistics it is possible to query the state of the interface at a given time.

5 Conclusions

The work developed allowed the creation of a simulation prototype to which it can be added any mobility model and calculate the metrics shown with the added option of choosing the duty-cycle of the radio interface.

We can conclude from the data collected that there is a big influence of the patterns used on the total contact time in the mobility model chosen. The random pattern isn't a good choice but the 2/3 pattern reduces the total contact time to roughly 50% but it may allow significantly energy savings in every mobile unit of this model.

5.1 Limitations

The study limits the range of the wireless interface abruptly and for now it does not include obstacles or models with obstacles.

Although the two chosen speeds allow generating faster results it is not possible to see the interaction between the metrics and the speed variation.

Furthermore the data collected does not take into account the startup of the interface and the exchanged protocols so, the connections of 1s may produce one contact but do not permit the transaction of messages.

5.2 Future work

This study need to be extended in order to get more generalized results and include more models to be analyzed.

The data have to be further processed to apply statistical tests and prove mathematically the results.

Furthermore the platform is capable of generating movement in 3D and can be integrated with geographical maps. The integration with other simulators like the NS2 [19] may permit the inclusion of protocol issues to the study.

References

1 Guo, H., Li, J., Washington, A. N., Liu, C., Alfred, M., Goel, R. et al.: Performance Analysis of Homing Pigeon based Delay Tolerant Networks In: Military Communications Conference, 2007. MILCOM 2007. IEEE, pp. 1-7 (2007)

2 Kevin, F.: A delay-tolerant network architecture for challenged internets In: Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications Vol. 33, pp. 27-34. ACM (2003)

3 Warthman, F. Delay Tolerant Networks (DTNs) - A tutorial. Report No. 1.1, (2003).

4 Cerf, V., Burleigh, S., Hooke, A., Torgerson, L., Durst, R., Scott, K. et al.: Delay-Tolerant Networking Architecture, <http://tools.ietf.org/html/rfc4838> (Abril 2007) [cited in 2008-06-05].

5 Scott, K.,Burleigh, S.: Bundle protocol specification, RFC 5050, <http://www.rfc-editor.org/rfc/rfc5050.txt> (November 2007) [cited in 2008-06-05].

6 Eddy, W. M.: Using Self-Delimiting Numeric Values in Protocols, RFC, <http://tools.ietf.org/html/draft-irtf-dtnrg-sdnv-00> (September 17, 2007) [cited in 2008-06-05].

7 Zheng, Q., Hong, X.,Ray, S.: Recent advances in mobility modeling for mobile ad hoc network research In, pp. 70-75 (2004)

8 Camp, T., Boleng, J.,Davies, V.: A survey of mobility models for ad hoc network research. Wireless Communications and Mobile Computing 2, 483-502, (2002)

9 Shah, R. C., Roy, S., Jain, S.,Brunette, W.: Data MULEs: modeling a three-tier architecture for sparse sensor networks In: Sensor Network Protocols and Applications. Proceedings of the First IEEE. 2003 IEEE International Workshop on, pp. 30-41 (2003)

10 Ariyakhajorn, J., Wannawilai, P.,Sathitwiriyawong, C.: A Comparative Study of Random Waypoint and Gauss-Markov Mobility Models in the Performance Evaluation of MANET In: Communications and Information Technologies, 2006. ISCIT '06. International Symposium on, pp. 894-899 (2006)

11 Seth, A., Kroeker, D., Zaharia, M., Guo, S.,Keshav, S.: Low-cost communication for rural internet kiosks using mechanical backhaul In: Proceedings of the

12th annual international conference on Mobile computing and networking, pp. 334-345. ACM (2006)
12 Demmer, M., Brewer, E., Fall, K., Sushant Jain, Melissa Ho,Patra, R. Implementing Delay Tolerant Networking. (Intel Research Berkeley Technical Report December 2004).
13 Christian, B.: Smooth is better than sharp: a random mobility model for simulation of wireless networks In: Proceedings of the 4th ACM international workshop on Modeling, analysis and simulation of wireless and mobile systems, pp. 19-27. ACM (2001)
14 Xiaoyan, H., Mario, G., Guangyu, P.,Ching-Chuan, C.: A group mobility model for ad hoc wireless networks In: Proceedings of the 2nd ACM international workshop on Modeling, analysis and simulation of wireless and mobile systems, pp. 53-60. ACM (1999)
15 Xiaoyan, H., Taek Jin, K., Mario, G., Daniel Lihui, G.,Guangyu, P.: A Mobility Framework for Ad Hoc Wireless Networks In: Proceedings of the Second International Conference on Mobile Data Management, pp. 185-196. Springer-Verlag (2001)
16 Bai, F., Sadagopan, N.,Helmy, A.: IMPORTANT: a framework to systematically analyze the Impact of Mobility on Performance of Routing Protocols for Adhoc Networks In: INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies. IEEE Vol. 2, pp. 825-835 (2003)
17 Jing, T., Joerg, H., Christian, B., Illya, S.,Kurt, R.: Graph-Based Mobility Model for Mobile Ad Hoc Network Simulation In: Proceedings of the 35th Annual Simulation Symposium, pp. 337-344. IEEE Computer Society (2002)
18 Aruna, B., Yun, Z., Croft, W. B., Brian Neil, L.,Aruna, V.: Web search from a bus In: Proceedings of the second ACM workshop on Challenged networks, pp. 59-66. ACM (2007)
19 NS2: The Network Simulator, <http://www.isi.edu/nsnam/ns/> (2008) [cited in 2008/12/07].

Real-Time Communication in IEEE 802.11 Wireless Mesh Networks: A Prospective Study

Carlos M. D. Viegas, Francisco Vasques

Faculty of Engineering – University of Porto (FEUP)
Rua Dr. Roberto Frias, 4200-465 – Porto – Portugal
{viegas,vasques}@fe.up.pt

Abstract. The purpose of this paper is to present an on-going prospective study about IEEE 802.11 Wireless Mesh Networks (WMNs), focusing on available techniques to support real-time communication. This paper addresses two possible approaches to meet real-time guarantees in WMNs that follow the IEEE 802.11s standard, respectively providing resource reservation for real-time message streams and improving handoff procedures to reduce the aggregated delays. The target of this paper is to sketch the first steps towards a state-of-the-art study for the author's Ph.D thesis.

Keywords: IEEE 802.11s, Wireless Mesh Networks, Real-Time Communication.

1 Introduction

Over the past few years, wireless networks have gained increased attention. Within this context, the IEEE 802.11 family of standards has become a dominant solution for Wireless Local Area Networks (WLANs). The main reason being its high performance, low cost and fast deployment characteristics [9, 21]. With the rapid growth of both Internet and wireless communications, there is an increasing demand for wireless broadband access and higher speed rates [11]. However, this demand creates new challenges due to the fact that usually increasing the data rate means that the communication range should be decreased in order to support a range of innovative services and access to mobile users [21].

The Wireless Mesh Network (WMN) concept appears as a promising solution for wireless environments, due to its characteristics and fields of application [8]. Basically, a WMN is formed by a set of wireless Mesh Points (MPs) that work together to convey communication between end users (a detailed description of WMNs architecture is presented in Section 2). WMNs are decentralized, easy to deploy and characterized by dynamic self-organization, self-configuration and self-healing [9]. They can be used in multiple application domains, like broadband home networks, community and neighboring networking, enterprise networking, metropolitan area networks, transportation systems, building automation, health and medical systems and security surveillance systems [3].

WMNs provide greater flexibility, reliability and performance when compared to traditional wireless networks [19]. Also, WMNs have the capacity to extend the network communication coverage without any additional infrastructure [6] by using multi-hop techniques, where nodes can relay traffic by traversing multiple hops to reach its final destination.

WMNs have particular characteristics that turn real-time communication into a challenging task. Thus, QoS provisioning techniques in WMNs need to be specifically suited for the purpose due to the lack of central infrastructure, the high level of heterogeneity, node mobility, fast topology change, medium access contention and also multi-hop characteristics, where packets may suffer from higher latency [8].

In this paper we consider a real-time communication environment where the topology is limited by boundary MPs. The incoming traffic should respect a set of QoS requirements defined by its real-time message streams properties, such as periodicity (P_i), execution time (C_i) and deadline (d_i). In the literature, there are multiple available techniques to support real-time communication in WMNs, by means of traffic management techniques, namely: admission control, resource reservation, policing, scheduling algorithms and others. To meet the requirements of real-time communication in WMNs, this paper envisages the use of mainly the following approaches: (1) The end-to-end delay guarantee will be provided by resources reservation and service differentiation techniques. (2) To guarantee a satisfactory level of communication continuity when a node is changing among MPs (handoff), it will be used techniques that improve the handoff process by reducing handoff delays, like fast handoff and cross-layer handoff.

The remainder of the paper is organized as follows. In the Section 2, it is given an overview of wireless mesh networks (WMNs) and the IEEE 802.11s standard, describing its main characteristics and architectures. The real-time communication techniques to guarantee QoS in WMNs are depicted in Section 3. In Section 4, the handoff process is discussed by presenting its operation and techniques to reduce handoff delays. Finally, in Section 5 some conclusions about real-time in WMNs context are presented.

2 IEEE 802.11 Wireless Mesh Networking

Wireless Mesh Networks are characterized by its capability of relaying frames from one device to another. In contrast to single-hop networks, where usually most of the traffic is directed to and received from a central infrastructure, mesh networks, potentially, have no hierarchy. The wireless medium is a shared resource that is used by all nodes in the mesh network [23].

Usually, WMNs consist of two types of nodes: mesh routers and mesh clients. The mesh routers are usually equipped with multiple wireless interfaces and are responsible for relaying traffic and interconnect the network with other networks, acting as gateways/repeaters. However, the mesh clients have only a single wireless interface and can also act as router relaying traffic, but without gateway function [3].

In traditional wireless single-hop networks, all nodes are either in mutual reception or have a common central neighbor: the Access Point (AP). Although in wireless mesh networks there may exist multiple direct and indirect neighbors, do not necessarily have an intersection of neighbors, characterizing a decentralized control [23].

The IEEE 802.11s draft standard [2] has been proposed to apply the WMN concept to the IEEE 802.11 networks, by introducing multi-hop forwarding at MAC level [6] and allowing wireless interconnection of access points (APs) and consequently the deployment of IEEE 802.11 wireless LANs with multiple APs.

2.1 The IEEE 802.11s draft standard

The IEEE 802.11s draft standard specifies a wireless mesh network technology based on the IEEE 802.11 WLAN standard. The objective is to extend the coverage of traditional WLANs and to allow the support of a larger diversity of wireless technologies. In a traditional IEEE 802.11-based WLAN, a *Extended Service Set (ESS)* is constituted by a set of *Basic Service Sets (BSSs)* interconnected via a wired IEEE 802.3 Ethernet. The way how the BSSs are interconnected leads to a poor scalability and increases the cost of the network. In a IEEE 802.11s ESS, the BSSs can be interconnected both via wired or wireless connections. Thus, the IEEE 802.11s ESS can support a larger number of nodes, allowing larger mesh networks to be created.

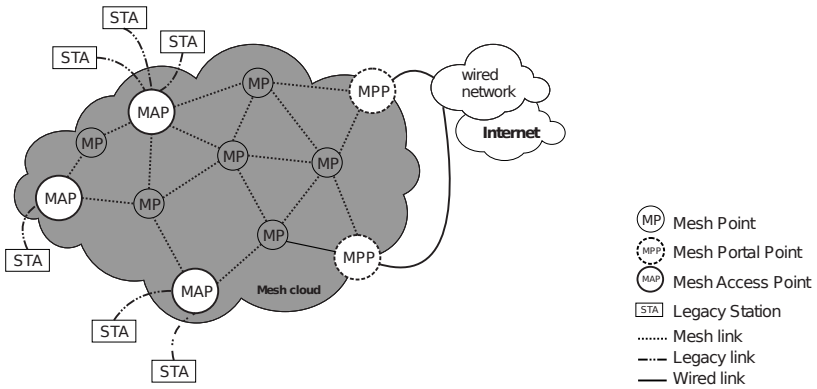


Fig. 1. Elements of a IEEE 802.11s Network.

In a IEEE 802.11s Mesh Network there are three types of nodes: *Mesh Points (MPs)*, *Mesh Access Points (MAPs)*, and *Mesh Portal Points (MPPs)*. A MP is a node that participates in the mesh routing process and that can forward frames. In addition, a MP device can have multiple radios, allowing the use of different radio channels or to access different radio technologies. Using multiple radio channels to communicate allows the increase of both throughput and redundancy of the mesh network. Some Mesh Points that have additional Access Point functionality are called Mesh Access Points. A MAP allows the support of other wireless *Legacy Stations (STA)*. Some other special MPs can act as a

Portal between the mesh network and other IEEE 802 networks. These nodes are called Mesh Portal Points and allow the extension of the mesh network coverage. Figure 1 illustrates the relationship between the different types of nodes in a mesh network. Note that nodes inside the mesh cloud, either MPs, MAPs or MPPs, are in essence MPs. Also, a MPP node can act as a Portal to a network and can act as a MAP to a such STA concurrently.

IEEE 802.11s Mesh Networks use a multi-hop wireless relaying infrastructure, where all nodes cooperatively maintain the network connectivity. Data frames can be routed from source nodes to destination nodes through a multi-hop communication network. Under the perspective of energy consumption, multi-hop communication represents an important approach to save energy on nodes with energy constraints, e.g. sensor nodes. However, it is also known that multi-hop communication has severe impact over the throughput capacity. The throughput degrades quickly, as long as the number of hops increase. Thus, to support real-time communication over a multi-hop network, it is necessary to create new medium access approaches.

3 Real-Time Communication in Wireless Mesh Networks

When considering the IEEE 802.11s draft standard, there are some relevant impairments to support real-time communication. Mainly, they are due to a medium access control technique designed for single-hop networks and that it is not well-suited for multi-hop networks. Also, the draft standard still doesn't specify any multi-channel mechanisms, and the scalability of routing algorithms is limited and difficult to provide adequate load balancing and QoS due to heterogeneous requirements [11, 21].

In order to provide real-time communication support, there is the need to ensure that the network is properly dimensioned and that enough resources are reserved in order to maintain the QoS parameters. For an absolute QoS guarantee, it may be required the reservation in advance of some resources, as the adequate reservations will help in maintaining delay, jitter and negotiated upper bound for packet loss rate requirements [10].

To support QoS requirements two methodologies will be envisaged, based on IntServ (*Integrated Services*) [4] and DiffServ (*Differentiated Services*) [17] techniques. The IntServ technique is aimed in providing per-flow QoS guarantees to individual applications, where several services classes are defined, and that applications should be able to choose a class based on their QoS requirements. It uses RSVP (*Resource reSerVation Protocol*) to allocate resources to the links along the data path from the sender to the destination. However, the IntServ scheme has scalability problems, where maintaining a large number of flows requires enormous resources [10].

The DiffServ technique consists on the specification of a restricted communication domain with specified requirements, delimited by boundary routers that control the ingress/egress network traffic. The ingress boundary router is required to classify the traffic according to a service level specification. The Diff-

Serv has a traffic conditioner where are included the traffic characteristics and the performance metrics (delay, throughput, etc.). In the interior nodes the traffic is processed at maximum available speed, as the traffic classification has been previously done by boundary routers [10, 22].

However, when considering the mobility of wireless mesh networks, neither the IntServ nor the DiffServ techniques work adequately with mobile nodes. This weakness is due to difficulties in reserving resources for mobile environments. As the IntServ works with RSVP by allocating resources to the links along the data paths, with the mobility of a wireless mesh node the path will change, and consequently there will be no reserved resources in a future router where the mobile node may connect. The main problem of DiffServ is the service level specification, i.e. when a mobile node moves to a new network and tries to establish new specification, resources must be available in the network to support the required QoS. If not enough resources are available, the mobile node will deal with degraded QoS [10].

It is clear that additional mechanisms must be proposed to achieve QoS in wireless mesh networks.

3.1 Resources Reservation

The first considered approach to support real-time communication in WMNs is the *Resource Reservation* technique. This technique consists in reserving resources in advance to guarantee that real-time requirements will be met. Usually, the reserved resources are bandwidth and time slots. This technique envisages to guarantee an end-to-end delay reduction and throughput increase.

A possible resource reservation technique in WMNs is the DARE protocol (*Distributed end-to-end Allocation of time slots for REal-time*) [5] that is proposed as a scheme to perform end-to-end reservations for real-time traffic flows. It operates at the MAC layer by reserving periodically time slots in all nodes along the route between the source node and its final destination. This protocol extends the concept of RTS (*Request to Send*) and CTS (*Clear to Send*) messages and proposes the RTR (*Request to Reserve*) and CTR (*Clear to Reserve*) messages to perform reservations. The RTR message, which includes requested duration and periodicity of a time slot as well as the address of the destination node, is transmitted along multiple hops since the source to the destination node, and if the destination node answers with CTR, the reservation is done and during the amount of reserved time the frames may be transmitted. In addition, during a reservation, the adjacent nodes of the real-time path are prevented to transmit, because they have been informed about such reservation. The DARE protocol is able to perform not only reservations but can recover them too in case of topology changes. According to simulation results, the DARE protocol offers a reliable and efficient support for QoS applications, by providing a constant throughput and a low and stable end-to-end delay for a reserved real-time flow.

3.2 Rate Adaptation

The second considered approach to support real-time communication in WMNs is the *Rate Adaptation* technique. It consists in a mechanism that uses the multi-rate capability of the network. The multi-rate capability can exploit the short inter-nodes distance in high-density networks owing the chance to use higher rates considering the rate-distance tradeoff. The effectiveness of a rate adaptation scheme depends on how fast it can respond to the variation of the wireless channel condition.

As a possible rate adaptation scheme in WMNs, the MTOP (*Multi-hop Transmission Opportunity*) [15] appears as a multi-rate adaptation mechanism that allows a frame to be forwarded a number of hops consecutively without contending for the medium. This mechanism is applied to multi-hop networks and takes advantage from different defer thresholds (multi-rate transmissions) to send frames to the next hops. Basically, the MTOP works after a TXOP (*Transmission Opportunity*) when it cannot allow frames to be transmitted in the given opportunity. By transmitting at different rates (1 Mbps and 11 Mbps, for example), it requires different defer thresholds (-105.1 dBm and -96.2 dBm, respectively). The difference between these defer thresholds is 8.9 dB and this is the multi-rate margin that MTOP exploits by allowing a frame to travel 1 or 2 more hops with a single medium access. This technique opens several interesting directions of research, as it can be employed in multi-radio/multi-channel networks.

3.3 Multi-Channel

The third considered approach to support real-time communication in WMNs is the *Multi-Channel* technique. The Multi-Channel consists in exploiting the multiple channels available in the wireless domain to transmit different frame types.

