

Smart Innovation, Systems and Technologies 36

George A. Tsihrintzis

Maria Virvou

Lakhmi C. Jain

Robert J. Howlett

Toyohide Watanabe *Editors*



Intelligent Interactive Multimedia Systems and Services in Practice

The logo for KES International, featuring the letters 'KES' in a stylized blue font above the word 'International' in a smaller blue font.

The Springer logo, which is a stylized chess knight icon, followed by the word 'Springer' in a serif font.

Smart Innovation, Systems and Technologies

Volume 36

Series editors

Robert J. Howlett, KES International, Shoreham-by-Sea, UK
e-mail: rjhowlett@kesinternational.org

Lakhmi C. Jain, University of Canberra, Canberra, Australia, and
University of South Australia, Adelaide, Australia
e-mail: Lakhmi.jain@unisa.edu.au

About this Series

The Smart Innovation, Systems and Technologies book series encompasses the topics of knowledge, intelligence, innovation and sustainability. The aim of the series is to make available a platform for the publication of books on all aspects of single and multi-disciplinary research on these themes in order to make the latest results available in a readily-accessible form. Volumes on interdisciplinary research combining two or more of these areas is particularly sought.

The series covers systems and paradigms that employ knowledge and intelligence in a broad sense. Its scope is systems having embedded knowledge and intelligence, which may be applied to the solution of world problems in industry, the environment and the community. It also focusses on the knowledge-transfer methodologies and innovation strategies employed to make this happen effectively. The combination of intelligent systems tools and a broad range of applications introduces a need for a synergy of disciplines from science, technology, business and the humanities. The series will include conference proceedings, edited collections, monographs, handbooks, reference books, and other relevant types of book in areas of science and technology where smart systems and technologies can offer innovative solutions.

High quality content is an essential feature for all book proposals accepted for the series. It is expected that editors of all accepted volumes will ensure that contributions are subjected to an appropriate level of reviewing process and adhere to KES quality principles.

More information about this series at <http://www.springer.com/series/8767>

George A. Tsihrintzis · Maria Virvou
Lakhmi C. Jain · Robert J. Howlett
Toyohide Watanabe
Editors

Intelligent Interactive Multimedia Systems and Services in Practice

 Springer

Editors

George A. Tsihrintzis
Department of Informatics
University of Piraeus
Piraeus
Greece

Robert J. Howlett
KES International
Shoreham-by-Sea
UK

Maria Virvou
Department of Informatics
University of Piraeus
Piraeus
Greece

Toyohide Watanabe
Nagoya University
Nagoya
Japan

Lakhmi C. Jain
Faculty of Education, Science,
Technology and Mathematics
University of Canberra
Canberra
Australia

ISSN 2190-3018

ISSN 2190-3026 (electronic)

Smart Innovation, Systems and Technologies

ISBN 978-3-319-17743-4

ISBN 978-3-319-17744-1 (eBook)

DOI 10.1007/978-3-319-17744-1

Library of Congress Control Number: 2015936376

Springer Cham Heidelberg New York Dordrecht London

© Springer International Publishing Switzerland 2015

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media
(www.springer.com)

Preface

The term *Multimedia Services* is used when referring to services that make use of coordinated and secure storage, processing, transmission, and retrieval of information which exists in various forms. From its nature, the term *Multimedia Services* includes several levels of data processing and applies to such diverse areas as digital libraries, e-learning, e-government, e-commerce, e-entertainment, e-health, and e-legal services, as well as to their mobile counterparts (i.e., *m-services*). Our society is characterized by a constant demand for new and more sophisticated multimedia services. This demand constantly poses new challenges for advanced processing at all levels.

Over the past few years, an attempt was made to follow recent advances in a series of edited books on multimedia services in intelligent environments. The volume at hand is the sixth volume on the topic. In our previous books [1–5], we covered various aspects of processing in multimedia services.

More specifically, in [1] we were concerned mostly with low level data processing in multimedia services in intelligent environments. In [2], we were concerned with various software development challenges and related solutions that arise when attempting to accommodate multimedia services in intelligent environments. In [3], we presented various integrated systems that were developed to accommodate multimedia services in intelligent environments. In [4], we were concerned with a special class of software systems, known as *Recommender Systems*. Finally, in [5], a specific class of multimedia services was addressed and presented, which are called *Recommendation Services*.

In the current volume, we present some specific multimedia systems that have been developed and applied *in practice*. Such multimedia systems are important not only in their own, but also in that they highlight various difficulties that may arise when developing practical multimedia and, thus, provide practical solutions applicable to other multimedia development projects as well.

More specifically, the book at hand consists of an editorial and seven chapters. All chapters in the book have evolved as extended versions of selected papers presented in the *6th International Conference on Intelligent Interactive Multimedia*

Systems and Services (KES IIMSS 2013), Sesimbra, Portugal, June 26–28, 2013. The chapters in the book are as follows:

Chapter 1 introduces the research area and presents a brief abstract of each chapter included in the book. Chapter 2 is on “On the Use of Multi-attribute Decision Making for Combining Audio-Lingual and Visual-Facial Modalities in Emotion Recognition.” Chapter 3 is on “Cooperative Learning Assisted by Automatic Classification within Social Networking Services.” Chapter 4 is on “Improving Peer-to-Peer Communication in e-Learning by Development of an Advanced Messaging System.” Chapter 5 is on “Fuzzy-Based Digital Video Stabilization in Static Scenes.” Chapter 6 is on “Development of Architecture, Information Archive and Multimedia Formats for Digital e-Libraries.” Finally, Chapter 7 is on “Layered Ontological Image for Intelligent Interaction to Extend User Capabilities on Multimedia Systems in a Folksonomy Driven Environment.”

As societal demand drives multimedia services in intelligent environments to become increasingly more sophisticated, as new challenges arise which require ever more efficient tools, methodologies, and integrated systems to be devised, and as the application areas of multimedia services continue to expand at an explosive rate, the entire field of multimedia services in intelligent environments cannot be effectively covered in the six volumes published so far. Thus, it may be expected that additional volumes on other aspects of multimedia services in intelligent environments will appear in the future.

We are grateful to the authors and reviewers for their contributions. Thank are due to the Springer for their assistance during the evolution phase of this book.

Greece	George A. Tsihrintzis
Greece	Maria Virvou
Australia	Lakhmi C. Jain
UK	Robert J. Howlett
Japan	Toyohide Watanabe

References

1. Tsihrintzis, G.A., Jain, L.C. (eds.): *Multimedia Services in Intelligent Environments—Advanced Tools and Methodologies*. Studies in Computational Intelligence Book Series, vol. 120. Springer, Berlin (2008)
2. Tsihrintzis, G.A., Virvou, M., Jain, L.C. (eds.): *Multimedia Services in Intelligent Environments—Software Development Challenges and Solutions*. Smart Innovation, Systems, and Technologies Book Series, vol. 2. Springer, Berlin (2010)
3. Tsihrintzis, G.A., Jain, L.C. (eds.): *Multimedia Services in Intelligent Environments—Integrated Systems*. Smart Innovation, Systems, and Technologies Book Series, vol. 3. Springer, Berlin (2010)
4. Tsihrintzis, G.A., Virvou, M., Jain, L.C. (eds.): *Multimedia Services in Intelligent Environments—Advances in Recommender Systems*. Smart Innovation, Systems and Technologies Book Series, vol. 24. Springer, Berlin (2013)
5. Tsihrintzis, G.A., Virvou, M., Jain, L.C. (eds.): *Multimedia Services in Intelligent Environments—Recommendation Services*. Smart Innovation, Systems and Technologies Book Series, vol. 25. Springer, Berlin (2013)

Contents

1 Intelligent Interactive Multimedia Systems in Practice:	
An Introduction	1
1.1 Introduction	1
1.2 Chapters Included in the Book	2
1.3 Conclusion	4
Resources	4
KES International	4
KES Conference Series	5
KES Journals	5
Book Series	5
References	6
2 On the Use of Multi-attribute Decision Making for Combining Audio-Lingual and Visual-Facial Modalities in Emotion Recognition	7
2.1 Introduction	8
2.2 Related Work	9
2.2.1 Multi-attribute Decision Making	11
2.3 Aims and Settings of the Empirical Studies	12
2.3.1 Elicitation of Emotions and Creation of Databases	13
2.3.2 Creation of Databases of Known Expressions of Emotions	16
2.3.3 Analysis of Recognisability of Emotions by Human Observers	17
2.4 Empirical Study for Audio-Lingual Emotion Recognition	17
2.4.1 The Experimental Educational Application for Elicitation of Emotions	17
2.4.2 Audio-Lingual Modality Analysis	17
2.5 Empirical Study for Visual-Facial Emotion Recognition	20
2.5.1 Visual-Facial Empirical Study on Subjects	20
2.5.2 Visual-Facial Empirical Study by Human Observers	21

2.6	Discussion and Comparison of the Results from the Empirical Studies.	24
2.7	Combining the Results from the Empirical Studies Through MADM	27
2.8	Discussion and Conclusions.	32
	References.	32
3	Cooperative Learning Assisted by Automatic Classification Within Social Networking Services	35
3.1	Introduction	35
3.2	Related Work	37
	3.2.1 Social Networking Services	37
	3.2.2 Intelligent Computer-Assisted Language Learning	38
3.3	Algorithm of the System Functioning	39
	3.3.1 Description of Automatic Classification	39
	3.3.2 Optimization Objective and Its Definition	39
	3.3.3 Initialization of Centroids.	40
	3.3.4 Incorporation of Automatic Classification	40
3.4	General Overview of the System	42
3.5	Evaluation of the System	44
3.6	Conclusions and Future Work	46
	References.	46
4	Improving Peer-to-Peer Communication in e-Learning by Development of an Advanced Messaging System	49
4.1	Introduction	50
4.2	Related Work	51
4.3	Data Analysis System Design	52
4.4	Experimental Results	56
4.5	Conclusions and Future Work	60
	References.	61
5	Fuzzy-Based Digital Video Stabilization in Static Scenes	63
5.1	Introduction	63
5.2	Related Work	64
5.3	Method of Frame Deblurring.	66
5.4	Fuzzy-Based Video Stabilization Method	68
	5.4.1 Estimation of Local Motion Vectors.	69
	5.4.2 Smoothness of GMVs Building	72
	5.4.3 Static Scene Alignment	73
5.5	Experimental Results	74
5.6	Conclusion	81
	References.	82

6	Development of Architecture, Information Archive and Multimedia Formats for Digital e-Libraries	85
6.1	Introduction	85
6.2	Related Work	86
6.3	Overview of Standards and Document Formats.	88
6.4	Requirements and Objectives.	91
6.5	Proposed Architecture of Digital e-Library Warehouse	92
6.6	Proposed EPUB Format Extensions.	93
6.7	Client Software Design and Researches of Vulnerability	96
6.8	Conclusion	101
	References	102
7	Layered Ontological Image for Intelligent Interaction to Extend User Capabilities on Multimedia Systems in a Folksonomy Driven Environment	103
7.1	Introduction	103
7.2	Human Based Computation	104
	7.2.1 Motivation of Human Contribution	104
7.3	Background of Related Work.	105
	7.3.1 Object Tracking	106
7.4	Dynamic Learning Ontology Structure	107
	7.4.1 Richer Semantics of Attributes.	107
	7.4.2 Object on Layered Representation	108
	7.4.3 Semantic Attributes	110
	7.4.4 Attribute Bounding Box Position.	111
	7.4.5 Attributes Extraction and Sentiment Analysis	111
	7.4.6 Folksodriven Bounding Box Notation	112
7.5	Image Analysis and Feature Selection	112
	7.5.1 Object Position Detection.	113
7.6	Previsions on Ontology Structure.	114
7.7	A Case Study: In-Video Advertisement	115
	7.7.1 In-Video Advertisement Functionality.	116
	7.7.2 Web GRP	116
	7.7.3 Folksodriven Ontology Prediction for Advertisement	118
	7.7.4 In-Video Advertisement Validation	119
7.8	Relevant Resources	120
7.9	Conclusion	120
	References	121

Chapter 1

Intelligent Interactive Multimedia Systems in Practice: An Introduction

George A. Tsihrintzis, Maria Virvou, Margarita N. Favorskaya and Lakhmi C. Jain

Abstract This chapter presents a brief description of chapters devoted to researches in practical implementation of intelligent interactive multimedia systems and services. The issues of interactive human-computer communication potentially influence many aspects of life that requires the well-thought-out and circumspective use of multimedia systems and services.

Keywords Intelligent systems · Interactive multimedia · Social services · Interactive communication

1.1 Introduction

Nowadays, interactive multimedia systems are a reality that affect many aspects of human life. The development of intelligent component is a necessity for the implementation of theoretical aspects into practical systems, which people encounter

G.A. Tsihrintzis (✉) · M. Virvou
Department of Informatics, University of Piraeus, Piraeus, Greece
e-mail: ctrouss@unipi.gr

M. Virvou
e-mail: mvirvou@unipi.gr

M.N. Favorskaya
Department of Informatics and Computer Techniques, Siberian State Aerospace University,
Krasnoyarsky Kray, Russia
e-mail: favorskaya@sibsau.ru

L.C. Jain
University of Canberra, Canberra, Australia
e-mail: Lakhmi.jain@unisa.edu.au

L.C. Jain
University of South Australia, Adelaide, South Australia, Australia

daily. Many interesting practical decisions one can find in this book including multi-attribute decision making system, improvement of social networking services, Tesys environment, system for digital video stabilization, digital e-library warehouse, and Folksodriven system. The majority of efforts focus on designing intelligent systems, which can be directly be applied in many practical situations.

1.2 Chapters Included in the Book

The main purpose of this book is to present the research results on intelligent interactive multimedia systems and services implemented in practical applications. The book includes seven chapters including this introductory chapter.