A possible multi-channel technique to use in the WMNs is the FFMAC (*Fast Forward Medium Access Control*) [25] protocol. It provides real-time guarantees through multiple communication channels, defining a multi-hop path between a source and a destination node (through IEEE 802.11s HWMP[1] routing protocol). Then, it reserves one channel to exchange control frames (control channel) and the remainder as channels to exchange data frames (data channels). The frame exchanging between a source and a destination node in a multi-hop environment is performed by a forwarding model, where the source broadcasts a RREQ frame (*Route Request*) on the control channel to its neighbors and they rebroadcast to neighbors until reaching the destination. Then, the destination answers with a RREP (*Route Reply*) frame that is broadcasted by neighbors until it reaches the source. This way, a transmission path is established for transmissions between the source and the destination nodes. After the path establishment, the source sends a data frame and waits for ACK from the neighbor node. However, the

[1] *Hybrid Wireless Mesh Protocol*.

neighbor node sends ACK to the source and also to the next neighbor to reserve the medium. If the next neighbor answers with a CTS frame, the data is finally forwarded and acknowledged again. And thus, it repeats through neighbors until the frame reaches the destination. The simulation results proved that this forwarding technique is able to reduce the end-to-end delay and to increase throughput.

4 Wireless Mesh Network Mobility Support

Real-time data services, like voice over IP (VoIP) and streaming multimedia, demand continuous network connectivity to guarantee that deadlines will be met. For an efficient delivery of real-time services to the mobile users, the wireless mesh networks require mechanisms for the mobility management, where efficient roaming techniques (handoff) are essential for ensuring connectivity and uninterrupted service delivery [18, 20].

4.1 Handoff Process

The handoff is a mobility process which allows a mobile station to move from one access point to another (Figure 2), i.e. the physical layer connectivity and state information with respect to a mobile client is transferred from one AP to another. This process occurs when a station is experiencing a degradation of the communication signal (*Received Signal Strength Indication - RSSI*) due to a physical movement and decides to perform a handoff to a candidate AP that offers better signal quality [23].

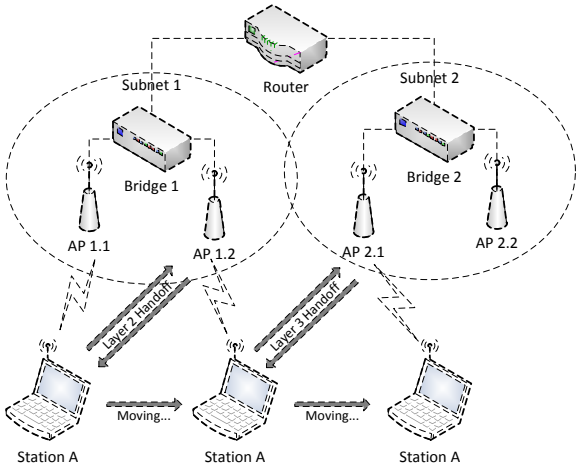


Fig. 2. A mobile station moving from an access point to another (handoff).

Considering the above network scenario (Figure 2), the handoff can be a layer 2 handoff if the station is moving among two APs in the same subnet or a layer 3 handoff if the station is moving among APs in different subnets [14]. The layer 2 handoff is transparent to the routing or appears simply as a reconfiguration without any mobility implications. On the other hand, in the layer 3 handoff the packets need to be switched to the new AP and new routing paths need to be established and QoS parameters renegotiated. The layer 2 handoff can be interpreted as a reconfiguration of the physical layer and the data link layer, while layer 3 handoff further affects reconfiguration of the network layer [23].

Regarding to the handoff characteristics, it can be *horizontal*, where is performed by using the same communication technology or *vertical*, where different technologies are involved [23]. In this paper we are concerned only with horizontal handoffs due to the use of the same technology (IEEE 802.11) in handoff procedures.

4.2 IEEE 802.11 Handoff Procedures

The IEEE 802.11 handoff can be classified in two types: *hard handoff* and *soft handoff*. A hard handoff occurs when a handoff process is triggered by a station after disconnection from an AP (break before make). Contrarily, a soft handoff occurs when the handoff process is triggered before disconnection from an AP (make before break). According to [13], the soft handoff is not supported by IEEE 802.11s due to resource limitations. Thus, the IEEE 802.11s standard only considers hard handoffs.

The handoff process in IEEE 802.11 networks is performed in three phases, briefly described in Figure 3 [14]:

1. **Scanning** phase, where a station searches for neighboring APs. The Scanning phase can be classified in *active* and *passive* scanning. In the active scanning mode the station broadcasts a probe request frame and then, after receiving the probe request, the APs respond with a probe response frame. In the passive scanning mode the station detects the neighboring APs by receiving beacon frames transmitted by these APs.
2. **Authentication** phase, for the station to be authenticated with the new AP. The Authentication phase has two authentication methods: *open system* and *shared key*. In the open system authentication, the station sends an authentication request frame and the requested AP responds with an authentication-response frame. In the shared key method, a four-way handshake is performed for the AP to check if the requesting station has the same security key.
3. **(Re)association** phase, where the station is finally connected to the new AP. The (Re)association is made by exchanging (re)association request frames from station to AP and (re)associated response frame from the AP to station.

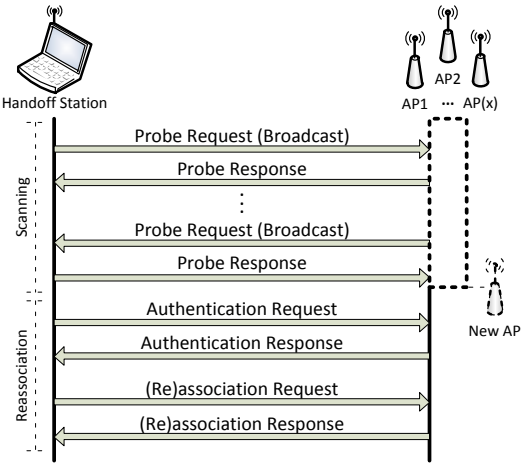


Fig. 3. IEEE 802.11 handoff process.

During the handoff process, there is a period where stations are unable to transmit due to the AP change. This interruption period mainly consists of two phases: the channel scanning delay, where a mobile station looks for APs in neighborhood, and reconnecting delay, where the mobile station proceeds to an association with a new AP [12]. These delays break the concept of connection continuity and should be avoided or reduced.

4.3 Reducing IEEE 802.11 Handoff Delays

A general handoff problem among WLAN environments is the lack of immediate upper layer awareness when the lower layer performs a handoff to a new access point in a different subnet [7]. There is a delay of several seconds for the upper layer to detect the node movement and to complete the duplicate address detection (DAD) and registration procedures.

The IEEE 802.11 standard considers both the layer 2 (link layer) and layer 3 (network layer) handoff delays. The link layer delay is divided into scanning, authentication and reassociation delay times. The scanning delay is the responsible for consuming most of the time in the overall handoff latency due to the probing scheme (and even worst due to the periodic beacons waiting scheme) [14]. The network layer delay includes the arrival, DAD, and binding update times. Some authors also consider the layer 5 (application layer) delay due to the reestablishment and modification of application layer properties such as IP address in a session [20].

Usually, the handoff execution applies a layer 2 handoff followed by a layer 3 handoff. Developing layer 3 handoff techniques, without considering the layer 2 handoff, will result in severe performance degradation and considerable handoff latency increase [16]. It is desirable to devise a mobility optimization technique

that can reduce these delays. In order to reduce the handoff delays, a well-defined coordination between layer 2 and 3 is required. A possible approach to reduce these delays is the cross-layer technique, where the link layer information is used to make an efficient network layer handoff. The utilization of link layer information reduces the delay in movement detection of a node, decreasing the overall handoff delay [20].

4.3.1 Cross-layer Handoff Techniques

As a cross-layer technique, is proposed in [13] the *Cross-Layer Transmission (CLT)* scheme to improve QoS in applications. This scheme consists of two phases: (1) *Off-line phase*, where a ratio of data frames relayed to a target MAP is computed based on Contention Window (MAC layer) and Received Signal Strength Indication - RSSI (PHY layer) values between an anchor node[2] and that target MAP when a station is moving through MAPs (handoff). This ratio is stored in each MAP candidates[3] (for run-time usage). (2) *Online phase*, where a feasible transmission ratio is derived based on the computed CW and RSSI values, when the handoff procedure starts. This transmission ratio will be used by an anchor node to control data relaying. By this way, to improve QoS in the station during the handoff procedure, the anchor node will start to transmit some data frames to the target MAPs (in advance) and at the same time it continually relays data frames to a source MAP. The objective of this scheme is to guarantee, during the handoff procedure, that real-time data frames will meet their deadlines. According to the simulation results, this CLT scheme avoids the real-time data frames missing their deadlines and also is able to improve the delay of non-real-time data frames.

In [24] is proposed an architectural design, named *Explicit multicast-based (Xcast-based) WMNs (XMesh)*, to facilitate inter-gateway handoff management in WMNs. The proposed architecture enables a parallel execution of multi-layers handoffs instead of its sequential execution (as usual in traditional approaches). Therefore, the handoff of layers 2, 3 and 5 is performed almost in parallel, thus reducing the aggregated delays in each layer. Also, a caching mechanism to guarantee a minimum packet loss is proposed, where it allows frames to be cached in a group of candidate MAPs in advance before an inter-gateway handoff. The integration of these two approaches reminds to a possible use *soft handoff*-like in IEEE 802.11s networks.

4.3.2 Fast Roaming: IEEE 802.11r

The IEEE 802.11r [1] is an emerging standard proposed to reduce the burden that authentication and QoS reservations generate to the handoff process. Briefly describing, this protocol defines a fast BSS transition (FT) that establishes the parameters necessary for data connectivity before the reassociation rather than

[2] An anchor node is a MAP which can relay data to both source and target MAPs.
[3] A candidate MAP consists in a MAP that possible can reassociate a mobile node from a handoff process.

after the reassociation when a handoff procedure occurs [14]. The FT is intended to reduce the disconnection time between a station and an AP during a BSS transition (handoff). Basically, it redefines the security key negotiation protocol, allowing both the authentication and reservation requests for wireless resources to occur in parallel.

As this is an emerging standard, it seems to be a promising technique to improve handoff delays and should be aim of further investigation.

5 Conclusions

This paper presented a study of proposed techniques to guarantee real-time communication in wireless mesh networks. This study envisaged the presentation of techniques to guarantee end-to-end delay by reserving resources and handoff management techniques to reduce the handoff delays. A combination of these techniques seems to be promising in order to deal with real-time communication in wireless mesh communication.

This study is a brief and initial state-of-the-art for my Ph.D research, where the future steps will be an implementation and simulation (through a network simulator software) of the presented techniques. The main idea is to combine resource reservation with improved handoff management techniques in order to analyze their viability to support real-time communication and a satisfactory mobility level within wireless mesh network environments.

References

1. IEEE 802.11r Standard for Information Technology - Local and Metropolitan Area Networks - Specific Requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications - Amendment 2: Fast Basic Service Set (BSS) Transition (2008)
2. IEEE 802.11s/D5.0 DRAFT Standard for Information Technology - Local and Metropolitan Area Networks - Specific Requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications - Amendment 10: Mesh Networking (2010)
3. Akyildiz, I.F., Wang, X., Wang, W.: Wireless Mesh Networks: a Survey. Comput. Netw. ISDN Syst. 47, 445–487 (mar 2005)
4. Braden, R., Clark, D., Shenker, S.: Integrated Services in the Internet Architecture: an Overview. RFC 1633, Internet Engineering Task Force (Jun 1994)
5. Carlson, E., Prehofer, C., Bettstetter, C., Karl, H., Wolisz, A.: A Distributed End-to-End Reservation Protocol for IEEE 802.11-Based Wireless Mesh Networks. IEEE Journal on Selected Areas in Communications 24(11), 2018–2027 (Nov 2006)
6. Carrano, R., Saade, D.C.M., Campista, M.E.M., et al.: Multihop MAC: IEEE 802.11s Wireless Mesh Networks, chap. 19. Encyclopedia on Ad Hoc and Ubiquitous Computing: Theory and Design of Wireless Ad hoc, Sensor, and Mesh Networks, World Scientific (2010)
7. Chen, Y.S., Hsiao, W.H., Chiu, K.L.: Cross-Layer Partner-Based Fast Handoff Mechanism for IEEE 802.11 Wireless Networks. In: IEEE 66th Vehicular Technology Conference, 2007. VTC-2007 Fall. pp. 1474–1478 (2007)

8. Farkas, K., Plattner, B.: Supporting Real-Time Applications in Mobile Mesh Networks. In: In Proceedings of the MeshNets 2005 Workshop. Budapest, Hungary (Jul 2005)
9. Hartmann, C., Meister, S.: A Quality of Service (QoS) Resource Management Architecture for Wireless Mesh Networks. In: 14th EUNICE open European Summer school 2008 (Sep 2008)
10. Jha, S., Hassan, M.: Engineering Internet QoS, chap. 2, 6, pp. 31–38. Artech House (2002)
11. Jiang, H., Zhuang, W., Shen, X., Abdrabou, A., Wang, P.: Differentiated Services for Wireless Mesh Backbone. IEEE Communications Magazine 44(7), 113–119 (Jul 2006)
12. Khan, R., Aissa, S., Despins, C.: MAC Layer Handoff Algorithm for IEEE 802.11 Wireless Networks. In: Computers and Communications, 2009. ISCC 2009. IEEE Symposium on. pp. 687–692 (2009)
13. Kuo, C.F., Tseng, H.W., Pang, A.C.: Cross-Layer Transmission Scheme with QoS Considerations for Wireless Mesh Networks. In: International Wireless Communications and Mobile Computing Conference, 2008. IWCMC '08. pp. 111–116 (Aug 2008)
14. Lee, B.G., Choi, S.: Broadband Wireless Access & Local Networks: Mobile Wimax and Wifi, chap. 16, pp. 509–533. Artech House Publishers (may 2008)
15. Lee, J.Y., Shen, T., Shin, K., Suh, Y.J., Yu, C.: Multihop Transmission Opportunity in Wireless Multihop Networks. In: 2010 Proceedings IEEE INFOCOM. pp. 1–9 (Mar 2010)
16. Mirchandani, V., Prodan, A.: Mobility Management in Wireless Mesh Networks. In: Misra, S., Misra, S.C., Woungang, I. (eds.) Guide to Wireless Mesh Networks, pp. 349–378. Computer Communications and Networks, Springer London (2009)
17. Nichols, K., Blake, S., Baker, F., Black, D.: Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers. RFC 2474, Internet Engineering Task Force (Dec 1998)
18. Pal, S., Kundu, S., Basu, K., Das, S.K.: Emancipating the IEEE 802.11 Network from Handoff Delay. CReWMaN (Jan 2006)
19. Pinheiro, M., Sampaio, S., Vasques, F., Souto, P.: A DHT-based approach for Path Selection and Message Forwarding in IEEE 802.11s industrial Wireless Mesh Networks. In: IEEE Conference on Emerging Technologies Factory Automation, 2009. ETFA 2009. pp. 1–10 (Sep 2009)
20. Sen, J.: Trends in Telecommunications Technologies, chap. Mobility and Handoff Management in Wireless Networks, pp. 457–484. InTech (Mar 2010)
21. Sgora, A., Vergados, D.D., Chatzimisios, P.: IEEE 802.11s Wireless Mesh Networks: Challenges and Perspectives. In: MOBILIGHT. pp. 263–271 (2009)
22. Sicker, D., McTasney, R., Grunwald, D.: Low Latency in Wireless Mesh Networks. In: Misra, S., Misra, S.C., Woungang, I. (eds.) Guide to Wireless Mesh Networks, pp. 379–424. Computer Communications and Networks, Springer London (2009)
23. Walke, B., Mangold, S., Berlemann, L.: IEEE 802 Wireless Systems: Protocols, Multi-Hop Mesh/Relaying, Performance and Spectrum Coexistence. John Wiley & Sons (nov 2006)
24. Zhao, W., Xie, J.: A Novel Xcast-based Caching Architecture for Inter-gateway Handoffs in Infrastructure Wireless Mesh Networks. In: 2010 Proceedings IEEE INFOCOM. pp. 1–9 (Mar 2010)
25. Zhu, G.M., Kuo, G.S.: A Fast Forward Medium Access Control Protocol for IEEE 802.11s Mesh Networks with Multiple Channels. In: IEEE Sarnoff Symposium, 2007. pp. 1–6 (2007)

Demystifying Cloud Computing

Paulo Neto

Faculdade de Engenharia da Universidade do Porto
Rua Dr. Roberto Frias, s/n 4200-465 PORTO, Portugal
pro10004@fe.up.pt

Abstract. The Cloud computing emerges as a new computing style which aims to provide on-demand network access to a shared pool of scalable and often virtualized resources (e.g., networks, servers, storage, applications, and services) that can be quickly provisioned and released. It became a mainstream in 2006 but so far there is still no consensus about its definition. This paper introduces and reviews the Cloud computing regarding to its definition, architecture, security and economical aspects. The purpose of this study is the creation of a baseline to start a PhD research within this subject.

Keywords: Cloud Computing, Grid Computing, Utility Computing

1 Introduction

With the fast improvement of computer processing and storage technologies in conjunction with the Internet success, the hardware resources have become cheaper and widely available. This technological trend has led to a new computing model called cloud computing, in which resources (e.g., CPU, storage and network) are provided as general utilities that can be leased and released by users through the Internet in an on-demand fashion.

Cloud computing can be seen as a platform that hosts applications and services and being driven by three significant trends. First, the wide shift to new Internet-based business models and Web 2.0 applications is driven by the growth in connected devices, real-time data streams, search operations, collaboration and social networking, and consumer-generated data. Second, global organizations are required to become integrated as they look to implement services-oriented architecture (SOA) applications and take advantage of software as a service (SaaS). Third, management, datacenter space and energy costs are driving to the requirement to improve the efficiency on the asset and human resources utilization.

Cloud computing provides computing as a utility. Just as electric companies provide electricity when and where needed, cloud computing vendors dynamically provision, configure, and de-provision IT (information technology) capability as needed, transparently and seamlessly. This allows IT consumers to focus on their specific problems and not on the computing resources they require.

There is an increasingly perceived vision that computing will be the 5th utility (after water, electricity, gas, and telephony) [3].

The remaining parts of this paper are organized as follows. Section 2 presents an overview and a definition of Cloud computing, Section 3 depicts a Cloud computing architecture, the underlying services and the deployment models and technologies. Section 4 compares Cloud with Grid computing emphasizing the role of the Web 2.0. The privacy and data security risks are discussed in Section 5. Section 6 describes some economical aspects and finally Section 7 presents some conclusions about this Cloud computing review.

2 Overview

In this section we present an overview of Cloud Computing including some definitions and a comparison with related concepts.

2.1 Definitions

The underlying concept of cloud computing is not a new one. John McCarthy in 1961, was the first to publicly suggest (in a speech given to celebrate MITs centennial) that computer time-sharing technology might lead to a future in which computing power and even specific applications could be sold through the utility business model (like water or electricity). This idea of a computer or information utility was very popular in the late 1960s, but given up by the mid-1970s as it became clear that the hardware, software and telecommunications technologies of the time were not ready yet for that challenge. However, since 2000, the idea has resurfaced in new forms. The first academic definition was provided by [6] in 1997 who called it a computing paradigm where the boundaries of computing will be determined by economic rationale rather than technical limits. When Eric Schmidt [16] explained his cloud computing view on the Search Engine Strategies Conference in 2006 and a couple of weeks later Amazon included the word cloud in the Elastic Cloud Computing (EC2), the term became a mainstream.

In October 2007 Google and IBM [7,12] announced a major research initiative to help students and researchers to address this new Internet-scale computing paradigm. There is still a little consensus how to define Cloud Computing [10]. The Berkeley researchers [14] define Cloud Computing as both the applications delivered as services over the Internet and the hardware and systems software in the datacenters that provide those services.