Chapter 2 introduces the integration of audio-lingual and visual-facial modalities in a bi-modal user interface for affect recognition. The recognition of emotions leads to affective user interfaces that can adapt users' behavior [1]. The research is built on the fact that human faces, people's voices, or people's actions show their emotions. In this chapter, the authors focus on the combination of the audio-lingual modality and the visual-facial modality by employing Multi-Attribute Decision Making (MADM). The MADM involves the making preference decisions (such as evaluation, prioritization, and selection) over the available alternatives that are characterized by multiple, usually conflicting attributes. The information from two modalities is combined using a multi-criteria decision theory [2, 3]. Each modality is considered as a criterion that a human observer would use in order to recognize another human's emotion. Through typical human-computer interaction session, the facial expressions corresponding to the "neutral", "happiness", "sadness", "surprise", "anger", and "disgust" emotional states were recognized. Also for the Greek language that can be used in the audio-lingual modality and a database of facial expressions of emotions, the basic affective vocabulary was constructed.

Chapter 3 presents a discussion about Social Networking Services (SNSs) such as Facebook as a potentially potent educational method for students, who spend a lot of time on these online networking activities [4, 5]. Many researchers investigated the use of SNSs in the context of learning, specifically language learning and proposed the exploitation of machine learning techniques to improve and adapt the set of user model stereotypes [6]. This chapter involves the automatic classification for cooperative learning among Facebook users. The automatic classification achieves to efficiently create user clusters. The proposed system recommends cooperation between user clusters in a way that both of them gain knowledge and promote their learning as opposed to helping only a specific user cluster to the detriment of others [7]. The approach presented in this chapter exploits the fact that Facebook has a large number of users, and automatic reasoning mechanisms based on recognized similarities among them can be used.

Chapter 4 studies the peer-to-peer communication in e-Learning as a model of recommended messages interesting for groups of users [8]. A problem of

improving peer-to-peer communication is tackled among learners by means of educational data mining, user modeling, and text mining [9, 10]. The chosen educational environment is Tesys, which is an online educational platform where students, professors and secretaries perform their responsibilities. The system support server side and client side processes. The server side is implemented in three modules: the data gathering module, the model building module, the message indexing module and the student classification and message retrieval module. The key idea is that text-mining and concept detection processes are run in the corpus of all messages in order to select a proper message indexing process into the argumentative classifier [11]. The messages exchanged between students and professors are indexed into a tree like structure (i.e., the classifier) by linking messages to student classes or labels. The designed system makes intensive use of text mining and stemming capabilities that are put together into the concept detection and question difficulty estimator tools.

Chapter 5 investigates development of digital video stabilization method based on the Takagi-Sugeno-Kang model for improvement of motion vectors clustering [12] and the application of scene alignment procedure into static scenes. The unintentional video camera motion provides the negative influence on following segmentation and recognition procedures in video sequences in various applications such as video surveillance, robotics, video encoding, etc. [13, 14]. The “IF-THEN” fuzzy rules were proposed in order to improve the selection of detected Local Motion Vectors (LMVs) in a scene [15]. Fast procedure of LMVs calculation was proposed permitting to avoid the challenges of luminance changing or moving foreground objects. Also the deblurring method to find and improve the blurred frames was developed. All methods and algorithms were realized by the designed software tool, which provides better calculation results in comparison with other software products.

Chapter 6 provides the advanced facilities for customers, who use stationary and non-stationary working places, by use Digital e-Library Warehouse (DLW) [16, 17] and library information archive. The chapter involves a description of library standards and document formats such as formats for storage and exchange of book catalog cards (based on MARC21, UNIMARC, MARCXML standards, or national standards), access standards (Z39.50 protocol), and common formats of documents (DOC, PDF, DJVU, HTML, CHM, FB2, and EPUB). The proposed architecture of the DLW is a simplified version of traditional warehouse (data shop-case) and includes three levels: Warehouse Level, Web Database Level, and End User Level [18]. Two novel possibilities in the EPUB format were proposed: the original page markup and the work with audio and video fragments. The first possibility is important for a group work of end users, when the current pages are uniformly assigned. The second possibility is useful for more accurate delivery of video and audio resources. Video sequences are divided into the scenes or music fragments with corresponding descriptions. Also common rates and types of vulnerabilities of the proposed DLW were analyzed.

Chapter 7 describes a method for enabling an ontological interaction on video clip shown on ubiquitous systems as a computer monitor, mobile or tablet.

Interactive multimedia systems help people in their ordinary life [19]. The main goal of this chapter is to enrich user interactions on video contexts. The proposed “Layered Ontological Image Interaction” on video contexts allows a user to directly access and process objects attributes as basic video components. After having extracted the underlying data structures contributed by users towards folksonomy tags, these tags are correlated with the source and time living in a structure called Folksodiven [20, 21]. The Folksodiven helps to bridge the semantic gap between words describing an image and its visual features that could be extremely helpful, when training data is limited. Such approach was examined in practical implementation.

1.3 Conclusion

This chapter has provided a briefly description of six chapters related to the practical implementation of various intelligent interactive multimedia systems and services. The contributions involve the recent achievements in designing multimedia systems in a bi-modal user interface for affect recognition, education aspect of the SNSs (Facebook), the peer-to-peer communication in e-Learning, digital video stabilization, advanced interaction with the DLW for a group work of end users, and ontological interactions on video contexts. Each chapter of the book includes a description of implementation with illustrative material in detail.

Resources

The readers may explore some resources from the KES international web page as a starting point to explore the field further.

KES International

<http://www.kesinternational.org>

KES International was founded for providing a professional community the opportunities for publications, knowledge exchange, cooperation and teaming. Involving around 5000 researchers drawn from universities and companies world-wide, KES facilitates international cooperation and generate synergy in teaching and research. KES regularly provides networking opportunities for professional community through one of the largest conferences of its kind in the area of KES.

KES Conference Series

The International Conference on Knowledge-based Intelligent Engineering and Information Systems (KES) is an annual event. In 1997, 1998 and 1999, the KES conferences were in Adelaide, Australia. In 2000, the conference was in Brighton, UK; in 2001, Osaka, Japan; in 2002, Crema, near Milan, Italy; in 2003, Oxford, UK; in 2004, Wellington, New Zealand; in 2005, Melbourne Australia; in 2006, Bournemouth, UK; in 2007, Salerno, Italy; in 2008, Zagreb, Croatia; in 2009, Santiago, Chile; in 2010, Cardiff, UK; in 2011, Kaiserslautern, Germany; in 2012, San Sebastian, Spain; in 2013, Kitakyushu, Japan; in 2014, Gdynia, Poland. In 2015, it will be held in Singapore. KES Membership numbers have grown from about 100 in 1997 to its present number of about 5000. The conference attracts delegates from many different countries, in Europe, Australasia, the Pacific Rim, Asia and the Americas. In addition to its annual conference, KES also organises symposia on specific technical topics, for example, Agent and Multi Agent Systems (KES-AMSTA-07, -08, -09 -10, -11, -12, -13, -14 and -15), Intelligent Decision Technologies (KES-IDT-09, -10, -11, -12, -13, -14 and -15), Intelligent Interactive Multimedia: Systems and Services (KES-IIMSS-08, -09, -10, -11, -12, -13, -14 and -15), and so on.

KES Journals

KES produces academic journals published by leading Netherlands scientific publisher, IOS Press.

- International Journal of Knowledge-Based Intelligent Engineering Systems, IOS Press, The Netherlands
<http://www.iospress.nl/journal/international-journal-of-knowledge-based-and-intelligent-engineering-systems/>
- Intelligent Decision Technologies: An International Journal, IOS Press, The Netherlands
<http://www.iospress.nl/loadtop/load.php?isbn=18724981>

Book Series

KES members are involved in book series published by leading publishers.

- Book Series on Smart Innovation, Systems and Technologies: www.springer.com/series/8767
- Book Series on Intelligent Systems Reference Library: www.springer.com/series/8578
- Book Series on Advanced Information and Knowledge Processing: www.springer.com/series/4738

References

1. Zeng, Z., Tu, J., Liu, M., Huang, T., Pianfetti, B., Roth, D., Levinson, S.: Audio-visual affect recognition. *IEEE Trans. Multimed.* **9**(2), 424–428 (2007)
2. Alepis, E., Virvou, M., Kabassi, K.: Development process of an affective bi-modal intelligent tutoring system. *Intell. Decis. Technol.* **1**, 1–10 (2007)
3. Kabassi, K., Virvou, M.: A knowledge-based software life-cycle framework for the incorporation of multicriteria analysis in intelligent user interfaces. *IEEE Trans. Knowl. Data Eng.* **18**(9), 1265–1277 (2006)
4. Mazman, S., Usluel, Y.: Modeling educational usage of Facebook. *Comput. Educ.* **55**(2), 444–453 (2010)
5. Promnitz-Hayashi, L.: A learning success story using facebook. *Stud. Self-Access Learn. J.* **2**(4), 309–316 (2011)
6. Basile, T., Esposito, F., Ferilli, S.: Improving user stereotypes through machine learning. *Commun. Comput. Inf. Sci.* **249**, 38–48 (2011)
7. Virvou, M., Troussas, C., Caro, J., Espinosa, K.J.: User modeling for language learning in Facebook. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) *Text, Speech and Dialogue, LNCS*, 7499, pp. 345–352. Springer, Berlin (2012)
8. Abu Tair, M.M., El-Halees, A.M.: Mining educational data to improve students' performance: a case study. *Int. J. Inf. Commun. Technol. Res.* **2**(2), 140–146 (2012)
9. Burdescu, D.D., Mihăescu, M.C.: Tesys: e-learning application built on a web platform. In: *International Joint Conference on e-Business and Telecommunications*. Setubal, Portugal, pp. 315–318 (2006)
10. Cilogluligil, B., Inceoglu, M.M.: user modeling for adaptive e-learning systems, computational science and its applications. In: Murgante, B., Gervasi, O., Misra, S., Nedjah, N., Rocha, A.M.A.C., Taniar, D., Apduhan, B.O. (eds.) *Computational Science and Its Applications—ICCSA 2012, LNCS*, 7335, pp. 550–561. Springer, Berlin (2012)
11. Mocanu, M., Popescu, P.-S., Burdescu, D.D., Mihaescu, M.C.: Advanced messaging system for on-line educational environments. In: Tsihrintzis, G.A., Virvou, M., Watanabe, T., Jain, L.C., Howlett, R.J. (eds.) *Intelligent Interactive Multimedia Systems and Services*, vol. 254, pp. 61–69. IOS Press, Amsterdam (2013)
12. Sugeno, M.: *Industrial applications of fuzzy control*. Elsevier Science Inc., New York (1985)
13. Rawat, P., Singhai, J.: Review of motion estimation and video stabilization techniques for hand held mobile video. *Int J Sig. Image Process* **2**(2), 159–168 (2011)
14. Shakoor, M.H., Moattari, M.: Statistical digital image stabilization. *J. Eng. Technol. Res.* **3**(5), 161–167 (2011)
15. Favorskaya, M., Buryachenko, V.: Video stabilization of static scenes based on robust detectors and fuzzy logic. *Front. Artif. Intell. Appl.* **254**, 11–20 (2013)
16. Yin, L.T.: *Handbook of research on digital libraries: design, development, and impact*. Her-shey, New York (2009)
17. Arulanandam, S., Jaganathan, S., Avula, D.: P2P and grid computing: opportunity for building next generation wireless multimedia digital library. *EURASIP J. Wirel. Commun. Netw.* **165**, 1–16 (2012)
18. Favorskaya, M., Damov, M.: Architecture and formats of digital e-library warehouse. *Front. Artificial Intell. Appl.* **254**, 21–30 (2013)
19. Sebe, N.: *Human-centered computing handbook of ambient intelligence and smart environments, Part IV*, pp. 349–370 (2010)
20. Dal Mas, M.: Elastic adaptive ontology matching on evolving folksonomy driven environment. In: *IEEE Conference on Evolving and Adaptive Intelligent System (EAIS 2012)*, pp. 35–40, Madrid, Spain, IEEE (2012)
21. Dal Mas, M.: Elastic adaptive dynamics methodology on ontology matching on evolving Folksonomy driven. *Environ. Evol. Syst. J.* **5**(1), 33–48 (2014)

Chapter 2

On the Use of Multi-attribute Decision Making for Combining Audio-Lingual and Visual-Facial Modalities in Emotion Recognition

**Maria Virvou, George A. Tsihrintzis, Efthimios Alepis,
Ioanna-Ourania Stathopoulou and Katerina Kabassi**

Abstract In this chapter, we present and discuss a novel approach that we have developed for the integration of audio-lingual and visual-facial modalities in a bi-modal user interface for affect recognition. Even though researchers acknowledge that two modalities can provide information that is complementary to each other with respect to affect recognition, satisfactory progress has not yet been achieved towards the integration of the two modalities. In our research reported herein, we approach the combination of the two modalities from the perspective of a human observer by employing a multi-criteria decision making theory for dynamic affect recognition of computer users. Our approach includes the specification of the strengths and weaknesses of each modality with respect to affect recognition concerning the 6 basic emotion states, namely *happiness*, *sadness*, *surprise*, *anger* and *disgust*, as well as the emotionless state which we refer to as *neutral*. We present two empirical studies that we have conducted involving

M. Virvou (✉) · G.A. Tsihrintzis · E. Alepis · I.-O. Stathopoulou
Software Engineering Lab, Department of Informatics, University of Piraeus, 185 34 Piraeus,
Greece
e-mail: mvirvou@unipi.gr

G.A. Tsihrintzis
e-mail: geoatsi@unipi.gr

E. Alepis
e-mail: talepis@unipi.gr

I.-O. Stathopoulou
e-mail: iostath@unipi.gr

K. Kabassi
Department of Restoration and Conservation of Cultural Heritage, Technological Educational
Institute of the Ionian Islands, 291 00 Zakynthos, Greece
e-mail: kkabassi@teiiion.gr

human users and human observers concerning the recognition of emotions from audio-lingual and visual-facial modalities. Based on the results of the empirical studies, we assign weights to criteria for the application of a multi-criteria decision making theory. Additionally, the results of the empirical studies provide information that may be used by other researchers in the field of affect recognition and is currently unavailable in the relevant literature.

Keywords Empirical studies • Affective computing • Facial expression analysis • Multi-modal interfaces • Audio-lingual affect recognition • Visual-facial affect recognition • Multi-criteria decision making

2.1 Introduction

The recognition of emotions can lead to affective user interfaces that take into account users' feelings and can adapt their behaviour accordingly. Neurologists and psychologists have made progress in demonstrating that emotion is at least as and perhaps even more important than reason in the process of decision making and action deciding [1]. Moreover, the way people feel may play an important role in their cognitive processes as well [2]. This important motivation has led to a wealth of recent research efforts toward the recognition of emotions of users while they interact with software applications. Picard points out that one of the major challenges in affective computing is to try to improve the accuracy of recognising people's emotions [3]. Improving the accuracy of emotion recognition may imply the combination of many modalities in user interfaces. Indeed, human emotions are usually expressed in several ways. Human faces, people's voices, or people's actions may all show emotions. As we articulate speech, for example, we usually move the head and exhibit various facial emotions [4]. Moreover, humans convey emotional information both intentionally and unintentionally via speech patterns; these vocal patterns are perceived and understood by listeners during conversation [5]. Hence the typical keyboard-mouse input device may now seem too limited for the goal of emotion recognition.

Ideally, evidence from many modalities of interaction should be combined by a computer system so that it can generate as valid hypotheses as possible about users' emotions. It is hoped that the multimodal approach may provide not only better performance, but also more robustness [6]. Similar views about the benefits of the combination of modalities have been supported by many researchers in the field of human-computer interaction [7–11]. However, progress in emotion recognition based on multiple modalities has been rather slow. Although several approaches have been proposed to recognize human emotions based on facial expressions or speech unimodally, relatively limited work has been done to fuse these two and other modalities to improve the accuracy and robustness of the emotion recognition systems [12].

In view of the above, it is the aim of this chapter to combine modalities in order to improve the accuracy and overall performance of emotion recognition. In this chapter, we focus on the combination of the audio-lingual modality and the visual-facial modality. Our approach for the combination of the two modalities is by employing Multi-Attribute Decision Making, henceforth referred to as MADM.

As a first goal of an attempt to combine the audio-lingual modality and the visual-facial modality, one has to determine the extent to which these two different modalities can provide emotion recognition from the perspective of a human observer. Moreover, one has to specify the strengths and weaknesses of each modality. In this way, one can determine the weights of the criteria that correspond to the respective modalities from the perspective of a human observer. Hence, empirical studies need to be conducted first concerning emotion recognition based on two modalities: the audio-lingual and the visual-facial.

Such empirical studies constitute an important milestone and yield important results. Not only do they provide the basis towards the combination of modalities into the affective user modelling component of a bi-modal system, but they also give evidence for other researchers to use since, currently, there are not many results from such empirical studies in the literature. Indeed, after an extensive search of the literature, one finds that there is a shortage of empirical evidence concerning the strengths and weaknesses of these modalities. In this chapter, we present each of the empirical studies and show and discuss results from their comparison. Moreover, we present and discuss a novel integration approach that we have developed, which is based on the results of the empirical studies as they have been incorporated in a multi-criteria decision making theory. Specifically, in this chapter, emphasis is placed on six basic emotions, namely *happiness*, *sadness*, *surprise*, *anger* and *disgust*, as well as the emotionless state, which we refer to as *neutral*.

The main body of the chapter is organised as follows. In Sect. 2.2, we present and discuss related work. We also show how important it is for the efficient application of MADM into the problem of affect recognition to have conducted empirical studies using human subjects. In Sect. 2.3, we present and discuss the aims of our empirical studies and their settings. In Sect. 2.4, we describe the empirical study concerning the audio-lingual modality. In Sect. 2.5, we describe the empirical study concerning the visual-facial modality. In Sect. 2.6, we discuss the results produced by the empirical studies. In Sect. 2.7, we describe the architecture of our bi-modal affect recognizer and we show how the results of the empirical studies have been combined using MADM in order to integrate information for affect recognition from the audio-lingual and visual-facial modalities. In Sect. 2.8, we present the conclusions drawn from the empirical studies and point to related future work.

2.2 Related Work

The issue of combining two modalities raises the problem of how these modalities may be combined. In fact, the mathematical tools and theories that have been used for affect recognition can lead to a classification of affect recognizers. Such classification

has been made in [13] where affect recognizers have been classified into two groups on the basis of the mathematical tools that these recognizers have used: (1) The first group using traditional classification methods in pattern recognition, including rule-based systems, discriminate analysis, fuzzy rules, case-based and instance-based learning, linear and nonlinear regression, neural networks, Bayesian learning and other learning techniques. (2) The second group of approaches using Hidden Markov Models, Bayesian networks etc. Indeed, a recent piece of research that uses the above approaches for the integration of audio-visual evidence is reported in [14]. Specifically, for person-dependent recognition, Zeng and his colleagues [11] apply the voting method to combine the frame-based classification results from both audio and visual channels. For person-independent tests, they apply multi-stream hidden Markov models (HMM) to combine the information from multiple component streams.

In contrast with all of the above tools, in this chapter, we present a solution that we have developed to the problem of *bi-modal* affect recognition based on the use of the Multiple Attribute Decision Making (MADM) methodology. MADM involves making preference decisions (such as evaluation, prioritisation, selection) over the available alternatives that are characterised by multiple, usually conflicting attributes [14]. In our case, the information from the two modalities is combined using a multi-criteria decision theory [14, 15]. More specifically, each modality is considered as a criterion that a human observer would use in order to recognise another human's emotion. Thus, we have focused on how humans perceive other humans' emotions and utilise this information to create criteria for the system to use. Our aim is to render bi-modal affect recognition as human-like as possible.

The motivation underlying this particular approach stems from our belief that affect recognition in computerized systems should be performed in ways that are similar to those of human-human interaction, so that affective human-computer interaction may become more natural. This is a belief that is also expressed by other researchers in the field. For example, Nasoz and Lissetti [16] express the belief that human-human interactions should form the basis for the design of human-computer interfaces since it is important to facilitate a natural and believable interaction between the user and the computer. The incorporation of other human reasoning theories, such as Human Plausible Reasoning, has been done successfully in past user interfaces [17].

MADM can provide a formal mathematical tool for the computer to judge which particular emotion the human computer-user is experiencing based on evidence from the visual-facial and audio-lingual modalities. In many cases, an emotion may be better recognized by a particular modality. For example, if a user is seen by a camera smiling without saying anything then this user is likely to have experienced *happiness*, based on the visual-facial modality while there is no evidence of emotion from the audio-lingual modality. On the other hand, if a user swears at the computer then this user is likely to have experienced *anger*, based on the audio-lingual modality irrespective of whether there is any additional evidence on the user's emotions from the visual-facial modality. A smaller and different part of our approach that combines two other modalities using MADM, namely the keyboard and the audio-lingual, is described in [18].

Decision making theories seem very promising although they have not yet been used for this purpose for combining modalities of interaction from the point of view of a human observer. Decision processes with multiple attributes deal with human judgement that takes into account several criteria and pinpoints to the best possible result among conflicting hypotheses. In affect recognition that is based on multiple modalities, the criteria that a human observer may use in order to recognize the emotions of a fellow human who s/he interacts with, may be regarded as criteria that may be used by the human observer to select the emotion that the fellow human is most likely to have.

The adaptation of a multi-criteria theory involves specifying the criteria that are usually taken into account by a human decision maker and calculating the importance of each criterion in the decision maker's reasoning process. Moreover, it involves incorporating the theory into the software. Therefore, the process of the application of a decision making theory requires conducting experiments, which aim at acquiring knowledge from human experts. The experiments play a crucial role in the resulting reasoning of a system. Indeed, if the experiments are not carefully designed and implemented, then there is a possibility that useful pieces of knowledge are missed out and the application of the decision making theory fails in the end [19].

So far, in the literature of human-computer interaction, MADM methods have been used for several purposes, such as selecting the best information source when a user submits a query [20], improving intelligent user interfaces [21], modelling user preferences in recommender systems [22], selecting the best route in mobile guides [23], or individualising e-commerce web pages [24, 25]. However, MADM methods have not been used for affect recognition by providing an integrating mechanism of different modalities.

2.2.1 Multi-attribute Decision Making

In more detail, a multi-attribute decision problem is a situation in which, having defined a set A of actions and a consistent family F of n attributes g_1, g_2, \dots, g_n ($n \geq 3$) on A , one wishes to rank the actions of A from best to worst and determine a subset of actions considered to be the best with respect to F [26]. According to Triantaphyllou and Mann [27], there are three steps in utilising any decision making technique involving numerical analysis of alternatives:

1. Determining the relevant attributes and alternatives.
2. Attaching numerical measures to the relative importance of the attributes and to the impacts of the alternatives on these attributes.
3. Processing the numerical values to determine a ranking of each alternative.

The determination of the relevant attributes and their relative importance is made at the early stages of the software life-cycle and is performed by the developer or is based on an empirical study which may involve experts in the domain.

There are several MADM theories. Among them, we have used in our research the Simple Additive Weighting (SAW) [14, 15] method, which is probably the best known and most widely used decision making method. SAW consists of translating a decision problem into the optimisation of some multi-attribute utility function U defined on A . The decision maker estimates the value of function $U(X_j)$ for every alternative X_j and selects the one with the highest value. The multi-attribute utility function $U(X_j)$ can be calculated in the SAW method as a linear combination of the values of the n attributes:

$$U(X_j) = \sum_{i=1}^n w_i x_{ij}, \quad (2.1)$$

where X_j is one alternative and x_{ij} is the value of the i attribute for the X_j alternative.

An empirical study, at the early stages of the development of a MADM system, aims at determining the weights of importance of the criteria that human experts take into account while observing users interacting with a computer. In the next section, we describe the empirical studies that we conducted concerning affect recognition by human observers.

2.3 Aims and Settings of the Empirical Studies

Affect recognition aims at recognizing human emotions. However, there are too many emotions for all to be distinct and equally basic. The thesis that there are basic emotions is not implausible [28]. In our empirical studies, we focus on the recognition of six basic emotions with respect to audio-lingual and visual-facial modalities. Specifically, the six basic emotional states are *happiness*, *sadness*, *surprise*, *anger* and *disgust* as well as the emotionless state, which we refer to as *neutral*. These six emotional states in our study were selected because they were fundamental in the majority of previous emotion recognition studies. Indeed, most of these emotional states have been selected by several researchers and theorists over the past years. For example, the emotions of anger happiness and surprise are considered in many research works [29–45].

For the purposes of our research we conducted two empirical studies concerning the audio-lingual and visual-facial modalities of interaction. The audio-lingual and the visual-facial empirical studies were conducted simultaneously in two different phases and their results were compared and combined in the end. There were two kinds of roles for the participants in both empirical studies: (1) The human participants who were used to express emotions in the modalities in question. The expressions of emotion in different situations were recorded in computer protocols. (2) The human participants who served as observers and recognisers of the emotions that were recorded in the computer protocols. The human participants of both roles were selected from the same groups for both experiments. In fact, there were 300 participants of various ages and educational background that

were used for the elicitation of emotions. Moreover, there were 50 participants of similar background that were used as observers. The experiment took place in Greece, therefore the participants of both roles were all Greek, so that their emotion expressions and recognition respectively, were compatible with the Greek culture and average temper of the people living in Greece. In our view, empirical studies of this kind should be culture-dependent for more reliable results.

The aims of each of the empirical studies were the following:

1. To elicit and record emotion expressions of human participants in the respective modalities.
2. To create databases of known expressions of emotions in the respective modalities. Specifically, in the audio-lingual modality, the aim was to create a database of typical words/phrases connected to each of the basic emotions. On the other hand, in the visual-facial modality, the aim was to create a database of facial expressions for each of the basic emotions.
3. To specify degrees of recognisability of each of the basic emotions by human listeners or visual observers for the respective modalities. This means that we aimed at estimating the extent to which a particular emotion (e.g. happiness) is recognizable by each of the modalities investigated. Such findings are useful towards combining the audio-lingual and the visual-facial modalities in the sense that we specify the strengths and weaknesses of each modality with respect to each of the basic emotions. The comparison of the results is important for the application of the multi-criteria theory.
4. To specify important parts (criteria) of recognisability in the respective modalities. For example, in the audio-lingual modality, the goal was to find out how recognition could be achieved based on the actual words that a user had said as well as the volume of the user's voice. In the visual-facial modality, the goal was to classify emotion recognisability from specific portions of the face (e.g. mouth, eyes, eye brows etc.).

The settings of the two empirical studies were the same. However, one important difference of the experiments was connected to the questions asked to human observers concerning recognisability of emotions via the two different modalities. The elicitation of emotions of subjects and the creation of databases of emotions were based on the human participants of the empirical studies who expressed emotions. On the other hand, the analysis of recognisability of emotions was based on human participants that were used as observers of the recorded expression of emotions.

2.3.1 Elicitation of Emotions and Creation of Databases

As a first step of the empirical studies we had to ensure that we could elicit emotions of human participants so that these could be recorded in databases for further analysis and use. The elicitation of users' emotions constitutes an important part of the settings of empirical studies concerning affective computing. For

example, Nasoz and Lisetti [16] conducted an experiment, where they elicited six emotions (sadness, anger, surprise, fear, frustration, and amusement) and measured physiological signals of subjects by showing them movie clips from different movies (The Champ for sadness, Schindler's List for anger, The Shining for fear, Capricorn One for surprise, and Drop Dead Fred for Amusement) and by asking them hard math questions (for frustration). They conducted a panel study to choose an appropriate movie clip to elicit each emotion. In contrast, our empirical studies aimed at recording emotions for different modalities; we did not aim at recording physiological signals of users but rather aimed at recording and analyzing visual-facial and audio-lingual expressions of emotions. For the purposes of our studies, we used an educational application to elicit emotions of subjects. Indeed, the educational procedure can provoke emotions. For example, Nasoz and Lisetti [16] have used hard mathematical problems for the elicitation of frustration of subjects.

In our empirical study we created and used an educational application for all six emotions that we were aiming at recording and analyzing. The educational application incorporated a monitoring module that was running unnoticeably in the background. Moreover, users were also video taped while they interacted with the application.

More specifically, the experimental system consisted of the main educational application with the presentation of theory and tests, a programmable human-like animated agent, a user monitoring component, and a database. While using the educational application from a desktop computer, participants were being taught a particular medical course. We have selected medicine as a subject area as it is an area of high interest for many people nowadays and we also considered it more appropriate for the elicitation of many different emotions.