The NIST also published a definition of cloud computing [13]:

> "Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."

From the hardware point of view, there is an illusion of infinite computing resources available on demand. The Cloud users don't need to concern themselves about the required resources for the future growth, and those resources will be paid in a pay-per-usage basis.

The Figure 1 shows the main characteristics of a cloud service. It needs to be paid in a pay-per-usage basis, the resources should be scaled up or down on demand (elasticity), and the resource management is owned by the provider.

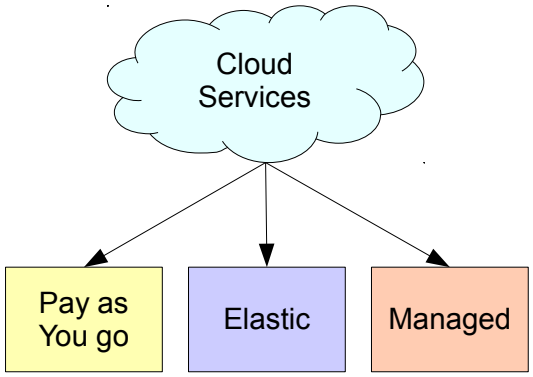


Fig. 1. Characteristics of a cloud service.

3 Cloud Computing Architecture

This section presents the cloud computing architecture, the typical deployment models as well as the underlying technologies.

3.1 Cloud Service Levels

Cloud services can be classified into four general types [4]:

Software as a Service (SaaS)
This is the most common type of cloud service and one that almost everyone has already used at some point. In the SaaS cloud model, the service provider supplies all the infrastructure along with the software product. Users interact with the service using a Web-based front end. These services cover a wide range, from Web-based e-mail like GMail to financial software like Mint.

Platform as a Service (PaaS)
Cloud service that provides software and product development tools hosted by the provider on the hardware infrastructure. The term PaaS is commonly used for cloud-based platforms to build and run custom applications. PaaS applications provide everything needed to build and deploy Web applications

and services accessible from anywhere on the Internet. The end users do not have to download, install, or maintain the system. Popular examples of this kind of a service are Google App Engine, Microsoft Windows Azure, Force.com, Morph and Bungee Connect [5].

Infrastructure as a Service (IaaS)
Cloud services that provide access to computing resources such as processing or storage which can be obtained as a service. The most popular examples of IaaS are Amazon Web Services (AWS) with its Elastic Compute Cloud (EC2) [1] for processing and Simple Storage Service (S3) [2] for storage , and and Rackspace.

data Storage as a Service (dSaaS)
Services that provide storage to be used by the consumer including bandwidth requirements [9].

Cloud computing can be viewed as a collection of services, which can be presented in layers services, as shown in Figure 2.

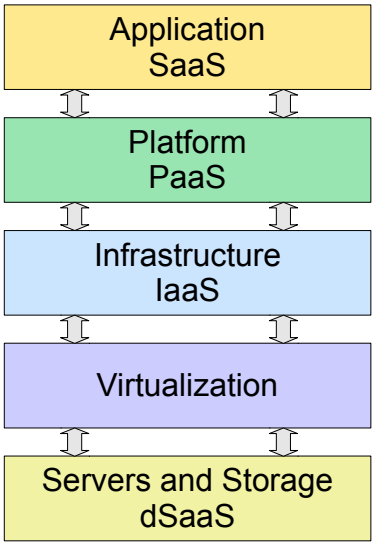


Fig. 2. Layered architecture of Cloud Computing.

3.2 Cloud Computing Deployment Models

There are four types of cloud computing deployment models:

Public Cloud
A public cloud is the traditional concept where the resources are leased

through the Internet from an off-site third-party provider who bills in a pay per usage basis.

Community cloud
A community cloud is usually used by a set of organizations with similar requirements and interests in order to share the infrastructure but keeping a certain additional level of security and privacy. Examples of community cloud include Google's "Gov Cloud".

Hybrid cloud
Even without consensus with this term, a hybrid cloud is probably the use of physical hardware and virtualized cloud server instances together to provide a single common service. For instance, a hybrid storage cloud can be used in a tier fashion for archiving, backup and replication functionalities.

Private cloud
In private clouds the idea of leasing instead of buying and manage is lost, but for big organizations it can be acceptable with the IT department (or the hosting entity) playing the provider role. The storage and server hardware providers also allow some pay per usage policy, like having additional processors and storage capacity that will be charged just when temporarily activated (Capacity on Demand). Thus even in private clouds the backend overprovisioning can be avoided.

3.3 Technologies Behind a Cloud

Numerous underlying precursor technologies enabled cloud computing to emerge, like:

- Internet
- Virtualization
- Software-as-a-Service
- LAMP and WAMP stacks
- Web Hosting
- Database
- Inexpensive CPUs and storage
- SOA (service-oriented architectures)
- Sophisticated client algorithms, including HTML, CSS, AJAX, REST
- Client broadband
- SOA (service-oriented architectures)
- Large infrastructure implementations from Google, Yahoo, Amazon, and others

4 Comparing with Grid Computing

The definition of Cloud Computing overlaps with many existing technologies, such as Grid Computing, Utility Computing, Services Computing, and distributed computing in general. Cloud Computing is indeed evolved out of Grid

Computing and relies on Grid Computing as its backend and infrastructure support. The evolution has been a result of a shift in focus from an infrastructure that delivers storage and processing resources (such is the case in Grids) to one that is economy based in order to deliver more abstract resources and services (such is the case in Clouds) [8]. Cloud and grid systems share the same basic goal to reduce the computing costs. Grids are mostly designed to be general purpose, so they exhibit a complete set of available system capabilities and the resulting interface available for users and applications remains low level [11]. In contrast to the Grid system interfaces, cloud system interfaces are simpler and they do not expose internal system characteristics. Typically the exposed capability set is usually much more limited than the set of capabilities available in the Cloud system itself. Figure 3 adapted from [8] shows an overview of the relationship between Clouds and other domains that it overlaps with. Web 2.0 covers almost the whole spectrum of service-oriented applications, where Cloud Computing lies at the large-scale side. Supercomputing and Cluster Computing have been more focused on traditional non-service applications. Grid Computing overlaps with all these fields.

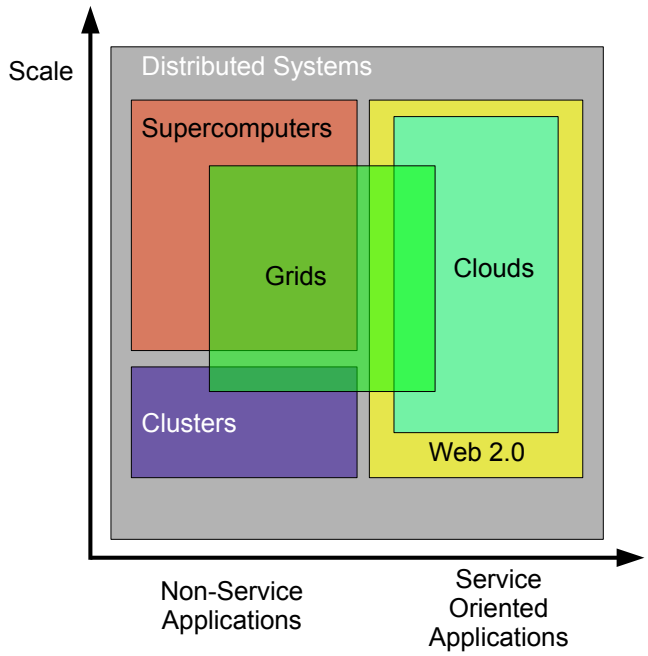


Fig. 3. Grids and Clouds interaction.

5 Privacy and Data Security Risks

Although the many advantages of cloud computing already presented, some security issues need to be carefully evaluated. Processing sensitive data outside the companies require additional cautions since the data bypass the physical, logical and personnel controls. The customers need to follow data retention policies enforced by the regulatory laws, thus the service provider need to be prepared and certified for those requirements. As long as the data is spread across multiple locations a special concern need to be taken regarding to the law on those specific locations (state/countries). In order to keep the data confidentiality, a widely tested encryption scheme should used because an encryption accident can make the data totally unusable. To protect against theft and or denial-of-service attacks by users, virtualization is the primary security mechanism in the today's clouds. It's a well known technology that protects against most attempts by users to attack one another or the underlying cloud infrastructure. Nevertheless neither all resources are virtualized nor the virtualization engines are bug free. Another important concern is the protection against the cloud provider. The virtualization technologies may allow the one who manages the lower layers to circumvent many security barriers. Although there is already strict legislation in US and EU to regulate and audit cloud computing providers [15], there is still some reluctance by several companies to adopt this new technology at least in a public cloud.

6 Cloud Computing Economics

One of the basic policies for any organization is the TCO (Total Cost of Ownership) reduction. For instance, provisioning a data center for the peak load it must sustain during the end of the month (batch jobs), leads to a resource underutilization during remaining time of the month. An organization in this situation may benefit from cloud computing paying by the hour for computing leading to cost savings even if the hourly rate to rent a machine from a cloud provider is higher than the rate to own one (including space, power, cooling and maintenance costs).

One of the basic policies for any organization is the reduction the TCO (Total Cost of Ownership). For instance, provisioning a data center for the peak load it must sustain during the end of the month leads to underutilization during the other days. In this situation an organization may benefit from cloud computing paying by the hour for computing leading to cost savings even if the hourly rate to rent a machine from a cloud provider is higher than the rate to own one (including space, power, cooling and maintenance costs).

Another example is a business startup that requires some computing resources at the beginning but will suffer an unexpected increase of the resources demand when it becomes popular. This type of organization doesn't need to invest upfront in computer resources that potentially will be used only months or years later.

Although the clouding computing costs are more expensive compared with owning a server for the same period, the cost of over-provisioning (having the server in almost idle status during many hours per day) or under-provisioning (increasing the system response time) is very high, driving the cloud computing solution very attractive from the business point of view.

In Figure 4 with a correctly antecipated peak load, there is a waste of resources due the lack of elasticity. The Figure 5 represents a typical underprovisioned environment where the demand overpass the capacity during the peak load and still presenting long periods of low resource usage. Figure 6 represents an on demand provisioning where the resources are dynamically adjusted based on the demand (elasticity) providing a great cost effective solution.

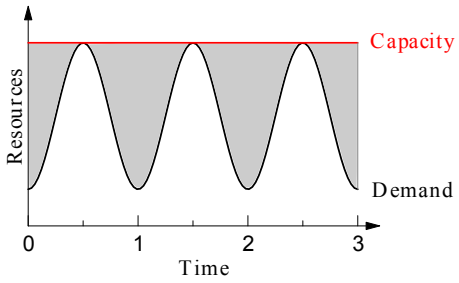


Fig. 4. Provisioning for peak load.

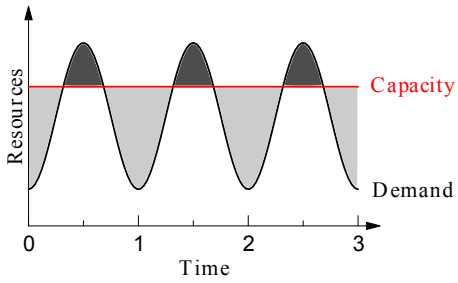


Fig. 5. Underprovisioning.

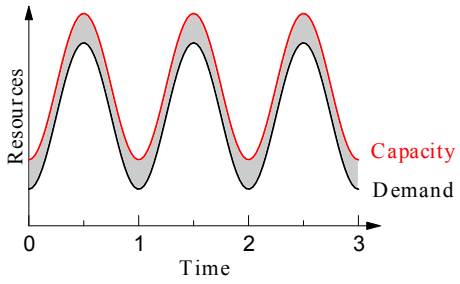


Fig. 6. On-Demand Provisioning.

In Figures 4,5 and 6, the grey area represents the waste of resources making the on-demand provision the most efficient.

Although the economical benefits explained before there are also some negative aspects using Cloud Computing. For instance transferring a large database over the network involves a significant cost (high bandwidth) and time. Some companies choose to ship the data in physical support (tape or disk) through a

courier in order to reduce the network costs and the time that takes the data import activity.

7 Conclusion

This paper reviewed the Cloud Computing technology from the architecture and deployment models point of view. The main goal of this new trend, i.e., reducing the IT costs has been achieved. However from the economical point of view there are still many constraints like transferring large chunks of data and assuring the correct data retention policies. On the other hand, there are many security concerns that need to be investigated in depth. Those security issues are spread across several technologies, like networking, software services, virtualization and storage, thus a future study will need to be narrowed within the main Cloud Computing domain. As a PhD student, my future study will be focused on the storage domain.

References

1. Amazon Elastic Compute Cloud (Amazon EC2). `http://aws.amazon.com/ec2/`, [Accessed December 2010]
2. Amazon Simple Storage Service (Amazon S3). `http://aws.amazon.com/s3/`, [Accessed December 2010]
3. Buyya, R., Yeo, C., Venugopal, S., Broberg, J., Brandic, I.: Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. Future Generation Computer Systems 25(6), 599–616 (2009)
4. Chaganti, P.: Cloud services for your virtual infrastructure, Part 1: Infrastructure-as-a-Service (IaaS) and Eucalyptus. IBM developerWorks (December 2009), [Accessed December 2010]
5. Chaganti, P.: Cloud services for your virtual infrastructure, Part 2: Platform as a Service (PaaS) and AppScale. IBM developerWorks (January 2010), [Accessed December 2010]
6. Chellappa, R.: Cloud computing: emerging paradigm for computing. INFORMS 1997 (1997)
7. Chiu, W.: Google and IBM Initiative for Computer Students. `http://www.ibm.com/ibm/ideasfromibm/us/google/images/podcast_google.pdf` (October 2007), [Accessed December 2010]
8. Foster, I., Zhao, Y., Raicu, I., Lu, S.: Cloud computing and grid computing 360-degree compared. In: Grid Computing Environments Workshop, 2008. GCE'08. pp. 1–10. Ieee (2009)
9. Furht, B., Escalante, A.: Handbook of Cloud Computing. Springer (2010)
10. Geelan, J.: Twenty-One Experts Define Cloud Computing. `http://cloudcomputing.sys-con.com/node/612375` (January 2009), [Accessed December 2010]
11. Jha, S., Merzky, A., Fox, G.: Using clouds to provide grids with higher levels of abstraction and explicit support for usage modes. Concurrency and Computation: Practice and Experience 21(8), 1087–1108 (2009)

12. Lohr, S.: Google and I.B.M. Join in Cloud Computing Research. The New York Times (October 2007), [Accessed December 2010]
13. Mell, P., Grance, T.: The NIST Definition of Cloud Computing. National Institute of Standards and Technology (2009)
14. Michael, A., Armando, F., Rean, G., Anthony, D., Randy, K., Andy, K., Gunho, L., David, P., Ariel, R., Ion, S., et al.: Above the clouds: A berkeley view of cloud computing. EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2009-28 (2009)
15. Sotto, L., Treacy, B., McLellan, M.: Privacy and Data Security Risks in Cloud Computing. Electronic Commerce & Law Report (2010)
16. Sullivan, D: Conversation with Eric Schmidt hosted by Danny Sullivan. (http://www.google.com/press/podium/ses2006.html) (August 2006), [Accessed December 2010]

Survey on Privacy Solutions at the Network Layer: Terminology, Fundamentals and Classification

Pedro Moreira da Silva[1], Jaime Dias[1], Manuel Ricardo[1],

[1] INESC Porto & Faculdade de Engenharia, Universidade do Porto,
Rua Dr. Roberto Frias, s/n 4200-465 Porto, Portugal
pmms@inescporto.pt

Abstract. Communication networks are by design targeted for performance and accountability not for preserving user's privacy. Secure communications using IPSec or SSL/TLS are designed just for providing authentication, integrity and confidentiality; the information leaked, such as IP addresses, is still a major privacy threat. This paper presents an overview of research done in this area, more precisely at the network layer level, and provides a description of the terminology used. Also, a possible taxonomy is suggested according to the message type and network architecture, both for message exchange and peer discovery. Although research in this field is going on for almost three decades, a practical, low-latency, robust and efficient solution is yet to be discovered.

Keywords: Privacy, Anonymity, Unlinkability, Unobservability, Network Coding, VPN.

1 Introduction

Communication networks are by design targeted for performance and accountability. The network managers analyse the traffic that flows on their network(s) for several reasons such as, billing, failure detection, ensuring certain quality of service levels and, even, traffic shaping. Also, at the network level, little is done to assure that the user's privacy is not compromised since packets are, in most cases, sent in clear. For several years, cryptographers' research was oriented to authentication, integrity and confidentiality, which provide the ground base for creating what is commonly referred as secure communications. Protocols such as IPSec [1] and SSL/TLS [2] provide secure communications through packet encryption, respectively at network and application layers. However, given that encryption is able to protect only the message that is being communicated, these protocols still leak valuable information such as the amount of traffic exchanged, the communicating end points identities (IP addresses) and communication flows. With this information it is possible, e.g., to successfully fingerprint requested web pages [3], [4] and discover the base station node(s) in wireless sensor networks [5]. The lack of privacy in the Internet is a well-documented fact in the literature [6] and has been a research topic for almost three decades.

Privacy solutions are designed, in general, for protecting a specific networking layer given that each layer presents its own threats to privacy preservation and is unable to conceal information leaked by other layers; e.g., even if the transport layer provides anonymity, online social networks may generate HTTP URIs that discloses user's information since they are user-specific [7]. For that reason, in order to provide a more detailed analysis, this paper only deals with solutions that aim to protect the network layer – the lowest layer used in the Internet.

1.1 Terminology

Attempting to standardize the key concepts in this area, Pfitzmann & Köhntopp published [8], and are still updating the document addressing feedback from the research community[1]. They provide the definitions for the following concepts:

Anonymity. To enable anonymity of a subject, there always has to be an appropriate set of subjects with potentially the same attributes. Thus, anonymity *"is the state of being not identifiable within a set of subjects, the anonymity set."*

The *anonymity set* is the set of possible subjects (acting entities) who might cause an action. For that reason, a sender can only be anonymous within a set of potential senders (*senders anonymity set*), which may be also a subset of all subjects. Regarding recipients, the anonymity set (*recipients anonymity set*) is the set of all potential addressees. These two sets may vary over time and can be disjoint, be the same or overlap. Hence, anonymity is the probability of a subject being identified by attackers within an anonymity set.

Unlinkability. As suggested in [9], unlinkability can be defined as absolute, if "*no determination of a link between uses*" is possible, or as relative, if there is "*no change of knowledge about a link between uses*". Usually, papers presenting anonymous systems, unless otherwise stated, by unlinkability mean relative unlinkability since what is being considered is the probability of an attacker being able to gain additional knowledge between uses and not the impossibility of obtaining such information (sometimes referred as provable unlinkability).

Unobservability. In contrast with anonymity and unlinkability, where not the item of interest (IOI) but its relationship between IDs and IOIs is protected, unobservability regards the protection of IOIs themselves. "*Unobservability is the state of items of interest (IOIs) being indistinguishable from any IOI (of the same type) at all.*" Unobservality can be considered a superset of confidentiality given that it intends to prevent any IOI identification at all and not only to thwart attackers from recovering the IOI using cryptographic attacks.

Although not considered in Pfitzmann & Köhntopp work, here we introduce *communication unobservability* to refer to the impossibility of knowing if a node is communicating or not in a certain period of time.

[1] http://dud.inf.tu-dresden.de/Anon_Terminology.shtml.

Pseudonimity. *Pseudonyms* are identifiers of subjects that reveal by themselves no additional information about the subject. The pseudonyms could be used either to identify a single subject or to identify a set of subjects; however, in the literature, generally, they are used only for single subjects. As so, "*Pseudonimity is the use pseudonyms as IDs.*"

Throughout this paper, the term privacy, unless otherwise stated, refers to anonymity, relative unlinkability, and communication unobservability. Pseudonimity is considered more a mean to achieve anonymity, unlikability and unobservability than a goal by itself.

The taxonomy used for classifying each solution presented in this paper is described in Section 2. Section 3 presents the solutions that provide privacy in a single entity controlled communication model; Section 4 describes the multi-entity controlled ones. The conclusions are presented on Section 5.

2 Taxonomy

For providing anonymity it is required that the sender does not, necessarily, communicate directly with the recipient, i.e., that one or more intermediate nodes are made part of the communication path. Depending on the traffic destination – within the solution's network or not – the last intermediate node may need to act as a translator between solution's protocol and the Internet protocol. Solutions are classified according to the kind of messages exchanged between nodes and the constructed network architecture. The network architecture is analysed from the point of view of how messages are exchanged and also, if present, from the peer discovery point of view – how peers are indexed and searched.

2.1 Messages

Solutions conceived for providing privacy either have a network architecture that is aware of the underlying network topology or create a virtual network that is used to abstract from it. Solutions that usually consider the underlying network topology are those that take advantage, when available, of some network functionalities such as multicast and broadcast. For classifying a solution according to the message type three categories are defined:

Unimessage. Messages that are exchanged between the network nodes in unicast and are never split, neither by the sender nor by nodes across the communication path, are classified as unimessage. For this kind of messages three fundamental schemes can be considered: fundamental path based, probability based and mimic traffic based.

Fundamental path based schemes are those in which messages travel along a predefined path. Usually, the path is constructed a priori, in an attempt to reduce the delay overhead, and the links are protected.

Fig. 1 illustrates one example, in which a message is sent encrypted by node A to node B; Node C is made part of the path to introduce some degree of privacy. Since a communication path can have several intermediate nodes, the previous node may not be the sender, just an intermediate node; the same applies for the next node, which needs not to be the destination node. Therefore, this scheme can provide sender and recipient anonymity.

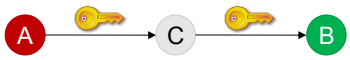


Fig. 1. Fundamental Path based schemes.

Probability based schemes, as the name suggests, are schemes in which the path is constructed according to a probability value. Any node in the path, including the sender, will generate a random value and, based on the random and probability values, will decide if the message is to be sent to the recipient or to another random node (see Fig. 2). This process is repeated until the message is sent to the intended recipient. Thus, since the recipient is always known, this scheme can provide only sender anonymity.

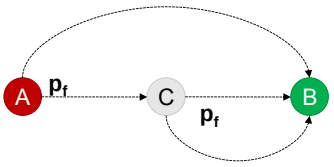


Fig. 2. Probability based schemes. p_f is the probability to forward to the recipient.

Mimic traffic based schemes, contrarily to the previous two schemes, are meant to hide the amount of traffic sent and when it was sent (communication unobservability), and not necessarily anonymity. Each node has its neighbours and, synchronized and periodically, all exchange messages; the message may be useful or dummy traffic. Since all nodes always exchange messages with their neighbours it is not possible to determine if a node is communicating or not. These schemes present a trade-off between delay and traffic overheads: the delay overhead is directly proportional to the interval between communications; the traffic overhead is inversely proportional to the interval between communications.

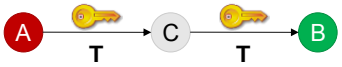


Fig. 3. Mimic traffic based schemes. T is the interval (*period*) between communications.

Split Message. Messages that are exchanged between the network nodes in unicast, split in two or more pieces, and sent through different but not necessarily disjoint communication paths, are classified as split message. Note that a message is considered split if the required information for retrieving its content comes from at

least two different messages, e.g., a message is encrypted but the decryption key is only sent in another message. For this kind of messages three fundamental schemes can be considered: information dispersal algorithm (IDA) based, rumour riding based and linear coding based.

IDA based schemes are those which the sender breaks the original message in several parts and sends them through different paths. Redundant parts may be generated to account for any packet drop or delays above a certain threshold. Usually, the content is encrypted before being broken and each part is sent in clear. If a global attacker is to be considered, each part can be encrypted as well.

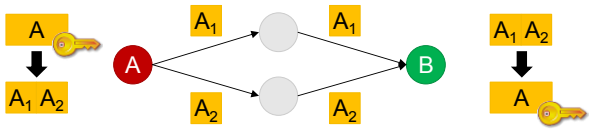


Fig. 4. Information dispersal algorithm based schemes.

Rumour Riding based schemes are used mostly for private queries. The query is sent encrypted in one message and the decryption key on another message. Nodes will forward encrypted queries and key rumours to their neighbours but keeping track of those that were received to prevent cyclic forwarding. At some point, a node will receive both the key rumour and the encrypted query, and will make that query on behalf of the sender (see Fig. 5). The answer is sent back reversing the path from where the rumour key and the encrypted query were received. Given that any node can be the decrypting one, only sender anonymity is provided. This scheme is impractical for large networks.

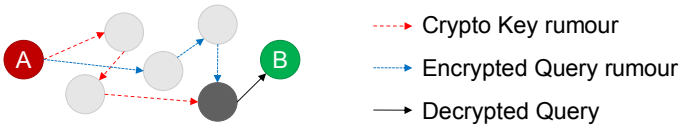


Fig. 5. Rumour ridding based schemes.

Linear coding based schemes are those which linearly combine messages: message is treated as an unknown variable and linear equations are created using some factors for each unknown; nodes in between may also produce new linear combinations. Once the recipient receives enough linearly independent messages it will solve the system of equations and recover the original message. Note that intermediate nodes may also solve systems of linear equations created by other intermediate nodes. Fig. 6 illustrates the case in which only the sender creates linearly combined messages.

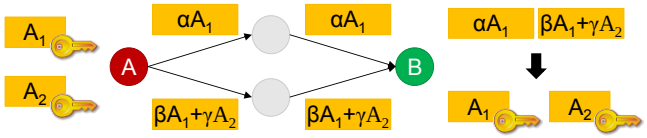


Fig. 6. Linear coding based schemes.

As in IDA based schemes, redundant messages may be sent. These are generally used for improving performance but also to difficult traffic analysis since it is harder to fingerprint the content and identify message flows.

Message Replication. The schemes based on message replication use multicast or broadcast for sending messages and achieving privacy. These are the schemes that are aware of the underlying network because multicast and broadcast are not available all over the Internet.

Multicast enables the recipient anonymity since any node that belongs to the multicast group is a potential recipient. However, the sender identity is made known. Fig. 7 illustrates the use of multicast in which the recipient is hidden in an anonymity set of cardinality 4.

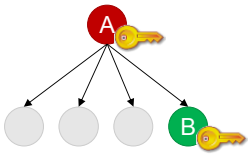


Fig. 7. Message replication based schemes (multicast).

Broadcast can be used to provide both sender and recipient anonymity because the source address and the destination address are the broadcast address. However, broadcast is only available on LANs. Fig. 8 illustrates the broadcast use for sender and recipient anonymity within an anonymity set of cardinality 5.

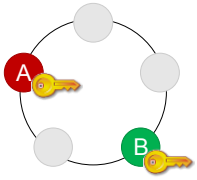


Fig. 8. Message replication based schemes (broadcast).

2.2 Network Architecture

Mainly peer-to-peer (P2P) based solutions need to provide peer discovery mechanisms because the network architecture and the relevant nodes is not known a priori. Either for message exchange or for peer discovery, the network architecture can be classified as belonging to one of the following three categories: centralized, decentralized and hybrid.

Centralized. A centralized solution holds one or more nodes that have a major role on the overall process, be it for routing or peer discovery. This architecture is usually more simple and efficient since the information is located and contained on a small set of well-known nodes albeit the servers are single points of failure.

Decentralized. In a decentralized solution all nodes are at same level. This architecture is usually more complex and less efficient since the information is spread across the network but the network is more resilient.

Hybrid. A hybrid solution is in between the prior two solutions and tries to minimize the downsides without losing too much simplicity, efficiency and resiliency.

3 Single entity controlled solutions

Commercial products that provide privacy services are usually single entity solutions. This kind of solutions is provided through the use of servers owned by those entities such as proxy and virtual private network (VPN) servers. The scheme is simple and easy to implement since clients just connect to a VPN server or a proxy server to do their regular activities. Given that requests are made by the server, the identity of the client is concealed. These are centralized solutions that aim to provide anonymity and, at some extent, unlinkability.

3.1 Proxy-based

Proxy-based solutions are normally used for privately accessing HTTP and FTP sites. The user configures his browser to use a specific proxy and all traffic is sent through that proxy. Proxify [10] and Socksify [11] are two examples of such solutions. Proxify is a web-based proxy that runs over HTTPS to encrypt the data exchanged between the website that acts as a proxy and the user; the request is then done by the website to the intended website/ftp site. Therefore, there is no configuration. Socksify is an HTTP proxy server that is SOCKSv5 compliant. As so, the browser (or a SOCKSv5 compliant application) needs to be configured to use this proxy.

3.2 VPN-based

Solutions based on VPN servers create network tunnels that are used to encapsulate packets; these solutions usually use IPSec or PPTP tunnelling protocols. The sender creates a packet using its local address that is encapsulated in another packet that will be sent to the VPN server over a public network. When the packet arrives the VPN server, it will be decapsulated and the inner packet will be treated as if it was originated from the local network (see Fig. 9).

Fig. 9. Encapsulated packet. Payload of the outer packet is sent encrypted.

Thus, if the inner packet has a destination on the Internet, it will be sent using VPN server's public IP address(es) concealing the user's identity. Contrarily to proxy-

based solutions, as long as the local network supports it, virtually any protocol can be used in the inner packet. Relakks [12] is an example of such commercial product.

Both proxy-based and VPN-based solutions are very easy to use and configure, relaying the packets through well-known servers that are assumed to be trustful. All user's traffic goes through those servers which can analyse it and, if the packets exchanged between the user and the end-server are not encrypted, it may capture user's credentials and other relevant information. These solutions work as an Internet Service Provider (ISP), the only noteworthy differences, from the privacy preservation point of view, are that one IP address may be shared by several users and that the legal obligations may not be the same.

4 Multi-entity controlled solutions

Contrarily to single entity controlled solutions, multi-entity ones are much more resilient to a global attacker given that, in general, no single entity can compromise the users' privacy. The work on privacy solutions was pioneered by Chaum which introduced two important privacy schemes: mix-nets [13] and DC-nets [14].

4.1 Mix-nets

Mix-net based solutions rely on mixers which send the packets in a different order from the one they were received. Chaum developed this scheme for providing anonymity in email which is a service that does not require low-latency networks. Also, in the original design, a mixer retained a packet until enough packets were received to conceal it. Being a fundamental path based solution, mix-nets provide sender and recipient anonymity. The anonymous remailers as Mixminion [15], since are only intended for email, are not analysed.

Onion Routing. Onion routing protocols, being Tor [16] the most used and widely known, use one encryption layer for each intermediate node (see Fig. 10).

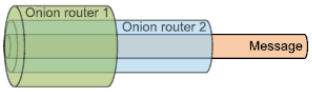


Fig. 10. Onion routing encryption layers (obtained from [17]).

Tor is a hybrid solution that serves end-nodes through special nodes: core routers; the core routers are the nodes that will mix the users' traffic and send it to the intended recipients. Tor is a low-latency solution that only supports TCP; paths are maintained keeping alive TCP connections between each adjacent node in the path. DNS queries, since they use UDP, are not directly supported by Tor.

Web Mixes. Web Mixes were proposed in [18] and consists of three logical parts: JAP (Java Anon Proxy) on the client-side, and, on the server-side, MIXes and cache-proxy. The JAPs are connected to the first MIX via Internet using a tunnel and MIXes are chain-interconnected via Internet tunnels as well; the last MIX is connected to the cache-proxy. JAPs, MIXes and cache-proxy regularly send dummy traffic to provide unlinkability and communication unobservability. This solution has hybrid architecture both regarding message exchange and MIXes indexing.

4.2 DC-nets

DC-nets, contrarily to mix-nets, have not seen such broad exploration. The main idea is to share coins flips between adjacent peers, organized in circle, and each peer broadcasts the flips seen; since each flip will be reported twice, the result must be even. If a node wants to transmit he flips the result. This scheme requires the existence of a reliable broadcast and provides unconditional sender and recipient anonymity; however, reliable broadcast is not available in every network, definitely not over the Internet. Given that at each round only one node can transmit, otherwise a collision will occur, and that the computation complexity of detecting attackers that intentionally create collisions to degrade the network performance, although solutions as [19] were developed, at the moment, the advantages/disadvantages ratio is too low.

4.3 Network routing based

These solutions provide privacy through the use of routing schemes and not through other techniques. Crowds [20] and buses for anonymous delivery [21] are two of such schemes. Only Crowds is described since the other has just theoretical interest.

Crowds. Crowds was developed for private Web browsing and it is a mix of fundamental path and probability based schemes. Each node constructs a path that behaves as a leaked pipe and the leaks – sending directly to the intended recipient – occur with a given probability. The paths are periodically reformed but Crowds does not provide anonymity against a global attacker or a local eavesdropper because it assumes that members of the crowd have no knowledge about the previous nodes.

4.4 P2P

The solutions presented in this section also perform mix-net like operations; however, contrarily to the ones presented in subsection 4.1, they are P2P solutions.

Tarzan. Tarzan [22] is the only solution presented in this paper that is based on mimic traffic. Tarzan constructs a P2P virtual network in which each node as a set of neighbours, called *mimics*, to whom it maintains persistent connections. The periodically sent messages are onion encrypted, to ensure that its content is known only by the intended recipient, and are only sent through mimics to avoid successful

traffic analysis with insufficient traffic. Communication with nodes outside Tarzan's network is done using a PNAT (pseudonymous network address translator).

Tarzan is a decentralized solution both for peer discovery and message exchange: peer discovery is done using a gossip protocol (mimics gossip about known peers) and message exchange is done hop-by-hop between mimics. It provides sender and receiver anonymity, communication unobservability and unlinkability.

MorphMix. MorphMix [23] is another P2P solution with architecture similar to Tarzan's. The main difference between both is the way messages are routed – MorpMix nodes only define the first node of a path that will be constructed along the way as the destination is known. This has the advantage of not requiring each node to have a complete view of the network. Yet, it disables recipient anonymity and creates a security breach that colluding nodes can use because the path can be constructed only with colluding nodes. MorphMix has a collusion detection mechanism but, as shown by [24], its local view is not enough to prevent all collusion attacks.

4.5 Network Coding

Network coding (NC) is general coding scheme that encompasses linear coding schemes. It started as a scheme that was only used by the network itself for obtaining better performance but presently it also used when referring to coding that is done by end-nodes. NC suffers from two serious privacy preserving problems: combination of packets with polluted ones (pollution attacks) and traffic analysis using the linear factors: they allow an attacker to track packet combinations.

Several NC solutions for P2P file sharing, such as [25] and [26], have been proposed but, at the best of our knowledge, only one has been presented for securing the network layer: the one presented by Fan et al. [27]. Their scheme uses homomorphic encryption functions (HEF) that allow intermediate nodes to combine packets without knowing the encrypted linear factors thwarting traffic analysis attacks that use them. Nevertheless, it is still susceptible to pollution attacks because only the recipient has the decryption key and only he can detect polluted packets.

4.6 Other Systems

P^5 [28] is a system that does not fit on any of the previous sections because it is based on message replication schemes with a hierarchical broadcast tree (see Fig. 11).

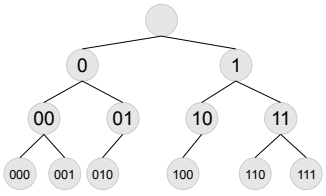


Fig. 11. Hierarchical broadcast tree.

P^5 stands for peer-to-peer personal privacy protocol. This solution allows the anonymity set to be chosen according to what is intended: privacy or performance. The performance is proportional to the length of the ID and the privacy is inversely proportional to ID's length: all leafs under a certain node receive the broadcast message. E.g., if 00 is the destination, the message will be received by 000 and 001.

5 Conclusions

In this survey it was presented a review of the main privacy techniques and solutions, and the terminology used in this research area. Also, a possible taxonomy was suggested based on very simple building blocks according to the network architecture and message types. Single entity controlled solutions, albeit being easier to configure and use, require the user to trust the solution provider; multiple entity ones are more complex and do not present such limitation. However, only Tor and JAP are widely deployed since the majority are more of theoretical interest and are impractical, mainly, for large and global networks because of either the broadcast requirements (DC-nets and P^5) or the traffic overhead (Tarzan and MorphMix).

The research on this area, although going on for almost three decades, is still very active and split message based solutions, mainly network coding, are very promising. A solution that presents a practical design, low-latency, robustness and efficiency is of high practical interest especially in networks with lower computational resources such as wireless sensor networks, given that Tor and JAP still have considerable computational requirements due to onion routing encryption.

Acknowledgments

The author would like to thank the support from the Portuguese Foundation for Science and Technology (FCT) under the fellowship SFRH/BD/69388/2010.