The information was given in text form while at the same time the animated agent read it out loudly using a speech engine. Participants could choose a specific part of the human body and all the available information was retrieved from the system database. An example of using the main application is illustrated in Fig. 2.1. An animated agent was present in these modalities to make the interaction more human-like. The educational application incorporated the capability to manipulate the agent behaviour with regards to movements and gestures of the agent on screen, as well as speech attributes such as speed, volume, and pitch. As a result, the system was enriched with an agent, who was capable to express emotions and, thereby the user was further encouraged to interact with more noticeable emotional evidence in his/her behaviour. Participants were also expected to take tests concerning part of the selected medical theory they had chosen. Both in the reading theory interaction, as well as in the interaction with the medical examinations, participants were expected to express their feelings freely, as a consequence of their interaction with the agent and with the educational environment in general.

Indeed, for the elicitation of the six basic emotions the educational application provided the following situations:

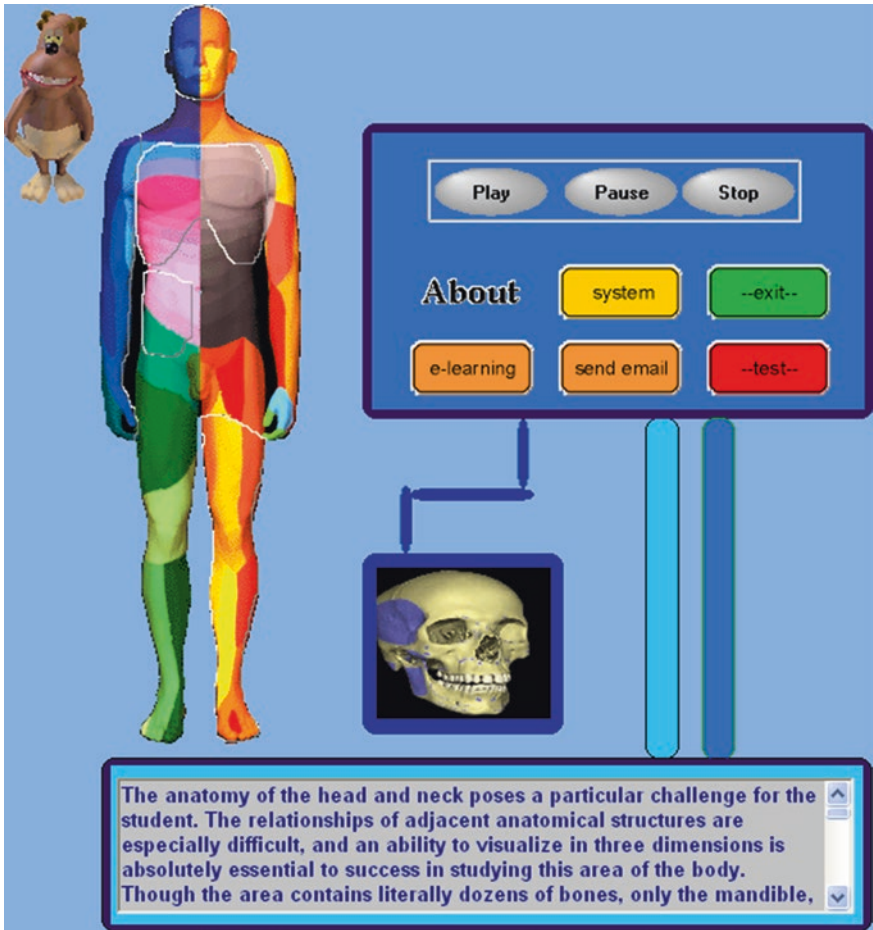


Fig. 2.1 A screen-shot of theory presentation in the medical educational application

1. For happiness: the agent would say jokes and would make funny faces or the user would receive excellent grades in his/her performance on tests.
2. For anger: the agent would be rude and unfair to users for their performance at tests.
3. For disgust: the medical application would show videos of surgical operations showing blood and human organs.
4. For sadness: the medical application would show pictures of patients while they were receiving treatments for serious diseases.
5. For surprise: the agent would pop up into the screen completely unexpectedly.
6. For neutral: the user would use the medical application under normal circumstances.

To ensure that users would experience the above emotions, all of the participants did not have a medical background and, therefore, they were not exposed frequently to this kind of material.

2.3.2 Creation of Databases of Known Expressions of Emotions

As stated above, we aimed at creating databases of known expressions of emotions for the visual-facial and audio-lingual modalities, respectively. This part of the empirical studies could have been skipped if there already existed databases of this kind that we could use for further analysis. Indeed, in the case of visual-facial modality, several face databases are available today, which have been developed by various researchers and could be accessed through the World Wide Web. These include: (1) The AR Face Database [46], which contains over 4,000 color images of 126 persons' faces in front view, forming various facial expressions under various illumination conditions and possible occlusion (e.g., by sun glasses and/or scarf). The main disadvantage of this database is its limitation to containing only four facial expressions, namely "neutral", "smiling", "anger", and "scream". (2) The Japanese Female Facial Expression (JAFFE) Database [47], which contains 213 images of the neutral and 6 additional basic facial expressions as formed by 10 Japanese female models. (3) The Yale Face Database [48], which contains 165 gray-scale GIF-formatted images of 15 individuals. These correspond to 11 images per subject of different facial expression or configuration, namely, center-light, with glasses, happy, left-light, without glasses, normal, right-light, sad, sleepy, surprised, and wink. (4) The Cohn-Kanade AU-Coded Facial Expression Database [49], which includes approximately 2000 image sequences from over 200 subjects and is based on the Facial Action Coding System (FACS), first proposed by Paul Ekman [30]. (5) The MMI Facial Expression Database [50], which includes more than 1500 samples of both static images and image sequences of faces in front and side view, displaying various expressions of emotion and single and multiple facial muscle activation.

Although many of the aforementioned face databases were considered for further analysis by human observers, we found that either the number of different facial expressions or the scope of the facial expression formation process did not entirely match our goals and studies. Furthermore, the aforementioned databases were constructed by photographing persons coming from diverse cultures. In our view, this would create a problem for the recognisability of emotions expressed because people in different cultures may have different ways of expressing emotions. Thus, we made the decision to create our own facial expression database [51] by photographing Greek people expressing the six emotional states mentioned previously.

In the case of audio-lingual modality there was no existing database that we could have used. Therefore, we certainly had to create our own database for the purposes of our research.

2.3.3 Analysis of Recognisability of Emotions by Human Observers

Recognizability of emotions by human observers consisted of specifying degrees of recognisability of each of the basic emotions by human listeners and visual observers for the respective audio-lingual and visual-facial modalities.

The empirical studies involved a total number of 300 male and female users of various educational backgrounds, ages, and levels of familiarity with computers. These users were asked to use the medical educational application and their actions were video recorded. Then, after they had completed their interaction with the application, participants were asked to watch the video clips concerning exclusively their personal interaction and to determine in which situations they experienced changes in their emotional state. Then, they associated each change in their emotional state with one of the six basic emotion states in our study and the data was recorded and time-stamped. In this way, we had recorded what the actual emotions of the users were.

2.4 Empirical Study for Audio-Lingual Emotion Recognition

The first empirical study that we have conducted concerns audio-lingual emotion recognition. In this empirical study, the audio-lingual modality of interaction is based on using a microphone as input device. The empirical study aimed at identifying common user reactions that express user emotions while they interact with computers. As a next step, we associated the reactions with particular emotions.

2.4.1 The Experimental Educational Application for Elicitation of Emotions

The participants were asked to use the medical educational application, which incorporated a user monitoring module. Figure 2.2 illustrates a snapshot of the monitoring application module while it records the user microphone input and the exact time of each event. In this study, we took into consideration only the data from the audio-lingual modality of interaction.

2.4.2 Audio-Lingual Modality Analysis

The basic function of this experiment was to capture the data inserted by the user orally. The data was recorded into a database of audio-video clips. The monitoring component recorded the actions of users with the microphone.

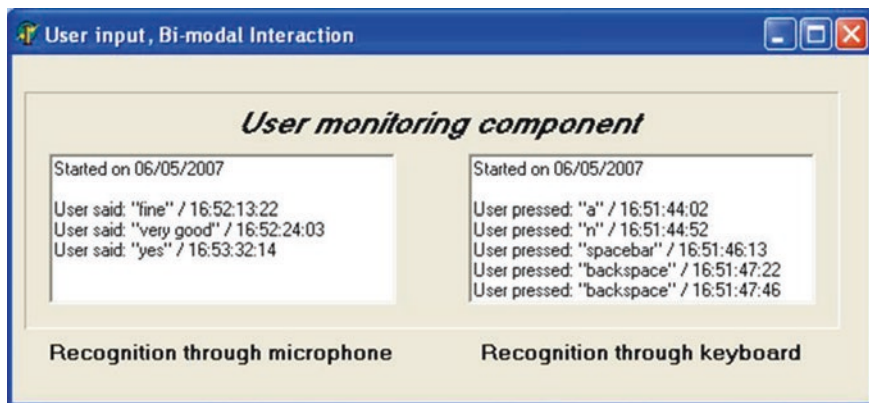


Fig. 2.2 Snapshot of operation of the user monitoring component

In the next step, the collected transcripts were given to 300 users as human expert-observers, who were asked to perform emotion recognition with regards to the six emotional states, namely happiness, sadness, surprise, anger, disgust and neutral. Successful recognition was considered, when an observer could recognize the emotion that the user had confirmed that s/he had experienced at a particular state of the video. All human expert-observers were asked to analyze the data corresponding to the audio-lingual input only. Therefore, they were asked to listen to the video tapes without seeing them. They were also given what the user had said in printed form from the computer audio recorder. The human expert-observers were asked to justify the recognition of an emotion by indicating the weights of the criteria that they had used in terms of specific words and exclamations, as well as the volume of voice.

This part of the study has supplied us with a significant number of words, phrases and exclamations, all associated with emotional states, which are used in the creation of a database of words for emotion recognition. Thus, a database has been constructed from the users' oral input and words, phrases and exclamations. At the same time changes in voice volume and voice pitch were also recorded in relation to the oral input.

Table 2.1 illustrates the results of the empirical study in terms of the oral input via microphone and the six basic emotions (neutral, happiness, sadness, surprise, anger, disgust) recognized by human expert-observers. For each emotion, we note the percentages of the users' oral reaction or the absence of audio input. Furthermore, Table 2.1 illustrates the changes in the users' voice while uttering a word or phrase or saying an exclamation in each emotional situation. For example, a user in surprise may have uttered an exclamation (58 %) rather than having spoken a word (24 %) or having remained silent (18 %). This action is recognised at a degree of 66 % accompanied by an increase in the user's voice volume. A summary of the results is illustrated in Table 2.1.

It should be noted that bored (neutral emotion) users and sad users were orally less expressive than users in other emotional states. However, in cases where bored

Table 2.1 Empirical study results

	Say an exclamation	Utter a certain word or phrase	Keep silent
Emotions	Change in volume (%)	Change in volume (%)	Change in volume (%)
Neutral	6	22	72
	45	37	
Happiness	31	45	24
	40	55	
Sadness	8	28	64
	52	44	
Surprise	58	24	18
	66	60	
Anger	39	41	20
	62	70	
Disgust	50	39	11
	64	58	

Human recognition of basic emotional states through microphone

Table 2.2 Percentages of successful emotion recognition by human experts

Emotional state	Percentage of recognition by human experts (audio data) (%)
Neutral	18
Happiness	46
Sadness	48
Surprise	62
Anger	79
Disgust	57

users actually said something, this could trigger a decrease in their voice volume. Users experiencing the emotions of surprise and disgust usually express them by saying an exclamation (58 and 50 %, respectively) while happy users and users in anger would likely say a word or phrase contained in our ‘emotional database’ of words and phrases (45 and 41 %, respectively). Particularly, with regards to the emotions of surprise and anger, users would increase the volume of their voice while saying something. Table 2.2 illustrates the percentages of successful emotion recognition by human experts concerning the participants’ emotional states and the audio modality of interaction. These percentages result after comparison of the recognized emotional states by the human experts and the actual emotional states as recorded by the participants themselves.

In the next step of our study, the participants were asked to specify, which input action from the microphone would help them find out what the emotions of the users were. From the input actions that appeared in the experiment, only those proposed by the majority of the human experts were selected. Considering the users’ basic input actions through the microphone we have 7 cases: (a) user speaks using strong language, (b) user uses exclamations, (c) user speaks with a high

voice volume (higher than the average recorded level), (d) user speaks with a low voice volume (lower than the average recorded level), (e) user speaks in a normal voice volume, (f) user speaks words from a specific list of words showing an emotion, (g) user does not say anything.

These input actions are considered as criteria for evaluating all different emotions and selecting the one that seems to be prevailing. More specifically, in the resulting audio-lingual emotion recognition system, each emotion is evaluated using the criteria (input actions) from the microphone. These criteria are weighted according to the analysis of the empirical studies so that for each emotional state, these seven input action criteria obtain specific values. For the evaluation of each alternative emotion, the system uses SAW for each particular category of users. The overall functionality of this approach (uni-modal recognition of emotions through audio-lingual data), described in [52], requires more comprehensive writing and is beyond the scope of the present chapter. In this chapter, we focus specifically on the combination of pre-given bi-modal information concerning emotion recognition, through a multi-attribute decision making theory, in order to improve the emotion recognition capability of a system.

2.5 Empirical Study for Visual-Facial Emotion Recognition

The basic aim of this experiment was to identify and quantify the most common facial expressions of a user during a typical human-computer interaction session. Firstly, we observed humans during human-computer interaction sessions and concluded that the facial expressions corresponding to the “neutral”, “happiness”, “sadness”, “surprise”, “anger”, and “disgust” emotional states arose very commonly and, thus, form the corresponding classes for our recognition/classification tasks.

2.5.1 Visual-Facial Empirical Study on Subjects

To acquire image data, we built a two-camera system. Specifically, two identical cameras of 800-by-600 pixel resolution were placed with their optical axes on the same horizontal plane and forming a 30° angle. This allowed us to video record and photograph faces in front and side view, simultaneously while users were interacting with the medical educational application.