References

1. Seo, K., Kent, S.: Security Architecture for the Internet Protocol, http://tools.ietf.org/html/rfc4301.
2. Dierks, T.: The Transport Layer Security (TLS) Protocol Version 1.2, http://tools.ietf.org/search/rfc5246.
3. Crotti, M., Dusi, M., Gringoli, F., Salgarelli, L.: Traffic classification through simple statistical fingerprinting. ACM SIGCOMM Computer Communication Review. 37, 5–16 (2007).
4. Herrmann, D., Wendolsky, R., Federrath, H.: Website fingerprinting: attacking popular privacy enhancing technologies with the multinomial naïve-bayes classifier. Proceedings of the 2009 ACM workshop on Cloud computing security. pp. 31–42 (2009).
5. Deng, J., Han, R., Mishra, S.: Intrusion Tolerance and Anti-Traffic Analysis Strategies For Wireless Sensor Networks. International Conference on Dependable Systems and

Networks. p. 637 (2004).
6. Clarke, R.: Internet privacy concerns confirm the case for intervention. Communications of the ACM. 42, 60–67 (1999).
7. Krishnamurthy, B., Wills, C.E.: On the leakage of personally identifiable information via online social networks. ACM SIGCOMM Computer Communication Review. 40, 112–117 (2010).
8. Pfitzmann, A., Köhntopp, M.: Anonymity, unobservability, and pseudonymity — a proposal for terminology. Designing Privacy Enhancing Technologies. pp. 1–9 (2001).
9. Danezis, G., Diaz, C.: A survey of anonymous communication channels. Journal of Privacy Technology. (2008).
10. Proxify, http://proxify.eu/.
11. Socksify, http://socksify.com/.
12. Relakks, https://www.relakks.com/.
13. Chaum, D.: Untraceable electronic mail, return addresses, and digital pseudonyms. Communications of the ACM. 24, 84–90 (1981).
14. Chaum, D.: The dining cryptographers problem: Unconditional sender and recipient untraceability. Journal of Cryptology. 1, 65–75 (1988).
15. Danezis, G., Dingledine, R., Mathewson, N.: Mixminion: Design of a type III anonymous remailer protocol. Security and Privacy, 2003. Proceedings. 2003 Symposium on. pp. 2–15 (2003).
16. Dingledine, R., Mathewson, N., Syverson, P.: Tor: The second-generation onion router. Proceedings of the 13th conference on USENIX Security Symposium-Volume 13. p. 21 (2004).
17. Ren, J., Wu, J.: Survey on anonymous communications in computer networks. Computer Communications. 33, 420-431 (2010).
18. Berthold, O., Federrath, H., Köpsell, S.: Web MIXes: A system for anonymous and unobservable Internet access. Designing Privacy Enhancing Technologies. pp. 115–129 (2001).
19. Dolev, S., Ostrobsky, R.: Xor-trees for efficient anonymous multicast and reception. ACM Transactions on Information and System Security (TISSEC). 3, 63–84 (2000).
20. Reiter, M.K., Rubin, A.D.: Crowds: Anonymity for web transactions. ACM Transactions on Information and System Security (TISSEC). 1, 66–92 (1998).
21. Bos, J., den Boer, B.: Detection of disrupters in the DC protocol. Advances in Cryptology—EUROCRYPT'89. pp. 320–327 (1990).
22. Freedman, M.J., Morris, R.: Tarzan: A peer-to-peer anonymizing network layer. Proceedings of the 9th ACM Conference on Computer and Communications Security. pp. 193–206 (2002).
23. Rennhard, M., Plattner, B.: Introducing MorphMix: peer-to-peer based anonymous Internet usage with collusion detection. Proceedings of the 2002 ACM workshop on Privacy in the Electronic Society. pp. 91–102 (2002).
24. Tabriz, P., Borisov, N.: Breaking the collusion detection mechanism of MorphMix. Privacy Enhancing Technologies. pp. 368–383 (2006).
25. Lee, U., Park, J.S., Yeh, J., Pau, G., Gerla, M.: Code torrent: content distribution using network coding in vanet. Proceedings of the 1st international workshop on Decentralized resource sharing in mobile computing and networking. pp. 1–5 (2006).
26. Gkantsidis, C., Rodriguez, P.R.: Network coding for large scale content distribution. INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE. pp. 2235–2245 (2005).
27. Fan, Y., Jiang, Y., Zhu, H., Shen, X.: An efficient privacy-preserving scheme against traffic analysis attacks in network coding. INFOCOM 2009, IEEE. pp. 2213–2221 (2009).
28. Sherwood, R., Bhattacharjee, B., Srinivasan, A.: P5: A Protocol for Scalable Anonymous Communication. PROC. IEEE SYMP. SECURITY AND PRIVACY. 58--70 (2002).

SESSION 7

RECONFIGURABLE COMPUTING / SEMANTICS WEB

Chairman: Carlos Manuel Dias Viegas

Ali Azarian
Pipelining Producer-Consumer Tasks using Custom Multi-Core Architectures

Daniel Sampaio
Digital Teacher: Proposing the use of WEB 2.0 tools for collaborative construction of teaching knowledge

Adela Ortiz
Poliontо: Ontology reuse with Automatic Text Extraction from Political Documents

Pipelining Producer-Consumer Tasks using Custom Multi-Core Architectures

Ali Azarian

Faculdade de Engenharia da Universidade do Porto,
Rua Dr. Roberto Frias, s/n, 4200-465, Porto, Portugal
azarian@fe.up.pt

Abstract. In recent years, there has been an increasing interest on using task level pipelining to accelerate the overall execution of applications mainly consisting of producer/consumer tasks. This paper introduces a multi-core architecture using inter-stage buffers to communicate data and to pipeline producer/consumer pairs of tasks. We present and analyze three main types of inter-stage buffer behaviors. The experimental results show speed-ups both in in-order and out-of-order producer/consumer tasks using the inter-stage buffers proposed in this paper.

Keywords: Multi-core, task level pipelining, FPGA, inter-stage communication.

1. Introduction

One of the major efforts in computer architecture is dedicated to improve the performance of computing systems. Recently, researchers have shown an increasingly interest in many-core or multi-core systems [7] which are one of the feasible options to improve the performance of computing systems. A multi-core computing system is composed of two or more independent cores, which can execute different instructions in parallel.

One major practical approach for improving processing power and efficiency in multi-core processors is the use of FPGAs (Field Programmable Gate Arrays) [8]. An FPGA is a device that contains a matrix of reconfigurable gate array logic circuitry. FPGAs are truly parallel in nature so different processing operations do not have to compete for the same resources. Hence, FPGAs provide many opportunities for task-level pipelining with multi-cores.

Task-level pipelining is an increasingly important topic area in multi-core processors as it provides additional speedups over the ones achieved by the use of parallelism in data-independent tasks. Thus, in recent studies a variety of methods have been proposed to assess task-level pipelining in multi-cores [9]. One of the most significant current discussions in task-level pipelining is inter-process communication between multi-cores. One of the simplest model of multi-core system is producer/consumer pairs which can communicate by using a FIFO channel [1][2]. In this model Producer/Consumer (P/C) pairs are performing in parallel.

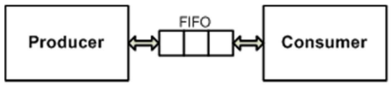


Fig. 1. Producer/Consumer Pairs using FIFO inter-process communication

Producer/Consumer (P/C) pairs which communicated by using FIFO channels has shown in Fig. 1. Produced data will be stored in FIFO channel while consumer is using data which is stored in FIFO channel in parallel manner.

```
for(int i=1; i<=N; i++){      for(int i=1; i<=N; i++){
  for(int j=1; j<=N; j++){     for(int j=1; j<=N; j++){
     a[i, j] = Fp();}             Fc( a[i, j] );}
}                               }
        Producer                      Consumer
```

Fig. 2. An in-order P/C pairs

The order of P/C pairs could be classified in two different types: in-order and out-of-order P/C pairs. In-order can be defined as a P/C pairs where every two consecutive Consumer iteration points are mapped onto two consecutive Producer iteration points [3] as shown in Fig. 2.

```
for(int i=1; i<=N; i++){      for(int i=1; i<=N; i++){
  for(int j=1; j<=N; j++){     for(int j=1; j<=N; j++){
     a[i, j] = Fp();}             Fc( a[j, i] );}
}                               }
        Producer                      Consumer
```

Fig. 3. An out-of-order P/C pairs

By contrast, out-of-order P/C pairs are generally defined as the change in consuming data which is consumes the produced tokens in a different order than they are produced. Through exchanging at the Consumer side the indexes of the array a (from *a[i,j]* to *a[j,i]*) as presented in Fig. 3.

This paper focused on introducing architecture for task-level pipelining by using an inter-stage buffer. We presented the speed-ups of this architecture for different inter-stage buffer schemes: considering only internal (local) buffer, only external buffer, and both.

This document has been structured in following order: Section 2 described and introduced three different architectures inter-stage buffer. In section 3 these architectures are evaluated using a number of benchmark applications to show the variety of speed up. Section 4 described the related work in multi-core systems and inter-process communications. Section 5 we presented some preliminary conclusions. In section 6 further work is described.

2. Multi-Core Architecture

To accelerating producer/consumer pairs using FIFO channels multi-core architectures have been proposed by many authors (see, e.g., [1][2]). As shown in Fig. 1, one core is assumed as a producer of data and the other core adopted as a consumer of the produced data. The inter-connection between cores is a simple FIFO channel.

If the order of requested data from consumer is entirely similar to the order of existed produced data in FIFO channel, in-order P/C pairs can implement by using this architecture and the speed up would be considerable. On the other hand, this architecture explained by the fact that it's not desirable architecture for out-of-order applications. In this issue, Consumer cannot receive the requested data till the previous produced data which has been stored in FIFO channel didn't consumed. The issue has grown in importance in light of recent studies to solve out-of-order applications problem [3].
In this study a new architecture is presented by using inter-stage buffer between P/C pars with three different behaviors. The main important purpose of inter-stage buffer is to prevent of producer/consumer misses by using local and external memories to store produced data.

2.1 Inter-stage buffer using internal memory

The first scheme for inter-stage buffer behavior is using an internal memory as a buffer to store produced data based on buffer size.

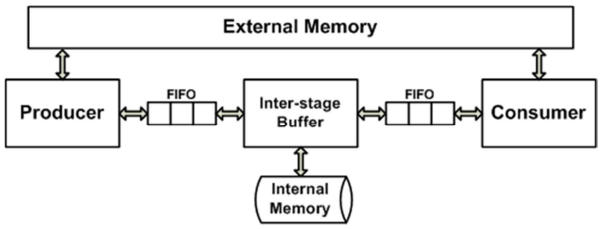


Fig. 4. inter-stage buffer using only internal memory

As shown in Fig. 4, Producer load input data from external memory and produced data will be store in internal memory based on buffer size. The amount of buffer size in this study is 1024. If buffer is full, producer will be stop till consumer consumes the stored data of buffer. As soon as each cells of buffer array would be empty, producer will produced a data and it will store on empty cell.

For the purpose of mapping produced index and data onto internal memory with the limitation of buffer size, a simple hash table is used. A hash table or hash map is a data structure that uses a hash function to map identifying values, known as keys.

```
while(1) {
 if(put == 1) {
  get(index_P);      // get the index from producer
  get(data_P);       // get the data from producer
  hash_P = index_P & MASK;
 if (flag_local[hash_P] == 0) {
  data_local[hash_P] = data_P;
  flag_local[hash_P] = 1;  }
}
 if (get==1) {
  get(index_C);      // get the index from Consumer
  hash_C = index_C & MASK;
if(flag_local[hash_C] == 1) {
  data_C = data_local[hash_C];
  put(data_C);       // send the data to Consumer
  flag_local[hash_C] = 0;
  get==1;
}
else get=0;
}
}
```

Fig. 5. Pseudo code of inter-stage buffer using only local memory

The most essential advantage of using only internal memory as an inter-stage buffer is founded in the speed of storing data on buffer; on the other hand, the limitation of buffer size to store produced data will be a reasonable cause of the latency on the producer. In this approach, Producer would have been idle during the buffer consuming. The pseudo code of inter-stage buffer using internal memory is presented in Fig. 5.

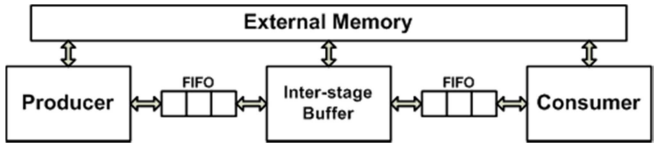


Fig. 6. Inter-stage buffer using only external memory

2.2 Inter-stage buffer using external memory

The second scheme for inter-stage buffer behavior is using external memory. As shown in Fig. 6. All of the three cores have access to external memory directly and produced data will be store in external memory entirely. Based on huge size of external memory, the producer wouldn't have been idle during the buffer consuming.

```
while(1) {
 if(put == 1) {
  get(index_P);        // get index form producer
  get(data_P);         // get data form producer
 if (flag_external[index_P] == 0)
{
  data_ external [index_P] = data_P;
  flag_ external [index_P] = 1;}
}
 if (get==1) {
  get(index_C);        // get the index from Consumer
 if(flag_ external [index_C] == 1) {
  data_C = data_ external [index_C];
  put(data_C);         // send the data to Consumer
  flag_ external [index_C] = 0;
  get==1;
  } else get=0;
}
}
```

Fig. 7. Pseudo code of inter-stage buffer using only external memory

However, according to latencies of access to external memory, the performance of this approach is lower than the one achieved by the previous behavior. The pseudo code for inter- stage buffer behavior is shown in Fig. 7.

2.3 Inter-stage buffer using internal and external memory

Although the inter-stage buffer using internal memory is faster than using external memory, the key problem with the first scheme is that the buffer size is limited and storing all produced data in the local buffer is impossible for most real applications. On the other hand, the main weakness of the second scheme is the high latency of accessing external memory.

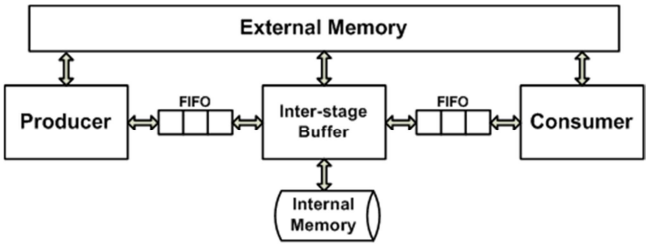


Fig. 8. Inter-stage buffer using internal and external memory

```
while(1) {
 if(put == 1) {
  get(index_P);        // get index form producer
  get(data_P);         // get data form producer
  hash_P = index_P & MASK; // computing the hash table

 if (flag_local[hash_P] == 0) {
  data_local[hash_P] = data_P;
  flag_local[hash_P] = 1;  }
else if (flag_external[index_P] == 0) {
      data_ external [index_P] = data_P;
      flag_ external [index_P] = 1;
}
}
 if (get==1) {
  get(index_C);        // get the index from Consumer
  hash_C = index_C & MASK;
 if(flag_local[hash_C] == 1) {
  data_C = data_local[hash_C];
  put(data_C);         // send the data to Consumer
  flag_local[hash_C] = 0;
  get==1;
  }
else if(flag_ external [index_C] == 1) {
      data_C = data_ external [index_C];
      put(data_C);     // send the data to Consumer
      flag_ external [index_C] = 0;
      get==1;
    }
  else get=0;
}
}
```

Fig. 9. Pseudo code of inter-stage buffer using local and external memory

To address these limitations an inter-stage buffer behavior using internal and external memory shames is presented in Fig. 8. In this scheme, first produced data will store in internal memory with buffer size 1024. If buffer has been full, produced data will be store in external memory.

Concurrently, if the requested data from consumer doesn't exist in internal memory, it will be seek in external memory based of index. The advantage of using local and external memory is that both of in-order and out-of-order application would be

supported by this scheme. In out-of-order issues, consumer will find the request data in internal or external memory In any case. The pseudo code of the inter-stage buffer behavior is presented in Fig. 9.

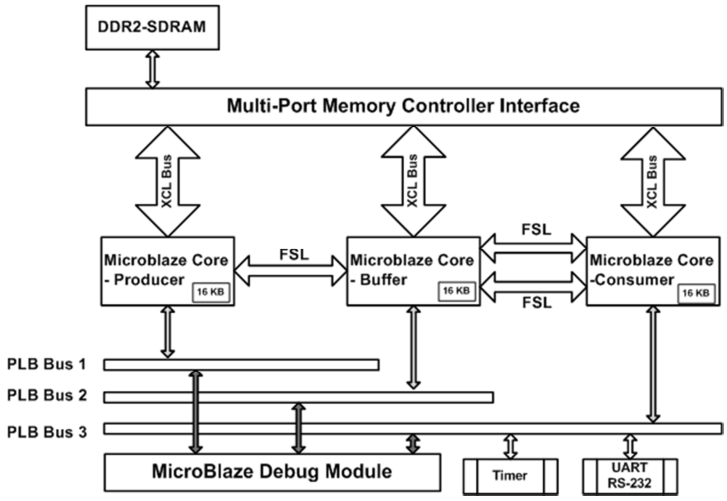


Fig. 10. Block-diagram of suggested architecture

3. Experimental results

This section presents experimental results using the three schemes for the inter-stage buffers previously presented. We use Xilinx Embedded Development Kits 12.3 (EDK) to create multi-core architectures consisting of Microblaze processors [10]. The Microblaze is a soft RISC processor core designed for Xilinx FPGAs. We measured the speed-up of each three schemes using the following benchmarks: gray image histogram, Fast Discrete Cosine Transforms (FDCT) [11] and a wavelet transform [12].

To establish the communication channel between cores Fast Simplex Link (FSL) channels has been used as shown in Fig. 10. FSL is a uni-directional point-to-point communication channel bus based on FIFO used to perform fast communication between any two-design elements on the FPGA when implementing an interface to the FSL bus [10]. All three Microblazes are connected to an MPMC (Multi-Port Memory Controller) for accessing a shared memory system using a DDR2-SDRAM. To access MPMC directly, Xilinx Cash Link (XCL) buses has been used.

Microblazes are connected to MDM interface using PLBs (Processor Local Buses) in order to provide bus infrastructure for connecting an optional number of PLB masters and slaves into an overall PLB system. In this block diagram, three IPs

Table 1. Speed-up of architectures using some benchmark applications

Benchmark	Architecture	Overall Latency	Speed-Up
Gray/Histogram	One Microblaze	116392034	1
	Two Microblaze	43070584	2.70
	Three Microblazes Using Just Local memory	44041318	2.64
	Three Microblazes Using Just Ext Memory	61343260	1.90
	Three Microblazes Using Local & Ext. Memory	44041324	2.64
FDCT	One Microblaze	83,374,184	1
	Three Microblazes Using Just Local memory	57,424,080	1.45
	Three Microblazes Using Just Ext Memory	102,802,928	0.81
	Three Microblazes Using Local & Ext Memory	76,977,093	1.08
Wavelet Transform	One Microblaze	62405741	1
	Three Microblazes Using Just Local memory	46447631	1.34
	Three Microblazes Using Just Ext. Memory	54180898	1.15
	Three Microblazes Using Local & Ext. Memory	44624902	1.40

connected to PLB, Microblaze Debug Module (MDM), Timer and RS-232 Port to monitor the results on the screen. In this experimental setup, the input image size in all benchmarks is 1024×768 and none of the above Microblazes are doesn't applied instruction/data cache.

Table 1 shows the clock latencies of performing tasks in consumer is measured and the speed-ups of each architecture and schemes are obtained. The speed up is a division of Overall latency of tasks by overall latency of one Microblaze architecture. For example in Gray/Histogram benchmark, in three Microblaze using just local memory architecture, the speed up is 2.64 (116392034/44041318) more than one Microblaze. It means that by using two Microblaze and split the task level codes in two cores, the speed up significantly increased more than 2 times.

In addition, the table 1 illustrates that there is a significant difference in speed-up between using one Microblaze and using two or three Microblazes for task level pipelining. For instance, FDCT benchmark is an out-of-order application and the speed-up is considerably increased compare with one Microblaze. As we mentioned in previous section, performing out-of-order applications on two Microblaze architecture is not reasonable. Therefore, in FDCT and Wavelet benchmark, three suggested schemes are presented.

4. Related work

Several research efforts have considered multiprocessor systems on FPGAs have appeared. For instance, Clark et al. [5] and Roxby et al. [6], presented shared memory

Multiprocessor case studies with ad hoc mechanisms for synchronization. This paper is not intended to contribute to the state-of-the-art of these architectures or compilers, but want to demonstrate flexible architecture for the validation of new ideas in the task level pipelining based on FPGAs.

A multitude of works efforts methods for solving out-of-order communication in Kahn Process Networks. We mention the ones from A. Turjan et al. [1]. A related approach has recently been proposed by Byrd et al. [4] who classified producer-initiated mechanisms as implicit or explicit, according to whether the producer must know the identity of the consumer when data are transmitted. The mechanisms in this study are evaluated for performance and sensitivity to network parameters, using a common simulated architecture and a set of application kernel benchmarks. Our work, instead, used the concept of this mechanism to introduce new architecture to decrease the latencies in task level pipelining.

5. Conclusion

The present study was designed to determine the effect of inter-stage buffer in measuring the speed-up and latencies. One of the most significant findings to emerge from this study based on evidence is that using inter-stage buffer in the middle of P/C pairs increased the speed of task level pipelining even for out-of-order applications such as FDCT and wavelet transform. In general, therefore, it seems that even without solving the out-of-order application in compile time, the suggested architecture decrease the latencies in consumer core and increase the speed of system.

6. Ongoing work

First, determining the order of application in compile time, then decreasing the frequency of idle cores to save the power consumption dynamically is an interesting discussion that is our ongoing work.

References

1. Turjan, A.; Kienhuis, B.; Deprettere, : A compile time based approach for solving out-of-order communication in Kahn process networks, Proceedings IEEE International Conference on Application- Specific Systems, Architectures, and Processors, p 17-28, 2002.
2. M Ben-Ari. Principles of Concurrent Programming. Prentice Hall, 1982.
3. Turjan, Alexandru; Kienhuis, Bart; Deprettere, Solving out-of-order communication in Kahn Process Networks, Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology, v 40, n 1, p 7-18, May 2005.
4. Byrd, G.T. ; Flynn, M.J. Source: Proceedings of the IEEE, v 87, n 3, p 456-66, March 1999
5. C. R. Clark, R. Nathuji, and H.-H. S. Lee. Using an fpga as a prototyping platform for multi-core processor applications. In WARFP-2005: Workshop on Architecture Research using FPGA Platforms, February 2005.

6. P. James-Roxby, P. Schumacher, and C. Ross. A single program multiple data parallel processing platform for FPGAs. FCCM 2004: 12th Annual IEEE Symposium on Field-Programmable Custom Computing Machines, pages 302–303, April 2004.
7. Xing, Jianguo ; Zhao, Wenmin; Hu, A. An FPGA-based experiment platform for multi-core system Proceedings of the 9th International Conference for Young Computer Scientists, ICYCS 2008, p 2567-2571, 2008,Proceedings of the 9th International Conference for Young Computer Scientists, ICYCS 2008
8. Wang, Jie ; Zhang, Shu-Yan; Liu, Tao; Ji, Zhen-Zhou; Hu, Ming-Zeng, Multi-core embedded processor based on FPGA and parallelization of SUSAN algorithm, Jisuanji Xuebao/Chinese Journal of Computers, v 31, n 11, p 1995-2004, November 2008.
9. Kim, D. ; Kim, K.; Kim, J.-Y.; Lee, S.; Yoo, H.-J. Memory-centric network-on-chip for power efficient execution of task-level pipeline on a multi-core processor, IET Computers & Digital Techniques, v 3, n 5, p 513-24, Sept. 2009.
10. FPGA and CPLD Solusion for Xilinx, http://www.xilinx.com/tools/microblaze.htm
11. M. S. Corrington, Implementation of fast cosine transforms using real arithmetic, National aerospace of electronic Conf., (NEACON), May 16-18, 1978, Dayton, OH.
12. C. Tenllado, J. Setoain, M. Prieto, L. Pin˜uel and F. Tirado; Parallel Implementation of the 2D Discrete Wavelet Transform on Graphics Processing Units: Filter Bank versus Lifting, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 19, NO. 3, MARCH 2008.

Digital Teacher: Proposing the use of WEB 2.0 tools for collaborative construction of teaching knowledge

Daniel Sampaio

Faculty of Engineering – University of Porto (FEUP)
Rua Dr. Roberto Frias, 4200-465 – Porto, Portugal
pro10011@fe.up.pt

Abstract. In Portugal, the several portals aimed at teachers do not as yet make use of nor fully exploit the potential of different tools provided by Web 2.0. Most of them have not yet made the transition from Web 1.0. Some of them include forums, wikis and support material (not made available through Web 2.0) – items that could be considered common in a portal. Many of the teachers already using the portals are not yet fully familiar with the concept of Web 2.0 and are not aware of its potential. In response to the need for teachers to use Web 2.0 tools in the classroom and to develop virtual learning communities, we propose the creation of the portal "*Professor Digital*". This paper present a portal that will enable teachers to access a wide range of personalized Web 2.0 tools presented in a user friendly way. They will be encouraged to participate in virtual learning communities through the use of the different Web 2.0 tools on the portal.

Keywords: Web 2.0 tools, Teacher learning communities, Collaborative teacher knowledge, Web 2.0 education, Web evolution.

1 Introduction

Today, the term Web 2.0 is used to describe applications that distinguish themselves from previous generations of software by a number of principles. These new, Web 2.0, applications take full advantage of the network nature of the Web: they encourage participation, are inherently social and open. Whereas Web 2.0 is not characterized by a new step of technology as is the Semantic Web [1], in the last years the Web changed from a medium to a platform, from a read-web to a read-write-web, thereby fulfilling Berners-Lee's original vision of the Web [2].

Innovative, full of potential and based primarily on interaction, collaboration and active user participation, Web 2.0 may have a very important place in education [3]. In order to keep pace with the development of this second phase of the Web, it is important that teachers use it effectively within their work. However, implementation is complex and raises many issues. To overcome these it is essential to be aware of the potential of available technologies and how they can be used in an educational setting. The overall aim of this work was to study and promote the use of Web 2.0 tools among teachers. Considering that teachers have an ever increasing need to develop their skills in educational technology, educational portals have an important

role to play in helping them to obtain training and support in this – something which should be developed.

It should be noted that, in Portugal, portals aimed at teachers have not adequately exploited the potential of Web 2.0 tools. Many teachers using them were not fully aware of the scope and potential of the technologies they were using. Initially, the idea of creating a new portal to more fully utilise Web 2.0 was considered. Then the concept of the portal "*Professor Digital*" (Digital Teacher) was conceived, which intended to provide for teachers a single customised set of Web 2.0 tools as well as information and support for them. In accordance with the spirit of Web 2.0, it was intended that the portal should encourage interaction and collaboration between teachers.

It was hoped that an environment could be created that would stimulate the use of Web 2.0 tools in teaching work, and that this would also enable the creation and accommodation of professional groups which could eventually grow into true online communities.

2 The Web 2.0 "revolution"

2.1 Main Features of Web 2.0

Web 2.0 is the term used to refer to the second generation of the World Wide Web. It reinforces and encourages exchange and cooperation between users on websites and virtual services. By collaborating in organizing and producing online content, it is intended that users make the web more dynamic and interactive. Web 2.0 is encapsulated in a new generation of Web applications which support multiple users who can access a set of highly dynamic and interactive tools. These strongly facilitate the collaboration and sharing of content among users.

We are experiencing a qualitatively different era of the Web, as evidenced by most Web applications and Web sites that have emerged and transpired over recent years. These Web applications are considered to be fundamentally different from traditional Web applications. One of the main ideas behind Web 2.0 is usability [4]. Web 2.0 applications approximate the look and feel of desktop applications and provide a far richer user experience and interaction capabilities. From the user's point of view, users have been offered new means of accessing information on the Web and sharing knowledge and ideas among others; from a technical point of view, Web applications have become more responsive and better in dealing with network latency. Web 2.0 also distinguishes itself from the early Web by how Web applications are built and delivered. The Web has changed from simple websites to Internet services, which gathers data from multiple sources and devices in real time, allows individuals to contribute ideas and contents, and delivers software as a continually updated service. One consequence of the Web as platform is that the applications and services remain in a kind of "perpetual beta" stage and are constantly refined and improved [5]. Therefore, a new type of Software Development Life Cycle has emerged and promises to continue affecting Web design and development in the future.

The most important factor in this evolution is to develop applications that take advantage of network effects and collective intelligence to improve their use [6]. Figure 1 gives details of some of the many Web 2.0 tools currently available, some of which are very popular and widespread. Many of the users of these will not be aware or have knowledge of the term “Web 2.0”.

Fig. 1. Web 2.0 tools

As illustrated in Figure 2, the creation of content becomes collaborative, and communication is made through the use of varying technologies. Configuration is simple, without the need for great technical expertise.

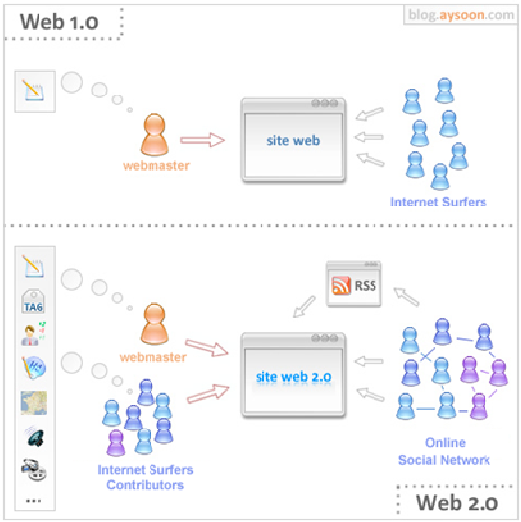


Fig. 2. Main difference between Web 1.0 and Web 2.0

Source: http://blog.aysoon.com/

Web 2.0 is dominated by user communities or social networks who actively create and distribute content free of authorship or copyright restrictions. This is in line with the concept that the Internet is evolving from a medium for information into a platform. As such there is a change in focus which develops as shown in Figure 3.

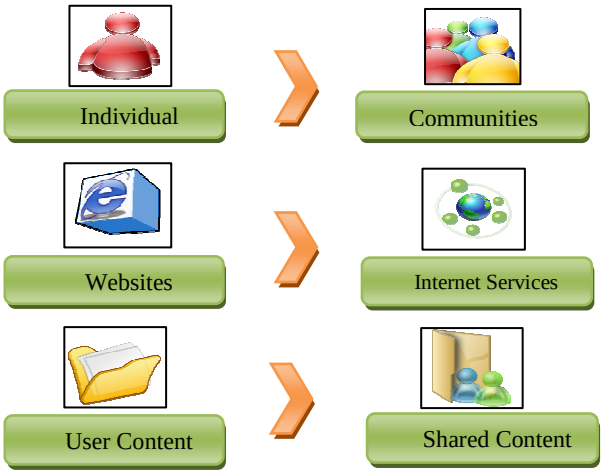


Fig. 3. Evolution for Web 2.0

We should pay particular attention to the use of the Internet and related services as a tool for both individual and collaborative learning. The web aids the learning process through searches, both free and structured, and also because it functions as a means of presenting and sharing work with all who would wish to access it online. With Web 2.0 software is available online to facilitate immediate editing and publishing, eliminating the need for the user to have the programs on their own machine. Wikis, podcasts and blogs are all examples of Web 2.0 features that allow users to work online.

In terms of learning, we can note the following major developments in the evolution of the Internet:

Web 1.0 – query/search

Web 2.0 – see/find/share

Web 2.0, a new phase that is actually at the maximum stage of its development, marks a true social revolution in terms of the importance it gives to the virtual interaction with other internet users.

Then come true learning communities, in which each participant is a real agent in the joint growth of knowledge – this is collective intelligence.

Regular websites, fundamentally for the display of information, have given way to blogs which are virtual places to share personal thoughts and ideas and which are open to comments from visitors. Also arising are virtual realities, three-dimensional worlds that simulate real-life social environments. “Second Life” is one of the best known of these.

Web 3.0 - see / search / share / find

We are now at a stage when Web 3.0 or the Semantic Web is emerging or developing. Web 3.0 is or will be characterized by the use of applications and content from multiple sources, combined or not, that the user will be able to use in their blog or website. Web 3.0 is the product of a technological revolution that makes it possible to organize and use more effectively the knowledge already available on the Internet. Data categorization and the building of interlinked databases will make searches much more directed and consequently more rapid and effective. An example of the future lies in the search engine *Wolfram Alpha (http://www.wolframalpha.com)*, which gives the user an answer instead of providing responses, as is the currently the case with search engines of today. They, the users, in each search, will form links between data, and in this fashion, "teach the machine".

The concept of the Web is, therefore, constantly evolving and, for now, Web 3.0 will be the future Web as far as "machine thinking" goes. After teaching the machine to learn, it is certain that we will continue to reinvent and develop ways to make learning more rich and effective.

So while it could in some ways be considered "revolutionary", Web 2.0 is only another stage in the development of the World Wide Web. It is important to take the opportunity to explore the possibilities that open in terms of learning and development opportunities among groups and communities and to watch how these integrate and relate.

2.2 Web 2.0/E-learning 2.0 and Virtual Communities (VC)

A Virtual Community is formed between individuals online who share common interests, values and objectives, and who, through common projects and exchanges establish a cooperative process [7]. The principal requirement of a VC is to be a group of people who establish between themselves social relations in a virtual environment.

The concept of VC was first coined in 1993 by Howard Rheingold[1], and they are defined as social groupings, originating through the internet, that consist of invisible interlocutors possessing various interests ranging from the scientific to the abstract.

The concept of virtual communities has been introduced to the teaching environment, leading to the development of virtual learning communities (VLCs) as online centres of education and social interaction. VLCs are used for intellectual, social and cultural exchanges, establishing relationships through linked computers. Several software packages designed to facilitate the development of VLCs are currently available through the Internet.

In VLCs users share information within a shifting network of colleagues through user profiles linking users to others posting similar information. In user profiles, each piece of data is a link; clicking on it displays everyone else in the network who included that element in their profiles. Other connections are more structured, based on user created groups that typically have descriptive titles [8].

[1] *Howard Rheingold* is a critic, writer, and teacher; his specialties are on the cultural, social and political implications of modern communication media such as the Internet, mobile telephony and virtual communities (a term he is credited with inventing)

Because it lies mainly on the contribution and active participation of each user, Web 2.0 becomes most important to education. In an attempt to follow the evolution of this second Web generation, e-Learning 2.0 is appearing and teachers are starting to use this social software in effective ways [9].

The goal is to construct a virtual environment where teachers can create and share contents, building on knowledge and professional development in a collaborative way. It is also intended to promote reflection on the integration and impact of Web 2.0 tools in these contexts.

Web is no longer an information repository or a place to search for resources. The new Web a place to find other learners, to exchange ideas and thoughts, to demonstrate creativity, and to create new knowledge. With the new tools and services provided by the new Web, it starts laying the foundation for innovative ideas such as Classroom 2.0, Library 2.0, School 2.0, University 2.0, and E-learning 2.0.

E-learning 2.0 can capitalize on many sources of content aggregated together into learning experiences and utilize various tools including online references, courseware, knowledge management, collaboration and search. E-learning 2.0 differs from traditional e-learning. Instead of learners simply receiving, reading, and responding to learning content in traditional e-learning; e-learning 2.0 allows learners to create content and to collaborate with peers to form a learning network with distribution of content creation and responsibilities. In addition, e-learning 2.0 allows learners to easily access content through search, aggregation, and tagging. It provides learners with opportunities to interact with the content and share their thoughts and comments with not only the instructors but also with other learners. E-learning 2.0, therefore, is evolving to one of the most exciting, dynamic, and challenging fields involving teaching and learning.

2.3 Constructing portals that integrate Web 2.0 tools

A portal is a technological application (website) that is intended to collect, organize and distribute contents answering to the functional requirements of a virtual community. The great advantage of the construction of portals inhabits in the use of Content Management System (CMS) tools, in this case - *Joomla*, which facilitate the dynamics of content creation, storage, administration and distribution, through a user friendly interface. CMS tools can become excellent supporters in teaching-learning process and in the organization and allotment of the information produced in environments with educational aims.

The generalized use of the Internet resulted in the creation of a new type of social organizations that allows the construction of virtual communities (human groups with common interests). The use of portals allows these groups to develop themselves, sharing contents and enriching experiences, and promotes the construction of knowledge, based on the interactions between the members of the communities.

With the advent of Web 2.0 websites have undergone a huge change in character. Users can now add and modify content and personalize portals. In accordance with the principals of Web 2.0, content must be free of ownership restrictions, permitting modification and reuse. Associated with opportunities for collaboration, interaction

and participation, Web 2.0 makes possible the generation and interaction of online communities through commentaries on blogs and social networking sites.

Web 2.0 sites do not exist in a finished form or version, since, due to their being part of a network, they are in a constant form of evolution or "permanent beta testing" through feedback and constant function testing from users.

In contrast to traditional software, Web 2.0 applications are no longer released in version-based software packages, one version at a time, but are constantly refined and improved.

Changes to services happen gradually, this is facilitated by the ability of Web applications to track the user's interaction with the service and thereby gathering data about interaction patterns that is nearly impossible to collect for desktop applications. While constant improvement of a service is not a bad thing, new or changed features may lead to confusion of learners who are using the service and may lead to distract from the task at hand. Changes in functionality and user interfaces require adapting previously written manuals. Sometimes parts of services are stopped completely, as it happened recently with the fee-based Google Video. This principle does not endorse a pedagogical theory. However, it has an effect on teachers/researchers that use a specific service. One advantage of the perpetual beta is that the developers are usually open to suggestions from adaptors. They often set-up developer discussion boards and use these to receive additional feedback.

3 The "*Professor Digital*" (Digital Teacher) portal

3.1 Presentation of the portal - Placement

Preliminary research into the integration of Web 2.0 tools in the main educational portals in Portugal has shown that they are still underexploiting the potential of the technology. The portals are not supporting or encouraging teachers to use them. The analysis looked at thirteen educational portals considered to be widely used – it focused on portals based in Portugal and used various parameters employed in site evaluation [10] as well as studying the use of Web 2.0 tools.

After looking at various portals, it was concluded that:

a) Registration – the majority of portals required registration through a login and password created by the user

b) Subscribers – The information presented, was in the main, aimed at primary and secondary school teachers of various subjects.

c) Services – The services most commonly available through the portals were: newsletters, electronic mail, content download and election participation.

d) Information – The information available was very varied, but was essentially related to educational matters.

e) Usability – In terms of usability, all the portals appeared to have user friendly interfaces.

f) Interaction – A search engine function was present in most portals. The portal "Professores Inovadores" permitted the creation of communities.

g) Navigation – Navigation of all portals was easy and intuitive with well- positioned menus.
h) Web 2.0 tools – The portals studied all underused Web 2.0 tools – all had forums, but only a minority used RSS and blogs and only one utilized a video conferencing tool.

We can conclude from this analysis that, in spite of having some good points, the education portals studied do not yet use Web 2.0 tools in a fully integrated way.

Some explanations were put forward to account for this. It is possible that the portals have not yet fully integrated the new technologies or adapted to the social and educational realities now existing. It is also possible that the slow take up of Web 2.0 tools in this area may be due to lack of skills or confidence among users. In addition, it is also possible that they may be reluctant to transfer materials to the new technology or change working practices they already have [11]. It is important to create conditions that facilitate this.

It was concluded that, in Portugal, education portals underused the potential of Web 2.0. Integration was still limited. It was felt that teachers could be missing out in much that could be useful to them in their work, and that they should be made more aware of the possibilities of the technology.

Portal development is an on-going process, and the incorporation of new technologies is gradual. Where appropriate, new technologies can greatly assist in teaching. It is important to promote and support the use of new tools such as Web 2.0. The idea of the "*Professor Digital*" (http://www.professordigital.net) was conceived as an attempt to mitigate problems with the integration of Web 2.0 tools in this area and to improve their take up.

This project, of a new educational portal, aims to integrate Web 2.0 tools to implement teaching knowledge in a collaborative way.