Videos of the resulting facial expressions were then showed to the participant to verify his/her facial expression. If the subject agreed that the expressed emotions were genuine, with regards to the facial expression, they were saved and labeled; as photographs. The final dataset consists of the images of 300 different individuals, each forming the six expressions: “neutral”, “happiness”, “sadness”, “surprise”, “anger”, and “disgust”, as described in [51].

2.5.2 Visual-Facial Empirical Study by Human Observers

In order to understand how humans classify someone else’s facial expression and set a target error rate for automated systems, we developed a questionnaire that was then answered by 300 participants acting as observers. Our aim was to identify differences between the “neutral” expression of a modality and its deformation into other expressions. We also aimed at quantifying these differences into measurements of the face (such as dimension ratio, distance ratio, texture, or orientation) and, thus, extracting corresponding features that convert pixel data into a higher-level representation of shape, motion, color, texture, and spatial configuration of the face and its components. With these goals in mind, the questionnaire was divided into two separate parts:

1. In the first part of the questionnaire, the participants were asked to map the facial expressions that appeared in 14 images into emotions. Each participant could choose from the 6 common emotions that we mentioned earlier, namely “anger”, “happiness”, “neutral”, “surprise”, “sadness”, and “disgust”, or specify any other emotion that he/she thought appropriate. Next, the participant had to specify the degree (0–100 %) of his/her confidence in the identified emotion. Finally, he/she had to indicate which features (such as the eyes, the nose, the mouth, the cheeks, etc.) had helped him/her make that decision. A typical print-screen of the first part of the questionnaire is illustrated in Fig. 2.3.
2. In the second part of the questionnaire, the participant had to classify the emotion from portions of the face. Specifically, we showed the participant the “neutral” facial image and an image of some facial expression of a subject. The

4a. What emotion does the image represent

Other...

4b. In what percent: %

4c. Which Facial Features helped you understand the emotion?
(you can choose more than one)

Mouth Forehead texture

Eyes Texture between the brows

Shape of the face Texture of the cheeks

Other...

Fig. 2.3 The first part of the questionnaire

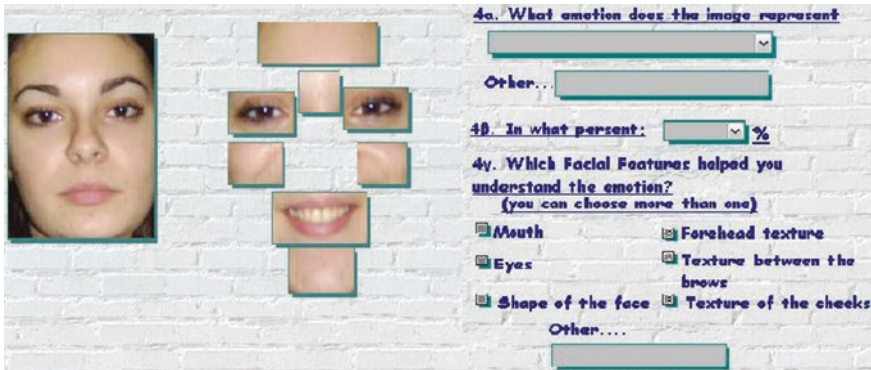


Fig. 2.4 The second part of the questionnaire

latter image was cut into the corresponding facial portions, namely, the eyes, the mouth, the forehead, the cheeks, the chin and the brows. A typical print-screen of this part of the questionnaire is shown in Fig. 2.4. Again, each participant could choose from the 6 emotions mentioned earlier or specify any other emotion that he/she thought appropriate. Next, the participant had to specify the degree (0–100 %) of his/her confidence in the identified emotion. Finally, he/she had to indicate which features (such as the eyes, the nose, the mouth, the cheeks, etc.) had helped him/her make that decision.

When it comes to recognizing an emotion from someone else’s facial expression, the majority of the participants consider it as a difficult task. Generally, “positive” emotions, such as “surprise” or “happiness”, achieved the highest correct recognition rate of 90 and 70 %, respectively. The questionnaire indicates the emotion of “disgust” as the most difficult to recognize, as it was correctly recognized at a rate of only 37 %. Other correct recognition rates by human experts, corresponding to all emotions in our questionnaire, are shown in Fig. 2.5 and Table 2.3.

From our study of the images and the questionnaire, we identified significant deviations from the “neutral” to other emotions, which can be quantified into a classifying feature vector. Typical such deviations are shown in Table 2.3.

The observations of facial changes that arise during formation of various facial expressions, as indicated in Table 2.3, led us to the identification of the most important facial features that can represent these changes in mathematical terms and allow us to form *feature vectors*. The aim of the feature extraction process is to convert pixel data into a higher-level representation of shape, motion, color, texture, and spatial configuration of the face and its components. Specifically, we locate and extract the corner points of specific regions of the face, such as the eyes, the mouth and the eyebrows, and compute variations in size or orientation from the “neutral” expression to another one. Also, we extract specific regions of the face, such as the forehead or the region between the eyebrows, so as to compute variations in texture. Namely, the extracted features are: (1) mouth ratio, (2) left eye ratio, (3) right eye ratio, (4) ratio of the face dimensions, (5) texture

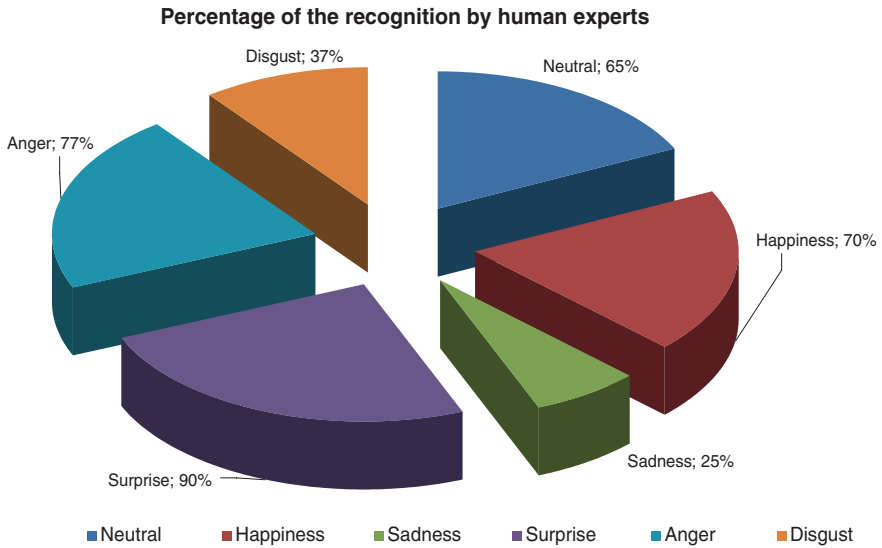








Fig. 2.5 Percentage of the recognition by human experts

Table 2.3 Emotions and facial expression

Results from empirical studies			
Emotion	Facial Image	Deviation from the “neutral expression”	Recognition percentage by human observers (%)
Neutral		<ul style="list-style-type: none"> • Lack of facial skin movement • All the variations depart from this expression 	65
Happiness		<ul style="list-style-type: none"> • Bigger-broader mouth • Slightly narrower eyes • Changes in the texture of the cheeks • Occasionally changes in the orientation of brows 	70
Sadness		<ul style="list-style-type: none"> • Changes in the direction of the mouth • Wrinkles formed on the chin (different texture) • Occasionally, wrinkles formed in the forehead and different direction of the brows 	25

(continued)

Table 2.3 (continued)

Results from empirical studies			
Emotion	Facial Image	Deviation from the “neutral expression”	Recognition percentage by human observers (%)
Surprise		<ul style="list-style-type: none"> • Longer head • Bigger-wider eyes • Mouth opened • Wrinkles in the forehead (changes in the texture) • Changes in the orientation of brows (the brows are raised) 	90
Anger		<ul style="list-style-type: none"> • Wrinkles between the brows(different textures) • Smaller eyes • Wrinkles in the chin • The mouth is tight • Occasional wrinkles over the brows. in the forehead 	77
Disgust		<ul style="list-style-type: none"> • The distance between the nostril and the eyes is shortened • Wrinkles between the brows and on the nose • Wrinkles formed on the chin and the cheeks 	37

measurement of the forehead, (6) texture measurement of the chin, (7) texture measurement of the region between the brows, (8) texture measurement of the left cheek, (9) texture measurement of the right cheek and, (10) brow orientation. The resulting feature vector is fed into a classifier (e.g. artificial neural network, svm-based, etc.) that will attempt to recognize the person’s facial expression. The feature extraction process and the recognition results of a neural network-based system are analyzed and presented at various stages of system development in [53–55]. However, a detailed analysis is beyond the scope of the present chapter that focuses on combining information from two modalities.

2.6 Discussion and Comparison of the Results from the Empirical Studies

On the basis of the results of the empirical studies that were described in the previous sections, we can compare the emotion recognisability achieved by the audio-lingual and the visual-facial modalities from the perspective of a human observer.

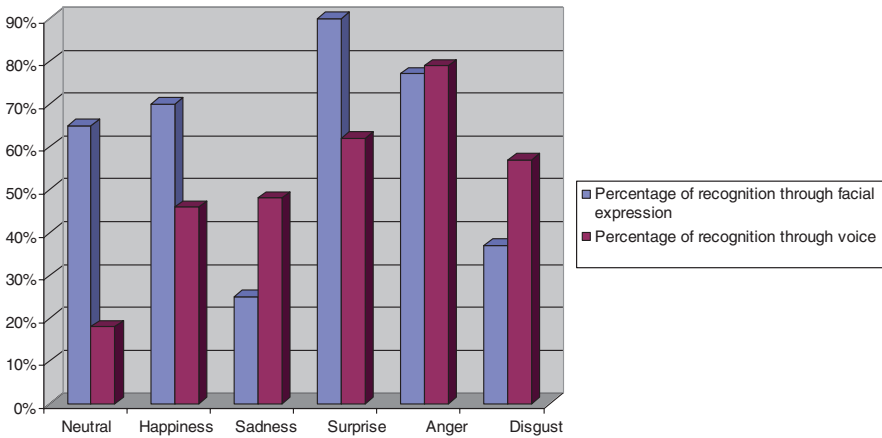


Fig. 2.6 Recognition of emotions through visual-facial and audio-lingual modalities

Figure 2.6 illustrates the percentages of successful emotional recognition through the audio-lingual and visual-facial modalities. The analysis of Fig. 2.6 leads to important conclusions.

In an overall comparison of the two modalities, the visual-facial modality appears to have stronger affect recognition potential than the audio-lingual modality. There are cases where both the audio-lingual and the visual-facial emotion analysis recognize an emotion significantly well. Such cases are for the emotions of anger and surprise where successful recognition is over 50 %. However, surprise can be recognized more easily by the visual-facial modality than the audio-lingual one, as the visual-facial analysis has been successful for 90 % of the cases whereas the audio-lingual modality has been successful for 62 % of the cases. In the case of the emotion of sadness, there is not satisfactory recognition by either modality since in both modalities affect recognition is under 50 %. However, regarding the emotion of sadness, the audio-lingual modality appears to have stronger affect recognition potential, since it achieves a recognition rate of 48 % rather than the rate of 25 % of the visual-facial modality.

When comparing the two modalities, the visual-facial modality is significantly better than the audio-lingual in the recognition of the following three emotional states:

1. Neutral (visual-facial: 65 % versus audio-lingual: 18 %).
2. Happiness (visual-facial: 70 % versus audio-lingual: 46 %).
3. Surprise (visual-facial: 90 % versus audio-lingual: 62 %).

The audio-lingual modality is significantly better than the visual-facial one in the recognition of the following two emotional states:

1. Disgust (audio-lingual: 57 % versus visual-facial: 37 %).
2. Sadness (audio-lingual: 48 % versus visual-facial: 25 %).

The emotional state that is easily recognizable by either modality is:

1. Anger (audio-lingual: 79 % versus visual-facial: 77 %).

The most relevant research work performed by other researchers in the past is that of De Silva et al. [8] who performed an empirical study and reported results on human subjects' ability to recognize emotions. The aims of the empirical study performed by De Silva et al. were similar to the aims of our empirical study, i.e. discovering the best modality for recognizing certain emotions. For this reason, they compared human recognition results in three tests: video only, audio only, and combined audio and video. In particular, De Silva et al. [8] showed video clips of facial expressions and corresponding synchronised emotional speech clips to 18 human subjects and asked them to recognize the emotions of the humans, who were appeared in the video and speech clips. De Silva et al. concluded that some emotions, such as sadness and fear, are better identified via audio clips. On the other hand, they concluded that other emotions, such as anger and happiness, are better identified via video clips. However, De Dilva et al. focused on the audio signals of voice rather than lingual keywords that conveyed affective information. In fact, in order for them to ensure that the lingual information would not interfere with the results of their empirical study, they used human subjects, who were not familiar with the languages used in the video clips, namely Spanish and Sinhala. In contrast with the De Silva et al. approach [8], in our research we have included the lingual aspect of users' spoken words on top of the pitch and volume of their voice and compare the audio-lingual results with the visual-facial results to identify the modality that conveys the largest chunk of information for human observers. Table 2.4 compares our findings with the findings in Da Silva et al. [8].

Table 2.5 illustrates percentages of successful emotion recognition through audio and facial data, based on the work of Busso et al. [12]. In their work, the

Table 2.4 Comparison of our results with De Silva and Huang [9]

Emotions	Audio modality		Video modality	
	Our results (%)	De Silva et al. [8] (%)	Our results (%)	De Silva et al. [8] (%)
Happiness	46	43	70	84
Sadness	48	36	25	16
Anger	79	43	77	66
Disgust	57	–	37	–
Surprise	62	44	90	56
Neutral	18	–	65	–

Table 2.5 Computerized recognition results from Busso et al. [12]

Emotions	Audio (%)	Facial (%)
Anger	68	79
Sadness	64	81
Happiness	70	100
Neutral	81	81

emotional states of humans are recognized through a sophisticated emotion recognition system and not by other humans. Moreover, only four (4) emotional states were recognized, namely anger, sadness, happiness, and neutral. For these reasons, this specific study could not give us significant empirical information for the purposes of our study.