Beyond the spreading and the availability of the new technologies and concepts, this process will involve experimentation and investigation in educational contexts, improving and evaluating its use in collaborative learning communities. This is the purpose of the Digital Teacher portal project.

The construction of this portal also intends to create an enthusiastic atmosphere through the use of attractive technologies and the possible organization of online events: the user, being also the creator of the page, feels the necessity of boosting his space. Once this application makes possible a process of allotment, when there is a problem or situation to solve, opinions and ideas of other users are looked.

While diverse existing portals are static and require the download of sites, with many menus, hyperlinks and sometimes disorganized contents, the Digital Teacher portal, applying the techniques of Web 2.0, will become simpler, easier to navigate, intuitive and with excellent usability. Among the Web 2.0 tools that will be available in the portal, the user will choose those that better fit his/her objectives. These technologies intend to give to the user the full control of his/her contents. If it is unquestionable that the tools are extremely important, the higher challenge of Web 2.0 does not rely on computers, but on the users' attitudes. They must understand and use adequately the power of these new technologies.

This portal will act as a unique and personalized point for teachers to access a diversified Web 2.0 tools and the information that they need to use them. It will

establish a net of virtual communities distributed by diverse themes, defined by the users, aiming to provide learning experiences based on social construction of knowledge, contextualized learning and collaboration, involving communication, interaction and discussion.

Aims and objectives

The portal was designed to meet the needs of teachers who wished to use Web 2.0 tools in their work individually or collaboratively, particularly to organise, store and present online material and allow its distribution. It was also designed to support the creation of virtual working and learning communities. It was hoped that the portal would provide a single access point to a diverse range of upgradeable Web 2.0 tools and the support and information needed to use them.
It was intended to build a portal for teachers which would constitute a single point of access to a diverse panel of customisable and upgradeable Web 2.0 tools, together with the information and support needed to use them. Users would be able to choose the best tools to meet their needs and objectives, and they would be supported in this by the administrator and the features of the portal itself.

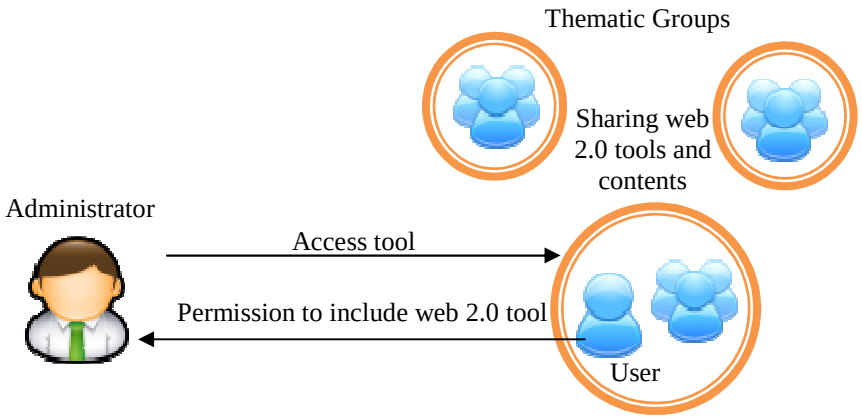


Fig. 4. "*Professor Digital*" Portal

Web 2.0 refers to a new set of web services and products that strengthen the concepts of exchange and contribution between users. The main change in Web 2.0 lies in the relation between user and information: whilst previously it is transmitted and consumed, now it becomes cooperative and participated, placed in sites where the content is selected, shared, modified, rebuilt and distributed, in communities where users communicate and interact.
The portal in this manner would assume a double function – on one side, it would make available Web 2.0 tools and on the other, stimulate and assist with their use technically and pedagogically.

Conclusion and Future Work

The main teaching portals in Portugal still don't implement the use of Web 2.0 tools in a wide and integrated form. We can advance some possible explanations for this fact.

Firstly, the portals may have failed the integration of these new tools, lacking adaptation to the new technical and social realities. This effort requires discovering and exploring the educational potentialities of Web 2.0 tools and how to make them available to the users in an integrated and accessible way.

On the other hand, the users may not yet possess enough digital culture and confidence to use these tools. In other cases, they are not able to transfer to the educational domain the digital abilities and practices developed in leisure contexts.

This is an opportune moment for a deep analysis of these tools, because it approaches a new age, the Web 3.0, and there is still little income of the age Web 2.0. Web 3.0 will attribute clear and more specific meanings to page contents, interpreting and contextualizing data. It is the attempt to invert the solution of how to improve the access to great volumes of information: the machine will play the role of the man and not the opposite. The implementation of Web 3.0 will result in the conjunction of new technologies and the experience and knowledge acquired by means of Web 2.0 use. Thus, when developing the integration of these tools in the portals, it's necessary to evaluate the modalities and the impact of its use by teachers.

In the context of The Information and Knowledge Society, which influences practically all areas of activity, we have assisted in recent decades a change in use of the information and communication technologies motivating the participants in collaborative growth of knowledge. The limitations of the traditional relationship between man and machine has led to the development of virtual environments for social interaction. These enable a wide range of social activities and relationships that enhance learning and group development.

This present work sets out to detail the development process and testing of the "*Professor Digital*" portal, which intends to provide a single point of access to a personalised panel of web 2.0 tools and to stimulate and assist their use in both collaborative and individual teaching work. The portal should promote group learning, developing communication and sharing of ideas and information, and permit the growth and establishment of online communities.

The portal presented here will encourage and assist in the use of these types of tools. For teachers, this will contribute to the development of their skills in the use of educational technology – in Portugal, this is in line with the Educational Technology Plan [12].

"Professor Digital" should the subject of future analysis and development, in accordance with the guidelines in place for studying the creation and dynamics of online communities [4,13,14]. It will also be essential to carry out a systematic and controlled evaluation of the portal and its use.

It remains only to reiterate and underline the potential of this environment (the portal) for studying and understanding some of these phenomena, and, ultimately, this latest phase in the development of the Internet.

Web 2.0 tools have a great potential for education, which is still not profited by educational portals. Collaboration is the keyword of Web 2.0, where it is possible not

only to find contents, but also to transform, to reorganize, to classify and to share them, developing cooperative learning, constructing collective knowledge. The use of the innumerable tools that Web 2.0 provides can motivate teachers to participate in the collaborative construction of knowledge in virtual educational communities. This is the main objective of the *"Professor Digital"* portal, which we aim to develop and evaluate.

While Sir Tim Berners-Lee was inventing a World Wide Web, he would not have an idea that the most accessible an ubiquitous source of knowledge and communication has been created.

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientic American, (2001)
2. Berners-Lee, T.: Weaving the Web. Orion Business Books, (1999)
3. Wang, Y., Zahadat, N.: Teaching Web Development in the Web 2.0 Era, In SIGITE '09 on SIG-information technology education, (2009)
4. Coll, C., Bustos, A.; Eengel, A.: Las comunidades virtuales de aprendizaje. In C. Coll & C. Monereo (Eds.) Psicología de la educación virtual, Morata, pp. 299-320, Madrid (2008)
5. Umbach, J.: Web 2.0 - the New Commons. Feliciter, p.192. (2006)
6. O'Reilly, T.: What is Web 2.0, Consulted December 2011 in http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html, (2005)
7. Palloff, R.M., Pratt, K.: Building learning communities in cyberspace - effective strategies for the on-line classroom, Consulted December 2011 in http://www.online2.org/wksp-projects/NCCE2004/resources-originals/bldg-learning-communities-cyberspace-notes.doc (1999)
8. C., Ullrich, C., Borau, K., Luo, H., Tan, X., Shen, L., Shen, R.: Why web 2.0 is good for learning and for research: principles and prototypes, In 17th international conference on World Wide Web, Beijing, China (April-2008)
9. Liyoshi, T., Kumar, M.: Opening up education: The collective advancement of education through open technology, open content, and open knowledge, In MA-MIT Press, Cambridge (2008).
10. Carvalho, A.: Indicadores de Qualidade de Sites Educativos. In CRIE. Avaliação de Locais Virtuais de Conteúdo Educativo - Cadernos SACAUSEF II, Lisboa (2006)
11. Attwell, G.: Web 2.0 and the changing ways we are using computers for learning: what are the implications for pedagogy and curriculum?. In Elearningeuropa, (2007)
12. Costa, F.: Competências TIC. Estudo de Implementação, Vol. I: GEPE - Ministério da Educação, Lisboa (2008)
13. Rodríguez, J.: (2007): Como as comunidades virtuais de prática e de aprendizagem podem transformar a nossa concepção de educação, In Sísifo - Revista de Ciências da Educação, 03, pp.117-124, (2007)
14. Rodríguez, J.: Comunidades Virtuales de Práctica y de Aprendizaje, In Publicacions y Edicions Universidad de Barcelona, Barcelona (2008)

Poliontо: Ontology reuse with Automatic Text Extraction from Political Documents

Adela Ortiz
PRODEI-The Doctoral Program in Informatics Engineering, FEUP- Faculty of Engineering of the University of Porto, Rua Dr. Roberto Frias, 4200-465, Porto, Portugal
pro10024@fe.up.pt

Abstract. This article presents a research work to analyze the development of domain ontologies through automatic text analysis. The goal is to create ontology in the political domain (Poliontо). The work progressed through two vectors: one concerned an effort to achieve automatic extraction of ontological terms and their relations through the use of specialized text analysis software. Second, we propose use ontology reengineering methods for the integration of the Portuguese ontology obtained by Text2onto with the English ontology model (government.owl). The article describes some steps taken to accomplish the Political domain ontology, presents and justifies the information extraction results with Text2onto. Therefore it provides a presentation of the reengineering methodology proposed to create a new ontology in the political domain.

Keywords: ontology, information extraction, political domain, ontology reuse

1 Introduction

One of the concepts that support the ideas and research behind web semantics is that of ontology. It is not the purpose of this article to make a detailed approach and description of what ontologies are but we can take the words of Guarino [1], for whom domain ontology is "a way to conceptualize explicitly and formally the concepts and constraints related to a domain of interest".
Information extraction (IE) is a key component in the process of ontology building. Given the complexity of the methodologies for ontology creation, the automatic extraction of information can provide an important leverage for an effective creation process.
The reuse process focuses on a method for ontology reengineering which attempts to capture the conceptual model of the implemented source ontologies in order to transform them into a new, more correct and more complete ontology.
Actually, we motivate to develop a new ontology in the political domain. The main objective of this paper is to describe the resources for the development of a political ontology.

One of the crucial issues in our ontology development requires the reuse of ontologies. Consequently, the need to integrate political documents thought automatic text extraction method. Before, we need adapt both ontologies to a Portuguese language and culture by ontology localization method. We propose include multilingual info in the ontology model with the Neon Label translator.
The information extractions process, we used a specific group of algorithms to extract various types of relationships. To form the corpus we used documents related to Political domain. Ontology reuse has been seen as a viable alternative to having an ontology model in the political domain. For the discovery of ontologies, we start by finding a government ontology only English language. We decided to analyze existing domain ontology proposed by DAML repository (government.owl). We start by finding with all the desired concepts, and then it is necessary to recognize and compare the same concepts in accordance to those classes that were considered important from the automatic text extraction of our political documents.

2 Related Work

Many researches have proposed the idea of creating ontologies from automatic text extraction.
For that purpose, Text2Onto was selected because it correctly and effectively answered the most important requirements such as: automatic extraction, usability, scalability, accessibility, interoperability and reusability [2].
In the process of drafting the ontology is used the Protégé tool, developed by Stanford University.
Our ontology reuse methodology is inspired from the papers [3] and the specific part of ontology localization by [4].
There are two more ontologies in the political domain as taxonomy example [5], [6], but these ontologies are a reference with few concepts and relationships for our work. We refer here to one that looks most relevant to our work (government.owl), to wich we present along this paper [7].
However, their classes and relationships are still somehow different our approach, so we need to made a comparison and selection more deeply. There are not approaches to ontology political domain have recently published. Actually, there are not ontologies in the political domain in Portuguese.
We found ontology (government.owl) in English with the goal of reusing it. We propose use ontology reengineering methods for the integration of the Portuguese ontology obtained by Text2onto with some concepts and relationships of English ontology model (government.owl).

3 Methodology

3.1 High Level

The reengineering methodology proposed consists of three steps:
Reengineering: We derive a possible conceptual model with taxonomic structure, followed by relations between concepts and instances.
Correcting: The objective of this step is to evaluate the concepts of ontology model with the best ontology extracted with text2onto, correct the detected errors, and refine the ontology model in conformity with the requirements of the new Poliontо. Both ontologies are used to build a new one, but with modifications.
Translating: Include multilingual info in the revised conceptual model with the Neon Label translator [2]. We try to find semi-automatically relevant synonyms in the field under analysis and we repeat the process again.
Merging: The ontology extracted from political documents is integrated on the basis of the revised conceptual model, resulting in a final ontology that is then assigned as Poliontо by aligning the domain knowledge.

3.2 Details

In this section we describe with detail each of the steps taken for the design of the domain ontology

3.2.1 Determination of area and scope of ontology

The domain used in this ontology refers to the representation of a political model. Through an analysis of the scientific field it was possible to identify the key terms on this context.
In this section, we display maps of words and semantic relationships where we can explore concepts and export lists of concepts or bag of words in specific applications of concept maps.
We used search engines and knowledge base to identify other ontologies in the political domain.

- *Ontology repository* – It maintains a local copy of ontologies, and their different versions, if they exist. This kind of tools usually only provide browse functionalities. For the Political domain, we found good results in two repositories:
  - The IHMC Public Ontology Server

COE provides a simple text search through archived Concept Maps in the local files and the IHMC Public Ontology Server [8]. Also, COE uses concept maps to display, edit and compose OWL. Concepts may be used to identify and retrieve related ontologies, and be incorporated into new Cmaps. This application published and

exported the concept maps (. cmap) to various formats of ontology OWL / RDF,RDF / XML, Turtle and N3.

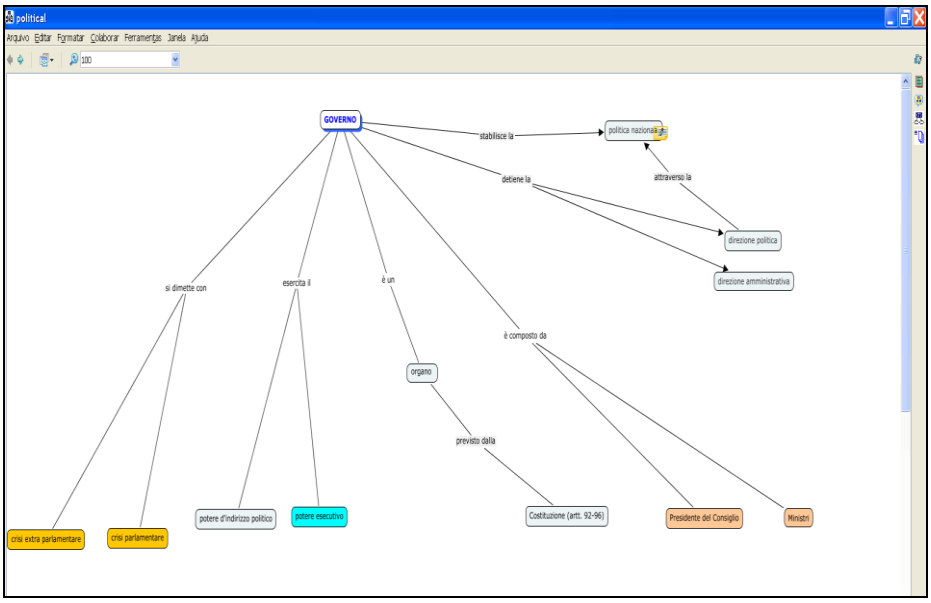


Fig. 1. Political Ontology of IHMC Public Ontology Server

Another useful vision example of our experiences with COE Cmap tool in the case of resource text2onto exporting its output in formats such as OWL ontologies. Subsequently, the ontology can be exported into COE Cmap later in order to be worked as a conceptual map.

Table 1. Concept Search Engine of Political domain with COE Cmap.

Experiences	Description	Process
#1	A domain ontology is created from the concept maps available through the detection of ontological concepts and relationships. In this method a conceptual map is transformed into an OWL ontology	Models studied and collected various maps, new concepts by extracting information from documents (corpus) of the political domain.
#2	A domain concept map is created from the OWL ontology In this method, the OWL ontology is transformed into a concept map.	We use our prototype ontology representing the political domain format to be transformed into OWL.

DAML ontology repository [9] has more than 160 ontologies implemented in DAML+OIL language. The formats (DAML OIL, OWL) are extensions of RDF. In this repository, we found the government.owl ontology.
Ontology search engine as Swoogle [10], Watson [11], Sindice [12] and Falcons [13]. We use these kinds of tools facilitate the findability of suitable ontologies, but they differ in the way they describe ontologies. The problem was to found the best ontology from few available in the political domain.

3.2.2 Automatic text extraction from political documents

The objective of the generation of ontology to extract automatically or semi-automatically the relevant concepts and relationships from a corpus determined to form an ontology.
Information extraction (IE) uses texts as input to output structured data, which could be named entities, such as organizations, people, keywords, or relations among them. To form the corpus we used documents in several formats like PDF, HTML and plain text. All these documents related to the specific domain addressed, that is, Politics. A collection of texts in electronic form (corpus) were used for natural language processing (NLP). The corpus in this case was formed through the use of texts in formats like pdf., txt and html. With this corpus different tests and extractions are performed. The results differ when single documents or multiple documents are used. Results are also highly influenced by the choice of certain algorithms. The extraction of concepts and instances is based on statistical approach. Different term weighting measures are used to compute how important a word is to a document in a collection or corpus how important a word is to a document in a collection or corpus. Text2Onto implements various measures to assess the relevance of a particular term with respect to the body in question.
The four algorithms in the class tab of Text2onto are [4],[14]:

- Entropy Concept Extraction: Entropy and the C-Value/NC-value method. C-value and NC-value method combines linguistic and statistical information. C-value enhances the common statistical measure of frequency of occurrence for term extraction, making it sensitive to a particular type of multi-word terms. NC-value: extraction of term context words (words that tend to appear with terms).
- Example Concept Extraction: This is based on candidate groups of characteristic terms for designated category.
- RTFConceptExtraction: Relative Term Frequency. The ranking by the frequency of terms use a TF (Term frequency) as a measure for determinate attribute, to give more classification for the attribute with more frequency in the corpus.
- TFIDFConceptExtraction: TFIDF (term frequency–inverse document frequency) is a weight (statistical measure) used to evaluate how important a word is to a document in a collection or corpus. TF-IDF weighting: give higher weight to terms that are rare. TF: Term frequency (increases weight of frequent terms). If term is frequent in lots of documents it does of frequent is term not has discriminative power. IDF: inverse document frequency.

Text2Onto uses the method of capture based on ontology learning to changes in the lists of concepts that are reflected by the application. The most frequent terms are analyzed in order to scope of main themes. This method proves useful for a first approach of the data extraction based on NLP methods. We used a specific group of algorithms to extract various types of relationships, like Instance-of, Subclass-of, Part-of and other general relations. We ran four series of experiments in order to calculate the average of Text2onto for each 100 classes extracted with each algorithm used. The four tests were calculated through the following statistical measures: Entropy Concept Extraction, Example Concept Extraction, RTFConceptExtraction and TFIDFConceptExtraction.
We selected Portuguese classes such as *"governo"*, *"legislativa"*, *"presidente"*, *"partido"*, and other concepts fundamental to the ontology of the Politics. After, we will specialize the activities of other concepts.