2.7 Combining the Results from the Empirical Studies Through MADM

The empirical data from the two modalities are combined using a multi-criteria decision theory, where each modality is a criterion. The percentages of emotion recognition for each emotion and for each modality are analyzed and then used as weights in order to improve the accuracy of the bi-modal system and determine the prevailing emotion (Fig. 2.7).

According to SAW, emotion recognition values are calculated as a linear combination of the inputs provided by the two distinct modalities and the main aim is to determine the prevailing emotion in order to improve the accuracy of the system. More specifically, using SAW, emotion recognition values are calculated by applying the corresponding criteria weights to the following Eq. (2.2)

$$EM_i = \sum_{m=1}^2 W_{mi} * V_{mi}, \quad i = 1, \dots, 6, \quad (2.2)$$

where EM_i represents the value of successful recognition of the i -th of the six basic emotions through the visual-facial and audio-lingual modality. Moreover, W_{mi} is the weight of the criterion V_{mi} which corresponds to the successful recognition of the i -th emotion through the visual-facial ($m = 1$) or the audio-lingual ($m = 2$) modality. The values of the criteria V_{mi} are calculated in run time by each modality separately and then are incorporated to the resulting system that combines them. Then the system decides for the prevailing emotion (REM in Eq. 2.3) as:

$$REM = \max(EM_i), \quad i = 1, \dots, 6, \quad (2.3)$$

Taking into account the vectors \bar{V}_{1i} and \bar{V}_{2i} , the utility function for the estimation of each emotion recognition is done based on the formulae presented in Table 2.7. Table 2.7 results from Table 2.6, in which the ability of each modality in recognizing an emotional state is illustrated. The values in Table 2.7 emerge from the values in Table 2.8 after normalization to 1 and can be incorporated into the MADM model.

The weights W_{mi} , on the other hand, are static and have resulted from the analysis of the results of the empirical study. In particular, the weights W_{mi} show the system efficiency in recognizing the i -th emotion through the m -th modality. For example, the neutral emotion ($i = 1$) is mostly recognized through the visual-facial modality ($m = 1$) and, therefore, the weight of recognition of the neutral emotion through the visual-facial modality (W_{11}) is higher than the weight of recognition of the neutral

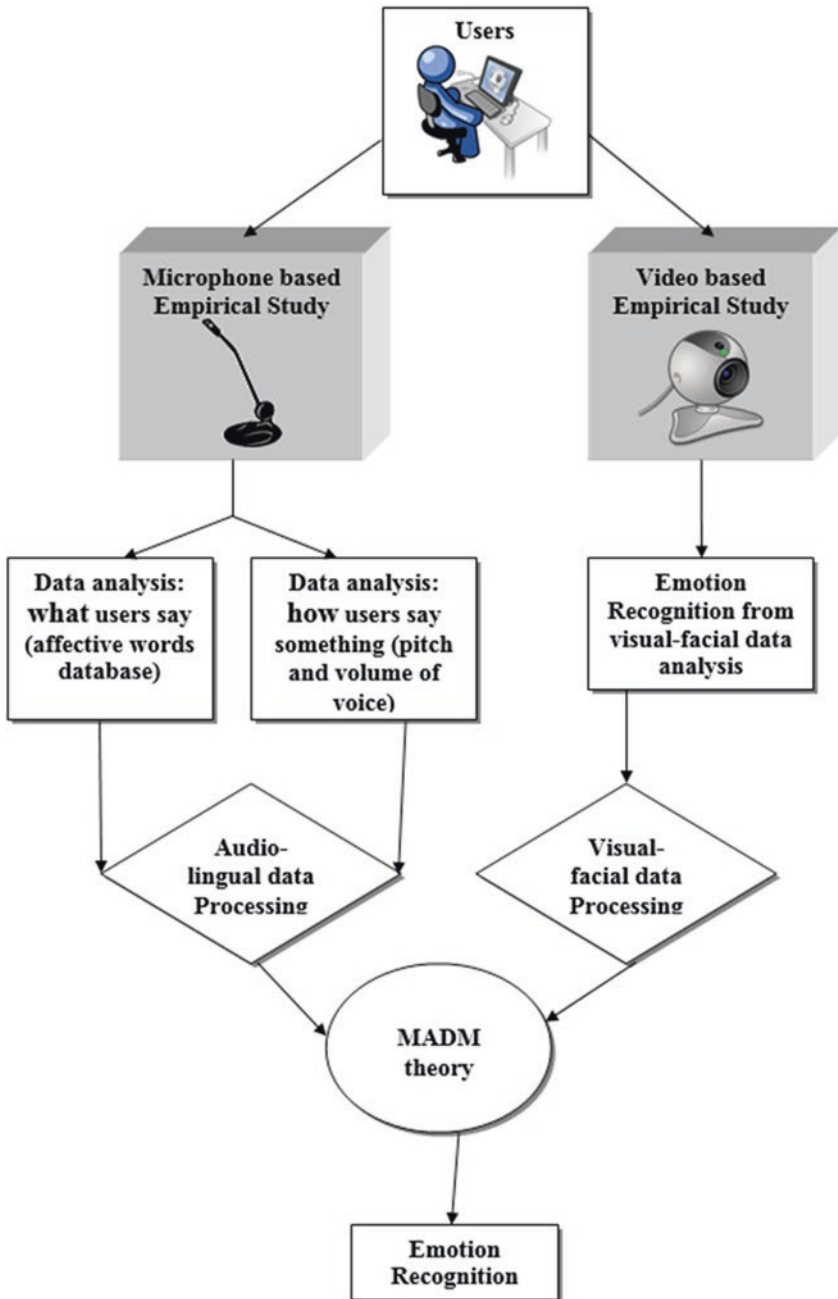


Fig. 2.7 Combining audio-lingual and visual-facial data through MADM

Table 2.6 Recognition of emotions through audio-lingual and visual-facial data

Emotions	Video modality (%)	Audio modality (%)
Neutral	65	18
Happiness	70	46
Sadness	25	48
Surprise	90	62
Anger	77	79
Disgust	37	57

Table 2.7 The formulae for emotion recognition and selection by the system

Emotion	$EM_i = W_{1i} * V_{1i} + W_{2i} * V_{2i}$
Neutral	$EM_1 = 0.78 * V_{11} + 0.22 * V_{21}$
Happiness	$EM_2 = 0.60 * V_{12} + 0.40 * V_{22}$
Sadness	$EM_3 = 0.34 * V_{13} + 0.66 * V_{23}$
Surprise	$EM_4 = 0.59 * V_{14} + 0.41 * V_{24}$
Anger	$EM_5 = 0.49 * V_{15} + 0.51 * V_{25}$
Disgust	$EM_6 = 0.39 * V_{16} + 0.61 * V_{26}$

emotion through the audio-lingual modality (W_{21}). The utility function for the first emotion (neutral emotion) is: $EM_1 = W_{11} * V_{11} + W_{21} * V_{21}$ and after replacing the corresponding weights from Table 2.7 becomes: $EM_1 = 0.78 * V_{11} + 0.22 * V_{21}$. The values for V_{11} and V_{21} are calculated in run time by the system and represent the results of emotion recognition through the visual-facial and the audio-lingual modality, respectively. More specifically, V_{11} and V_{21} are elements of vectors \bar{V}_{1i} and \bar{V}_{2i} correspondingly and they take their values in $[0,1]$. \bar{V}_{1i} is the vector that refers to modality $m = 1$ (visual-facial modality) and \bar{V}_{2i} is the vector that refers to modality $m = 2$ (audio-lingual modality). In the proposed system, each modality sends data to the emotion recognition decision system through vectors $\bar{V}_{1i} = (V_{11}, V_{12}, V_{13}, V_{14}, V_{15}, V_{16})$ and $\bar{V}_{2i} = (V_{21}, V_{22}, V_{23}, V_{24}, V_{25}, V_{26})$. Specifically, the data for the six emotions are represented as vector elements $V_{11}-V_{16}$ for the visual-facial modality and $V_{21}-V_{26}$ for the audio-lingual modality.

In cases where one modality or both modalities contain significant information about a specific recognized emotion, the multi-criteria approach confirms the recognition. However, there are cases where evidence from modalities leads to two possible recognized emotional states. These are cases where a modality fails to recognize an emotion correctly as an emotional state is confused with another. As an example, we may state that for the visual-facial emotion recognition the emotional state of anger projects as a facial expression that may be confused with the facial expression projected by the emotional state of sadness. Hence, visual-facial emotion recognition systems frequently fail to correctly identify these two emotions. The incorporation of the multi-criteria model, with weights that derive from the analysis of the empirical studies, gives the bi-modal system the additional capability to use evidence from more than one modalities of interaction that is complementary to a high extent. Stated in other words, when a modality cannot distinguish between two or more emotional states, the incorporation of evidence from a second modality might provide a solution.

Table 2.8 A possible response from the audio-lingual emotion recognition subsystem

<i>Audio-lingual emotion recognition system</i>						
Emotions	Neutral (%)	Happiness (%)	Sadness (%)	Surprise (%)	Anger (%)	Disgust (%)
Percentage of recognition	6	6	3	37	44	4

As a characteristic example, we consider the situation where a user is angry about something while s/he uses an educational application. Evidence from the user’s voice and face is provided by a video camera that records users’ interaction with the computer. Each modality uses the captured video data and analyzes them in terms of only optical and only acoustical information. Subsequently, both modalities make emotion recognition assumptions. In this particular situation, where a user is experiencing the emotional state of anger, we have the following evidence: the user raises his/her voice volume. At the same time linguistic information, such as the presence of specific exclamations, reveals an unusual emotional state of the user. Finally, characteristic changes appear on the user’s face.

The audio-lingual information (higher voice volume and presence of specific exclamations) is compatible to a high degree with emotional states of anger and surprise. However other emotional states may produce similar audio-lingual information. After the analysis of the six possible emotional states for the seven basic input action characteristics, described in Sect. 2.4, a possible recognized emotion response from the audio-lingual recognizer is the following (Table 2.8):

This response may accordingly result in a vector $\bar{V}_{2i} = (0.06, 0.06, 0.03, 0.37, 0.44, 0.04)$. The resulting response from the audio-lingual recognizer emerges also from the analysis of stereotypical data, both from user characteristics, as well as from user input actions that indicate possible emotional states. A thorough examination of both the stereotypical analysis for audio-lingual emotion recognition and the underlying mechanism for the combination of stereotypical emotional data is presented in previous works of the authors [52]. Table 2.8 indicates that there are two possible recognized emotional states, namely anger and surprise, with the emotional state of anger being dominant.

Continuing our example, snapshots from the video camera provide the other modality for emotion recognition, i.e. facial evidence of the user. Specifically, the visual-facial modality returns a six dimensional vector containing one ace to indicate a recognized expression and five zeroes. For example, when a ‘neutral’ facial expression is recognized, an output value of [1.00; 0.00; 0.00; 0.00; 0.00; 0.00] should be returned ideally. Similarly, when the emotional state of happiness is recognized, the ideal output should be [0.00; 1.00; 0.00; 0.00; 0.00; 0.00] and similarly for the other expressions. Practically, the output vector components will have values in the [0, 1] interval each and all of them adding up to one. Thus, the output vector components can be regarded as a set of six degrees of membership of the face image in each of the six emotional states, namely “neutral”, “happiness”, “surprise”, “anger”, “disgust”, and “sadness” [55]. For example, a typical output vector from the visual-facial modality could be: [0.10; 0.01; 0.44; 0.01; 0.40;

0.04]. This implies that the visual-facial recognizer considers the input expression as “neutral”, “happiness”, “sadness”, “surprise”, “anger”, and “disgust” with a confidence of 10, 1, 44, 1, 40 and 4 %, respectively.

In this particular situation the emotional states of anger and sadness are recognized with similar degrees of confidence. Furthermore, the visual-facial emotion recognition system would wrongly recognize sadness, rather than anger, as the most probable emotional state.

Thus, the two modalities return the output vectors $\bar{V}_{1i} = (0.10, 0.01, 0.44, 0.01, 0.40, 0.04)$ (visual-facial modality) and $\bar{V}_{2i} = (0.06, 0.06, 0.03, 0.37, 0.44, 0.04)$ (audio-lingual modality). Their elements represent the degree of confidence of correct emotion recognition for each of the six emotions. In this case, we can calculate the values EM_i for each one of the six emotions:

$$EM_1 = 0.78 * V_{11} + 0.22 * V_{21} \Rightarrow EM_1 = 0.78 * 0.10 + 0.22 * 0.06 \Rightarrow EM_1 = 0.091,$$

Correspondingly $EM_2 = (0.60 * 0.01 + 0.40 * 0.06) = 0.03$, $EM_3 = (0.34 * 0.44 + 0.66 * 0.03) = 0.169$, $EM_4 = (0.59 * 0.01 + 0.41 * 0.37) = 0.157$, $EM_5 = (0.49 * 0.40 + 0.51 * 0.44) = 0.462$, $EM_6 = (0.39 * 0.04 + 0.61 * 0.04) = 0.04$. Applying formula (2.3) $REM = \max(EM_i) = 0.462 = EM_5$ reveals that the utility function is maximized for the first emotion. Therefore, the combination of information from the audio-lingual and visual-facial modalities recognizes the correct emotional state (anger) for the user interacting with the system. Table 2.9 illustrates the conclusive normalized results of emotion recognition, after the application of MADM, taking into consideration weights that resulted from the analysis of our empirical studies.

In view of the above, every time the user interacts with the system and an emotion is expected to occur, the values of the utility functions for all the six emotions are estimated (EM_1 – EM_6) and the one that maximizes the utility function is selected as the prevailing emotion. At this point we should emphasize the fact that, in some cases, a modality may perform quite successfully in recognizing a certain emotional state based on uni-modal data, while information from another modality may seem unusable. However, the objective purpose of an emotion recognition system is to provide considerable and accurate information for a set of emotional states of users. In our case (six discrete emotional states), the two modalities complement each other to a large extent in their ability to recognize emotions. Clearly, properly combined complementary information from several modalities statistically improves the overall emotion recognition accuracy of a system.

Table 2.9 Resulting emotion recognition from the combined bi-modal data

	Video (%)	Audio (%)	Bi-modal result	Bi-modal normalized (%)
Neutral	10	6	9.12	10.03
Happiness	1	6	3	3.30
Sadness	44	3	16.94	18.64
Surprise	1	37	15.76	17.36
Anger	40	44	42.04	46.27
Disgust	4	4	4	4.40

2.8 Discussion and Conclusions

In this chapter, we have described and discussed a novel approach towards combining two modalities in bi-modal affect recognition using Multi-Attribute Decision Making (MADM). The bi-modal interaction consists of the visual-facial and the audio-lingual modalities. The main focus of this chapter has been on two empirical studies that we conducted concerning the two modalities and linking the results to the integration of the modalities through MADM.

The empirical studies constitute an important milestone for our approach described in this chapter. However, the settings and results of our empirical studies provide an important research investigation and results for other researchers to use or to compare with their own results. The field of affect recognition is not currently well understood and there are not many results of empirical work yet in the literature. Specifically, our empirical studies provide results as to how affect recognition is achieved via the visual-facial and audio-lingual modalities for the perspective of a human observer rather than physiological signals [56]. The two modalities are to a high extent complementary to each other and, thus, can be used in a bi-modal affective computer system designed to perform affect recognition taking into account the strengths and weaknesses of each modality.

From the empirical studies, we found that certain emotion states such as neutral and surprise are more clearly recognized from the visual-facial modality rather than the audio-lingual one. Other emotion states, such as anger and disgust are more clearly recognized from the audio-lingual modality. These results are only partially in accordance with a previous empirical study [8] which, however, did not take into account the lingual aspect in the audio modality. This is not surprising, since certain words convey emotions and this constitutes evidence that can be added to such information as the pitch and volume of the voice. Moreover, we have constructed a basic affective vocabulary for the Greek language that can be used in the audio-lingual modality and a database of facial expressions of emotions.

In our research so far, we have performed significant research work leading to the construction of a bi-modal affect recognizer. As a plan for the near future we will conduct extensive experiments for the evaluation of our affect recognizer.

References

1. Leon, E., Clarke, G., Callaghan, V., Sepulveda, F.: A user-independent real-time emotion recognition system for software agents in domestic environments. *Eng. Appl. Artif. Intell.* **20**, 337–345 (2007)
2. Goleman, D.: *Emotional Intelligence*. Bantam Books, New York (1995)
3. Picard, R.W.: Affective computing: challenges. *Int. J. Hum. Comput. Stud.* **59**, 55–64 (2003)
4. Stathopoulou, I.-O., Tsihrintzis, G.A.: Visual affect recognition. *Front. Artificial Intell. Appl.* **214**, 1–267 (2010)
5. Morrison, D., Wang, R., De Silva, L.C.: Ensemble methods for spoken emotion recognition in call-centres. *Speech Commun.* **49**, 98–112 (2007)
6. Pantic, M., Rothkrantz, L.J.M.: Toward an affect-sensitive multimodal human-computer interaction. *Proc. IEEE* **91**, 1370–1390 (2003)

7. Chen, L.S., Huang, T.S., Miyasato, T., Nakatsu, R.: Multimodal human emotion/expression recognition. In: Proceedings of the 3rd International Conference on Face & Gesture Recognition: IEEE Computer Society (1998)
8. De Silva, L., Miyasato, T., Nakatsu, R.: Facial emotion recognition using multimodal information. In: IEEE International Conference on Information, Communications and Signal Processing (ICICS'97), pp. 397–401 (1997)
9. Huang, T.S., Chen, L.S., Tao, H.: Bimodal emotion recognition by man and machine. In: ATR Workshop on Virtual Communication Environments Kyoto, Japan (1998)
10. Oviatt, S.: User-modeling and evaluation of multimodal interfaces. In: Proceedings of the IEEE, pp. 1457–1468 (2003)
11. Zeng, Z., Tu, J., Liu, M., Huang, T., Pianfetti, B., Roth, D., Levinson, S.: Audio-visual affect recognition. *IEEE Trans. Multimed.* **9**, 424–428 (2007)
12. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S.: Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proceedings of the 6th International Conference On Multimodal Interfaces, State College, PA, USA, ACM (2004)
13. Liao, W., Zhang, W., Zhu, Z., Ji, Q., Gray, W.D.: Toward a decision-theoretic framework for affect recognition and user assistance. *Int. J. Hum.-Comput. Stud.* **64**, 847–873 (2006)
14. Alepis, E., Stathopoulou, I.-O., Virvou, M., Tsihrintzis, G., Kabassi, K.: Audio-lingual and visual-facial emotion recognition: Towards a bi-modal interaction system. In: Proceedings of International Conference on Tools with Artificial Intelligence, ICTAI, 2, art. no. 5670096, pp. 274–281 (2010)
15. Fishburn, P.C.: Additive utilities with incomplete product set: applications to priorities and assignments. *Oper. Res.* **15**, 537–542 (1967)
16. Nasoz, F., Lisetti, C.L.: MAUI avatars: mirroring the user's sensed emotions via expressive multi-ethnic facial avatars. *J. Vis. Lang. Comput.* **17**, 430–444 (2006)
17. Virvou, M., Kabassi, K.: Adapting the human plausible reasoning theory to a graphical user interface. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* **34**(4), 546–562 (2004)
18. Alepis, E., Virvou, M.: Object oriented design for multiple modalities in affective interaction. *Intell. Syst. Reference Libr.* **64**, 87–99 (2014)
19. Kabassi, K., Virvou, M.: A knowledge-based software life-cycle framework for the incorporation of multicriteria analysis in intelligent user interfaces. *IEEE Trans. Knowl. Data Eng.* **18**, 1265–1277 (2006)
20. Naumann, F.: Data fusion and data quality. In: Proceedings of the New Techniques and Technologies for Statistics (1998)
21. Kabassi, K., Virvou, M.: A knowledge-based software life-cycle framework for the incorporation of multicriteria analysis in intelligent user interfaces. *IEEE Trans. Knowl. Data Eng.* **18**(9), 1265–1277 (2006), art. no. 1661516
22. Schütz, W., Schäfer, R.: Bayesian networks for estimating the user's interests in the context of a configuration task. In: UM2001 Workshop on Machine Learning for User Modeling, pp. 23–36 (2001)
23. Bohnenberger, T., Jacobs, O., Jameson, A., Aslan, I.: Decision-theoretic planning meets user requirements: enhancements and studies of an intelligent shopping guide. In: Pervasive Computing, pp. 279–296 (2005)
24. Chin, D.N., Porage, A.: Acquiring user preferences for product customization. In: Proceedings of the 8th International Conference on User Modeling 2001, Springer, Berlin (2001)
25. Kudenko, D., Bauer, M., Dengler, D.: Group decision making through mediated discussions. In: *User Modeling*, 2003, pp. 147–147
26. Vincke, P.: *Multicriteria Decision-Aid*. Wiley, New York (1992)
27. Triantaphyllou, E., Mann, S.H.: An examination of the effectiveness of multi-dimensional decision-making methods: a decision-making paradox. *Decis. Support Syst.* **5**, 303–312 (1989)
28. Ortony, A., Clore, G., Collins, A.: *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge (1998)
29. Ekman, P., Friesen, W.V.: *Unmasking the Face. A Guide to Recognizing Emotions from Facial Clues*. Prentice-Hall, Englewood Cliffs (1975)

30. Ekman, P., Friesen, W.V.: *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto (1978)
31. Ekman, P.: *Emotion in the Human Face*. Cambridge University Press, New York (1982)
32. Ekman, P., Levenson, R.W., Friesen, W.V.: Autonomic nervous system activity distinguishes between emotions. *Science* **221**(4616), 1208–1210 (1983)
33. Ekman, P., Davidson, R.J.: *The Nature of Emotion. Fundamental Questions*. Oxford University Press Inc., Oxford (1994)
34. Ekman, P.: Basic emotions. In: Dalglish, T., Power, M.J. (eds.) *Handbook of Cognition and Emotion*. Wiley, Sussex (1999)
35. Tomkins, S.S.: Affect as amplification: some modifications in theory. In: Plutchik, R., Kellerman, H. (eds.) *Emotion: Theory, Research and Experience*, vol. 1: *Theories of Emotion*. Academic Press, New York (1980)
36. Tomkins, S.S.: Script theory: differential magnification of affects. In: Howe, J.H.E., Dienstbier, R.A. (eds.) *Nebraska symposium on motivation*, vol. 26. Lincoln University of Nebraska Press, Lincoln (1999)
37. Plutchik, R.: *The Emotions: Facts, Theories, and a New Model*. Random House, New York (1962)
38. Plutchik, R.: A general psychoevolutionary theory of emotion. In: Plutchik, R., Kellerman, H. (eds.) *Emotion: Theory, Research and Experience*, vol. 1: *Theories of Emotion*. Academic Press, New York, pp. 3–33 (1980)
39. Oatley, K., Johnson-Laird, P.N.: Towards a cognitive theory of emotions. *Cogn. Emot.* **1**, 29–50 (1987)
40. McDougall, W.: *An Introduction to Social Psychology*. Luce and Co., Boston (1926)
41. Izard, C.E.: *The Face of Emotion*. Appleton-Century-Crofts, New York (1972)
42. Izard, C.E.: *Human Emotions*. Plenum, New York (1977)
43. Frijda, N.: *The Emotions*. Cambridge University Press, New York (1987)
44. Arnold, M.B.: *Emotion and Personality*. Columbia University Press, New York (1960)
45. Weiner, B.: An attributional theory of achievement motivation and emotion. *Psychol. Rev.* **92**, 548–573 (1985)
46. Martinez A.M.: *The AR face database*, CVC Technical Report (1998)
47. Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J.: Coding facial expressions with gabor wavelets. In: *Proceedings of the 3rd International Conference on Face and Gesture Recognition*. IEEE Computer Society (1998)
48. The Yale Database. <http://cvc.yale.edu/projects/yalefaces/yalefaces.html>
49. Kanade, T., Tian, Y., Cohn, J.F.: Comprehensive database for facial expression analysis. In: *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000*: IEEE Computer Society (2000)
50. Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. In: *IEEE International Conference Multimedia and Expo (ICME'05)*, Amsterdam, The Netherlands (2005)
51. Stathopoulou, I.-O., Tsihrintzis, G.A.: Facial expression classification: specifying requirements for an automated system. In: *Knowledge-Based Intelligent Information and Engineering Systems*, pp. 1128–1135 (2006)
52. Alepis, E., Virvou, M., Kabassi, K.: Development process of an affective bi-modal Intelligent Tutoring System. *Intell. Decis. Technol.* **1**, 1–10 (2007)
53. Stathopoulou, I.-O., Tsihrintzis, G.: Emotion recognition from body movements and gestures. In: *Smart Innovation, Systems and Technologies*, 11 SIST, pp. 295–303 (2011)
54. Lampropoulos, A.S., Stathopoulou, I.-O., Tsihrintzis, G.A.: Comparative performance evaluation of classifiers for facial expression recognition. *Stud. Comput. Intell.* **226**, 253–263 (2009)
55. Stathopoulou, I.-O., Tsihrintzis, G.A.: NEU-FACES: a neural network-based face image analysis system. In: *Proceedings of the 8th International Conference on Adaptive and Natural Computing Algorithms, Part II Warsaw*. Springer, Poland (2007)
56. Stathopoulou, I.-O., Alepis, E., Tsihrintzis, G., Virvou, M.: On assisting a visual-facial affect recognition system with keyboard-stroke pattern information. *Knowl.-Based Syst.* **23**(4), 350–356 (2010)

Chapter 3

Cooperative Learning Assisted by Automatic Classification Within Social Networking Services

Christos Troussas, Maria Virvou and Kurt Junshean Espinosa

Abstract Social networking services tend to promote social relations among people sharing interests, activities, backgrounds, or real-life connections. They consist of a representation of each user, namely a profile, along with of his/her social links and a variety of additional information. Such kind of information can be used as a means for highlighting the educational aspect of social networking services. To this direction, the authors of this chapter used Facebook as a testbed for this research and implemented a multi-language learning application. The main focus of this chapter is on the automatic classification of Facebook users, utilizing the aforementioned application, based on their profiles. In that way, coherent user clusters can be created and efficient cooperative learning among them can be achieved. Finally, the educational process can be further ameliorated given that cooperation among user clusters is recommended by the system.

Keywords Automatic classification · Facebook · Social networking services · Social networks · Computer assisted language learning · Intelligent tutoring systems

3.1 Introduction

In the recent few years, the field of education tends to be increasingly enriched by computer-assisted methods. Currently, Social Networking Services (SNSs) are being adopted rapidly by millions of users worldwide, most of whom are students [1]. The

C. Troussas (✉) · M. Virvou
Software Engineering Laboratory, Department of Informatics,
University of Piraeus, 125, Grigoriou Lampraki Street, 18532 Piraeus, Greece
e-mail: ctrouss@unipi.gr

M. Virvou
e-mail: mvirvou@unipi.gr

K.J. Espinosa
Department of Computer Science, University of the Philippines Cebu,
Gorordo Avenue, Lahug, 6000 Cebu, Philippines
e-mail: kpespinosa@up.edu.ph

SNSs gradually promote new trends of learning through the majority of advantages that they offer. The use of SNSs in instructional contexts is a hot matter in recent scientific literature; however, there is a long way to this direction. The use of SNSs, such as Facebook, in instructional contexts can be regarded as a potentially potent educational method given that students tend to spend anyway a lot of their spare time on these online networking activities [2]. However, the use of SNSs for educational reasons along with the modeling of students, who are trying to achieve significant educational outcomes, is not yet well explored.

Major developments have been observed in the areas of telecommunications and transportation. These significant advances have permitted the rise of the phenomenon of globalization, by which regional economies, societies, and cultures have become integrated through a global network of people, a fact that necessitates foreign language learning. Furthermore, European Union accentuates the need of learning foreign languages and thus students from Europe need to invest a good deal of time on these educational activities. As a result, the educational applications are addressed to very large and heterogeneous audience, namely learners of different background and with different needs [3]. For this reason, the instruction process should be characterized by high adaptivity and should render students capable of comprehending and dealing with large cognitive material.