3.3. Selecting Ontology for Reuse

We gave particular attention to the hierarchical classification of ontology model (government.owl) as a taxonomy that would later lead to the ontology. The field analysis (Figure 1 and 2) made it possible to identify the relevant key terms in the context developed by Stanford University, specifically we used the components of the ontologies editor in the OWL extension.
In this section we obtained ontology in the DAML ontology library. We evaluate this ontology, which, based on the ranking we consider their popularity, concepts coverage and knowledge richness. This ontology had 84 classes, 78 object properties, 3 data properties, 53 subclasses axioms count. The ontology (government.owl) is reused for the purpose of developing ontology for Politic.
The terms that are part of the "Political Organization", such as "Political Party", for example, are defined in the ontology as subclasses of this super class. An "organization" can be represented by a "Government Organization", "International Organization", "Judicial International Organization" and "Legislative Organization". In Figure 2, the left side of this image, the ontology can be seen as a set of terms of taxonomy, structured as a tree.
Figure 2 shows the first fragment of the ontology model with the tool Protégé OWL. This graph was generated using the OntoViz Tab in Protégé [15]. This tool allows visualizing all the ontologies with sophisticated graph visualization software called Graphviz-Tool Dot from AT&T. We pick a set of basic classes such as: Government and Organization.

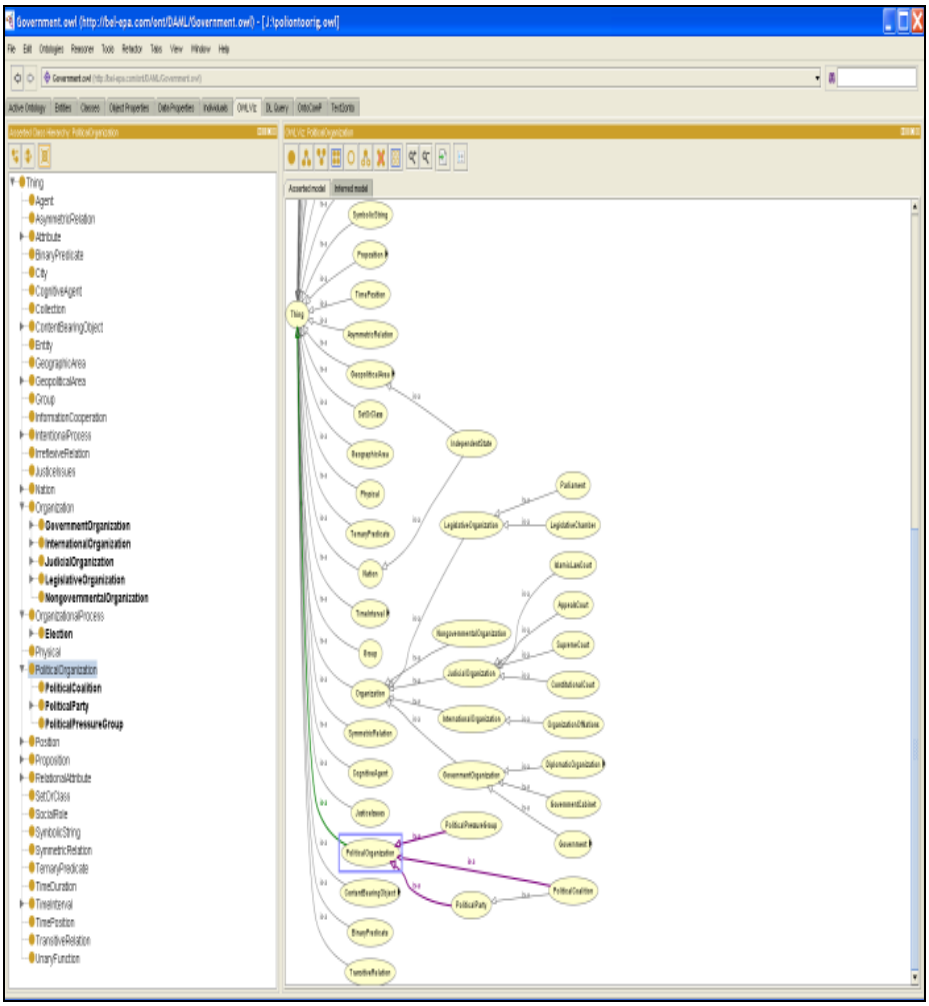


Fig. 2. Ontology for reuse (Goverment.owl).

3.4 Evaluating the results of Automatic Text Extraction from Political Documents

The most frequent terms are analyzed in order to scope of main themes. This method proves useful for a first approach of the data extraction based on NLP methods. We used a specific group of algorithms to extract various types of relationships, like Instance-of, Subclass-of, Part-of and other general relations.

We ran four series of experiments in order to calculate the average of Text2onto for each 100 classes extracted with each algorithm used. The four tests were calculated through the following statistical measures: Entropy Concept Extraction, Example Concept Extraction, RTFConceptExtraction and TFIDFConceptExtraction [14]. Our

experiments show that the best result that values the tool in its tests was an F-Measure of 30%, for the classes when classified with respect to ontology includes 100 concepts (Table I). Each line represents an F-measure in percent of good political concepts for each 10 concepts extracted from political documents corpus.

Table 2. Results of the analysis of automatic text extraction from political documents

Entropy Concept Extraction	Example Concept Extraction	RTF Concept Extraction	TFIDF Concept Extraction
0,00%	4,00%	3,00%	4,00%
1,00%	2,00%	4,00%	4,00%
2,00%	2,00%	2,00%	5,00%
2,00%	2,00%	2,00%	3,00%
1,00%	3,00%	3,00%	3,00%
2,00%	3,00%	4,00%	3,00%
2,00%	2,00%	2,00%	3,00%
3,00%	1,00%	2,00%	2,00%
2,00%	2,00%	2,00%	3,00%
2,00%	0,00%	1,00%	0,00%
17,00%	21,00%	25,00%	30,00%

The following graphic shows the results for each algorithm used in each group of 10 concepts analyzed in the political domain.

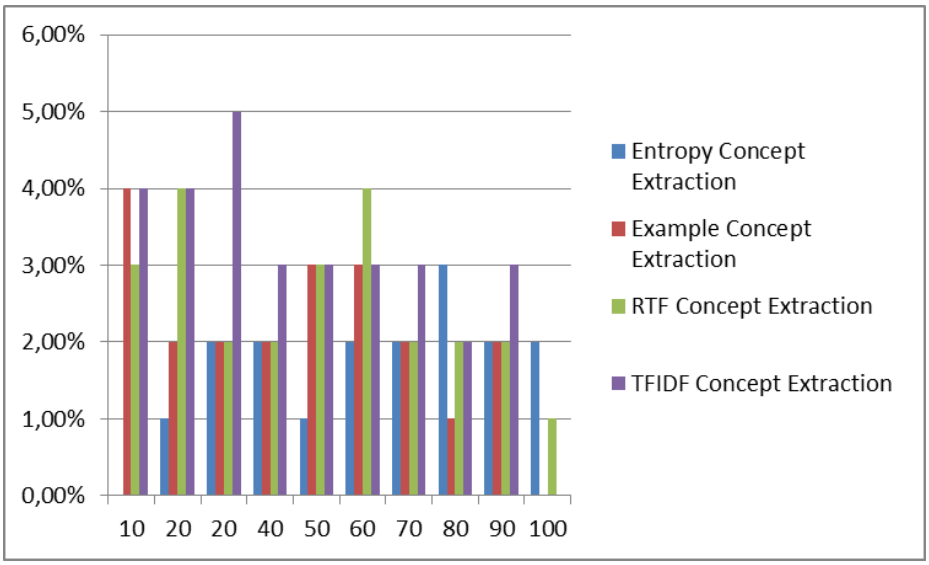


Fig. 3. Percent of good concepts found for four algorithms in each 10 concepts in ontology generated.

We collect political news stories from Portugal. The new stories ranged from 692971. Each of one of the news stories is short with, on average, 221 Portuguese characters. From political domain, we choose the first 11357 news stories as our training set and the following 2974 news stories as our testing set.

For example, if the new story about politic contained the next sentence "*Cavaco Silva optou também por não fazer comentários à greve nacional dos docentes".* The article should be classified in the Politic domain.

We know that President "*Cavaco Silva"* is a politician and that most newsworthy people regularly appear in a specific domain. The micro-average of the Text2onto 30% is acceptable for each 100 concepts extracted. We found that there are many new politicians' names in the news.

We collected 35 politicians' names and found associated classes in the first part of our test. Further experiments using the ontology extraction and other ontology methods are still in progress.

As an example, this section presents some results from the same source, using different algorithms. The Probabilistic Object Model (POM) shows the concept tab for the classes (Figure 4 and 5).

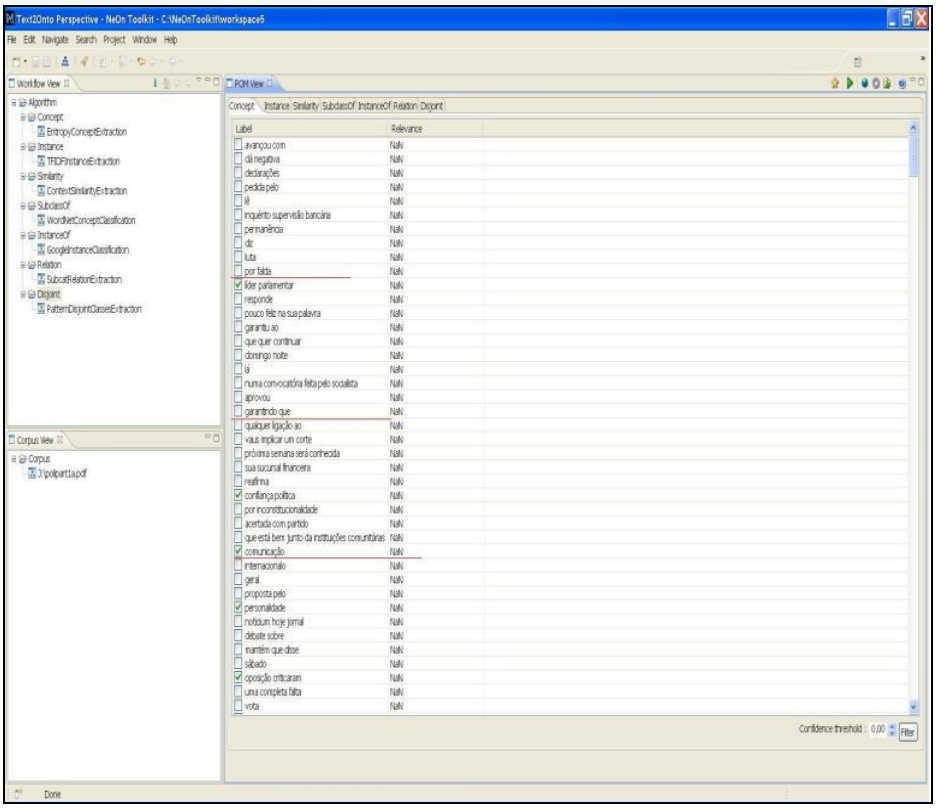


Fig. 4. Some results obtained with the Entropy Concept Extraction algorithm (Test 1

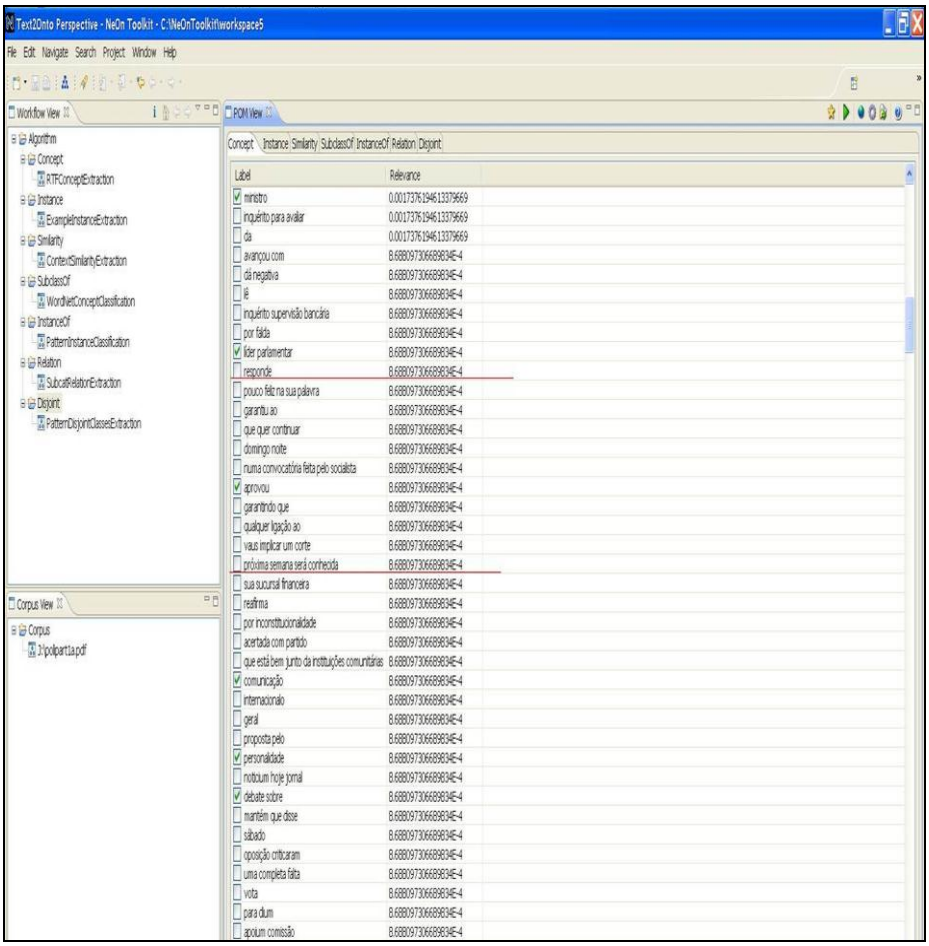


Fig. 5. Some results obtained with the RTFconceptExtraction algorithm (Test 3)

Acknowledgments. This paper is part of my work in the course of Language Processing and Information Extraction in the Doctoral Program in Informatics Engineering. The Prof. Luis Sarmento has been teaching new ideas.

5 Conclusions

The focus of the work demonstrated the applicability of the reengineering approach with the automatic text extraction from Political documents. Thus, the reported results provide evidence on the concepts and relationships expected. We concentrated on the presentation of the ontology reuse methodology. We admitted the complexity of the

ontological sources employed, and the need for automatic means. The experiment was restricted to political ontology domain containing a manageable number of at most several hundred of concepts in this area. We used text2onto to calculate sequences of concepts, for analyzing frequency of political concepts in the document for each one of four algorithms available to this tool, some interesting findings. In this paper we evaluate in our experiments that the TFIDF Concept Extraction algorithm of Text2onto can extract more concepts in the political domain than the others three algorithms evaluated. We proposed reengineering workflow can be executed manually and automatically.

References

1. Guarino, N., Formal Ontology in Information Systems, Proceedings of FOIS'98, Trento, Italy, 6-8 June, IOS Press (1998), pp. 3-15.
2. Espinoza M., Montiel-Ponsoda E., Corcho O., Aguado G., Gómez-Pérez A., Work on Multilingual Ontologies within the NeOn Project, C21 Cost Action Towntology (2008)
3. Gruber T. R., A translation approach to portable ontologies, Knowledge Acquisition (1993), Vol. 5, no. 2, pp. 199-220.
4. Cimiano P., Völker J., A Framework for Ontology Learning and Data-driven Change Discovery, Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB), Lecture Notes in Computer Science, Springer (2005), Vol. 3513, pp. 227-238.
5. D'aquin M., Sabou M., Motta E., Reusing Knowledge from the Semantic Web with the Watson Plugin, 7th International Semantic Web Conference (ISWC2008), Karlsruhe, Germany (2008).
6. Wikipedia, Ontology Learning, http://en.wikipedia.org/wiki/Ontology_learning
7. P. Hayes P., Eskridge T. C., Saavedra R., Reichherzer T., Mehrotra M., Bobrovnikoff D., Collaborative Knowledge Capture in Ontologies, In: K-CAP 05: 2005, Banff, Canada (2005), pp. 99-106.
8. COE Cmap tool, http://www.ihmc.us/groups/coe/
9. DAML Ontology repository, http://www.daml.org/ontologies
10. Swoogle, http://swoogle.umbc.edu/
11. Watson, http://kmi-web05.open.ac.uk/WatsonWUI/
12. Sindice, http://sindice.com/
13. Falcons, http://iws.seu.edu.cn/services/falcons/objectsearch/index.jsp).
14. Haase, P., Völker J., Ontology Learning and Reasoning: Dealing with Uncertainty and Inconsistency, Uncertainty Reasoning for the Semantic Web I: ISWC International Workshop (2008), p.p 366-384.
15. Protégé, http://protegewiki.stanford.edu/wiki/Protege_Plugin_Library
16. Cuel R., Cristani M., A Survey on Ontology Creation Methodologies, Int. J. Semantic Web Inf. Syst. (2005), Vol. 1, no. 2, pp. 49-69.
17. Kremer R., Concept Mapping: Informal to Formal, Proceedings of the Third International Conference on Conceptual Structures, Knowledge Acquisition Using

Conceptual Graphs Theory Workshop, Tepfenhart, W., Dick, J. & Sowa, J. (Eds.) (1998), University of Maryland College Park, MD, pp. 152-167.

18. Garcia A., Noreña A., Betancourt A., Garcia L., CMAPS Supporting the Development of OWL Ontologies, Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC2008) , Karlsruhe, Germany, 2008.
19. Espinoza M., Montiel-Ponsoda E., Corcho O., G. Aguado G., Gómez-Pérez A., Work on Multilingual Ontologies within the NeOn Project, C21 Cost Action Towntology October 20th, 2008
20. Nédellec C., Golik W., Aubin S., Bossy R., Building Large Lexicalizedfrom Text: A Use Case in automatic Indexing of Biothechnology Patents, Proceeding of EKAW (2010) Knowledge Engineering and Management by The Masses. 17th International Conference, Lisbon Portugal.

PAPERS IN ALPHABETICAL ORDER

AUTHORS IN ALPHABETICAL ORDER

Proceedings of the 6th Doctoral Symposium in Informatics Engineering

This Symposium is part of the course in Methodologies for Scientific Research integrated in the Doctoral Programme in Informatics Engineering (ProDEI) and it aims to provide doctoral students a platform to present their future research proposals, share their ongoing work, discuss their ideas and collect feedback from peers, professors and experienced researchers.

With this Symposium students improve their practice in writing and presenting scientific papers and knowledge in methodologies for scientific research as well as can receive advice and find new research directions on the areas of Informatics Engineering and Computers Sciences. This edition of the Symposium included the following major topics:

- Artificial Intelligence
- Computer Graphics
- Data Processing
- e-Learning
- Information Systems
- Mobile Computing
- Networking
- Reconfigurable Computing
- Software Engineering
- Semantics Web and Evolution

ISBN 978-972-752-129-6
9789727521296