Cooperative learning has important instructional benefits promoting students' interpersonal relationships and social interaction with their peers. Hence, the tutoring process can be considerably assisted by cooperation among students. When adaptive individualized e-learning systems could stimulate the tutoring process by disclosing the virtues and shortcomings of each learner, they could dynamically design the curriculum and individualize the understanding and didactic strategy. Machine learning techniques are utilized for obtaining models of individual users interacting with tutoring systems and assemble them into communities or stereotypes with common characteristics, so that the students reap the benefits of cooperation. Cooperative learning concerns the formation and construction of correlative data sets, based on inquiring large databases. The emerging correlations and patterns can shed light on specific characteristics of already existing group, or they can form the construction of the cluster/group/category itself.

In view of the above, the automatic classification for cooperative learning among Facebook users is presented. Automatic classification achieves to efficiently create user clusters. After the aforementioned user clusters have been created, the system recommends cooperation between them. As such, every cluster cooperates with another cluster in a way that both of them gain knowledge and promote their learning as opposed to helping only a specific user cluster to the detriment of others.

This chapter is organized as follows. First, related scientific work is presented. In Sect. 3.3, the K-means algorithm is discussed, which is in the basis of system's architecture. Following, a description of our Facebook application, namely a general overview accompanied by screenshots of the system, is situated in Sect. 3.4. In Sect. 3.5, the evaluation results concerning the automatic classification of our Facebook learning application are presented. Finally, Sect. 3.6 involves a discussion about the usability of the resulting system with future plans.

3.2 Related Work

In this section, the related scientific work, related, first, to social networking and, second, to Intelligent Computer Assisted Language Learning (ICALL), is presented.

3.2.1 *Social Networking Services*

The advent of SNSs has made an impact in various areas including educational context [4]. Many researchers investigated the use of SNSs such as Facebook in the context of learning, specifically language learning.

In [5], the authors discussed the effects of the inclusion of a Facebook activity as part of the language classroom. They examined students' feedback toward the activity and the use of language in their interaction. Based on their findings, the said activity proved to be advantageous because it kept the students connected with their friends while they take the language course.

In [6], the authors investigated the benefits of SNS communities (i.e. mixi and Facebook) for learning Japanese as a second language (L2). They found out that the SNSs has provided "a portal for L2 learners to access other information and sources, and present a safe introduction to wider communication in the L2". Seeing the usefulness of the SNSs communities as a learning tool, they recommend that "researchers and educators should direct their attention to such new tools, and the communities that form around them".

In [7], the authors have found out that simple activities in Facebook had helped the less language-proficient students to become more actively involved in the language learning process. In [8], the authors investigated the possibility of gathering learners' experience and views pertaining to issues on English language learning problems in secondary school, college or university, gathering learners' views on English language teaching and learning in secondary school, college and university and gathering learners' suggestions on ways to enhance English language learning and teaching.

In [9], the authors focused on the utilization of Facebook to enhance students' interaction using English in the English Speaking Zone. In undertaking this study, the authors had two research objectives: to understand the ways that social network namely Facebook can be used to provide the students with opportunities to use English as a medium of communication, and to reflect on and improve our own practice with an aim to create a Facebook environment that can continue to be used to enhance opportunities to use the language. In [10], the authors proposed an educational application in Facebook for learning the grammatical phenomenon of conditionals. This application is addressed to Greek students, who want to learn conditionals in Filipino and vice versa. The users are modeled so that they can receive advice from the system.

In all the aforementioned scientific literature, it can be assumed that the SNSs possibly played a crucial role in language learning because it provides a good interactive environment for learning. However, none of these researches took advantage of the automatic classification serving for the recommendation of profitable cooperation between user clusters.

3.2.2 Intelligent Computer-Assisted Language Learning

In [11], the authors proposed the exploitation of machine learning techniques to improve and adapt the set of user model stereotypes by making use of user log interactions with the system. To do this, a clustering technique is exploited to create a set of user models prototypes; then, an induction module is run on these aggregated classes in order to improve a set of rules aimed as classifying new and unseen users. Their approach exploited the knowledge extracted by the analysis of log interaction data without requiring an explicit feedback from the user.

In [12], the author presented a snapshot, of what has been investigated in terms of the relationship between Machine Translation (MT) and Foreign Language (FL) teaching and learning. Moreover, the author outlined some of the implications of the use of MT and of free online MT for the FL learning. In [13], the authors investigated, which human factors are responsible for the behavior and the stereotypes of digital libraries users, so that these human factors can be justified to be considered for personalization. To achieve this aim, the authors have studied, if there is a statistical significance between the stereotypes created by robust clustering and each human factor, including cognitive styles, levels of expertise, and gender differences.

In [14], the authors focused on machine learning approaches for inducing student profiles, based on Inductive Logic Programming and on methods using numeric algorithms, to be exploited in this environment. Moreover, an experimental session has been carried out from the authors, comparing the effectiveness of these methods along with an evaluation of their efficiency in order to decide how to best exploit them in the induction of student profiles.

In [15], the authors explained that user modeling poses a number of challenges for machine learning that have hindered its application in user modeling, including the need for large data sets, the need for labeled data, the concept drift, and computational complexity. In [16], the authors constructed a learning agent that models student behavior at a high level of granularity for mathematics tutor, by using traces from previous users of the tutor to train the machine learning agent. In [17], the authors described the implementation of student modeling through machine learning techniques, which aims to ameliorate future multiple language learning systems.

However, after a thorough investigation in the related scientific literature, one came up with the result that there is not any application in social networking services, including Facebook, which concerns automatic classification for profitable

cooperative learning of foreign languages. Thus, we designed and implemented a prototype application, which automatically classify Facebook users into clusters and recommends cooperation among these clusters.

3.3 Algorithm of the System Functioning

The K-means algorithm is a clustering algorithm, which is used to classify or to group given objects based on specific attributes or parameters into K number of groups. K is a positive integer number [18–20]. The K-means algorithm is the most popular clustering algorithm. It is an iterative algorithm, which has basically two steps: cluster assignment step and move centroid step.

3.3.1 Description of Automatic Classification

The K-means algorithm takes two inputs:

1. K which is the number of clusters that you want to find in the data.
2. Unlabeled training set $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$, where $x^{(i)}$ is an element of R^n .

Basically in formal form:

Randomly initialize K cluster centroids $\mu_1, \mu_2, \dots, \mu_k$ element of R^n .

Repeat {

for $i = 1$ to m

$c(i) :=$ index (from 1 to K) of cluster centroid closest to $x^{(i)}$

for $k = 1$ to K

$\mu_k :=$ average (mean) of points assigned to cluster k

} until convergence (μ_k does not change)

3.3.2 Optimization Objective and Its Definition

Given the following notations,

- $c^{(i)}$ is an index of cluster (1, 2, ..., K), to which example $x^{(i)}$ is currently assigned,
- μ_k is a cluster centroid k (μ_k element of R^n),
- $\mu_{c^{(i)}}$ is a cluster centroid of cluster, to which example $x^{(i)}$ has been assigned,

the optimization objective is Eqs. 3.1–3.4.

$$J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K) = \frac{1}{m} \sum_{i=1}^m \left\| x^{(i)} - \mu_{c^{(i)}} \right\|^2, \quad (3.1)$$

$$\min J\left(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K\right), \quad (3.2)$$

$$c^{(i)}, \dots, c^{(m)}, \quad (3.3)$$

$$u_1, \dots, u_K. \quad (3.4)$$

The cluster assignment step does the minimization of the cost function (also called distortion function) while the “move centroid step” chooses the value of μ that minimizes the cost function J with respect to the location of cluster centroids u_1, \dots, u_K .

3.3.3 Initialization of Centroids

At the start of K-means algorithm, the cluster centroids are randomly initialized. Given $K < m$, randomly pick K training examples. Then set u_1, \dots, u_K equal to these K examples. However, this could also lead to local optima. But this can be addressed by running multiple initializations. Particularly,

```

for  $i = 1$  to  $R$  {
  Randomly initialize K-means
  Run K-means. Get  $c^{(i)}, \dots, c^{(m)}, u_1, \dots, u_K$ .
  Compute cost function (distortion)
   $\min J(c^{(i)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K)$ 
}

```

where R is a number of K-means execution (e.g. 100, 1000).

Then pick clustering that gave the lowest cost $J(c^{(i)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K)$.

3.3.4 Incorporation of Automatic Classification

For the incorporation of the automatic classification into the resulting Facebook language application (Fig. 3.1), let us make the following basic steps:

- For the initialization of the system the algorithmic techniques receive as input, pre-stored data or data from empirical studies. In our system, we have used several fundamental characteristics, which in accordance with the authors’ expertise in the domain tend to influence the educational procedure:
 - Age.
 - Score at the preliminary test.
 - Gender.
 - Foreign language knowledge.
 - Levels of education.

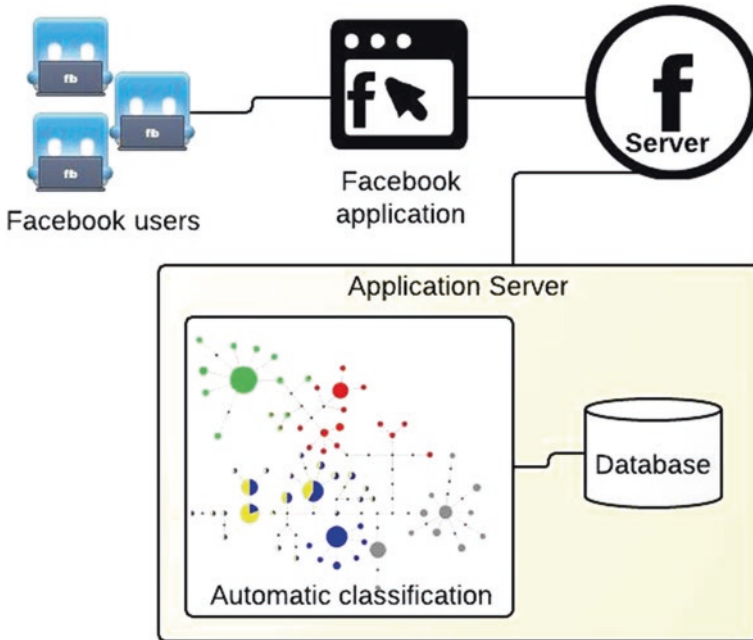


Fig. 3.1 Automatic classification

- Working experience.
- Computer knowledge.

Using the prototype application, the following characteristics were extracted from each user. Basically, all of them except age, score, and duration of computer use, were gathered from their Facebook profile. Following, a general reasoning for the use of several user characteristics is presented:

- Gender. We would like to examine the degree of differentiation in learning between the genders.
 - Foreign language knowledge. This particular data gives significant information concerning the language learning ability of the user. It is widely acknowledged that the more languages the user knows, the more apt s/he is in learning a new one.
 - Levels of education. This characteristic includes all the levels of education the user has attended. The hypothesis lies in the fact that the language learning ability is proportionate to the educational attainment.
 - Working experience. This data is indicative of the maturity of the user and thus can reflect his/her experience in learning new things, such as a new language.
- Machine learning techniques are used as a next step in order to describe efficiently the circumstances that underlie the student's actions in terms of their behavioral patterns and preferences.

Based on the aforementioned characteristics, the system creates clusters of the already existing students. These clusters contain valuable information about their members, considering their behavior, their preferences and generally their interaction with the system.

Figure 3.1 illustrates the general overview of our system. In particular, Facebook users study foreign languages by utilizing the Facebook application. The application server contains all the significant components of our system, including the automatic classification and a database.

3.4 General Overview of the System

Figure 3.2 illustrates the preliminary test, which attempts to evaluate the student’s understanding and mastery of the subject matter. It uses the English language as the medium of instruction so that the student will be evaluated solely on his/her knowledge state in the application domain. Particularly, in this figure the student is asked to identify the conditional type of the sentence structure presented. Correspondingly, he/she then selects the correct answer from the given choices. In this way, the system attempts to cluster the students in order to better assist them in the educational process.

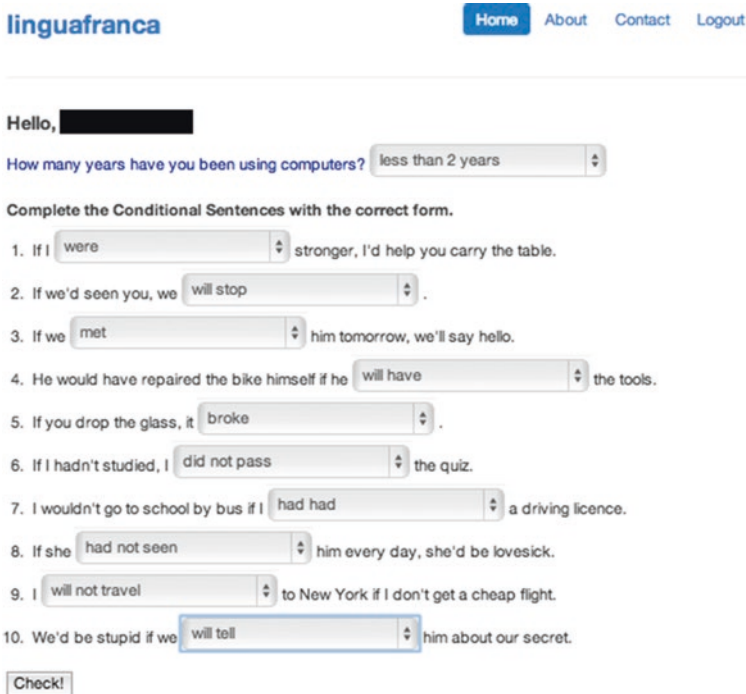


Fig. 3.2 Preliminary test

Figure 3.3 illustrates the clustering procedure, which is performed by the K-means algorithm. There are three clusters created, based on the aforementioned user characteristics, which serve as input of the clustering algorithm. In this way, each user knows exactly, in which cluster s/he belongs and has the possibility of collaborating with his/her peers (users of the same cluster) through instant or asynchronous messaging through the Facebook platform.

Figure 3.4 indicates how the database of application server looks like. The black boxes, as appeared in Figs. 3.2 and 3.3, are used in order to ensure anonymity and conceal the names of real Facebook users, who used our language learning application.

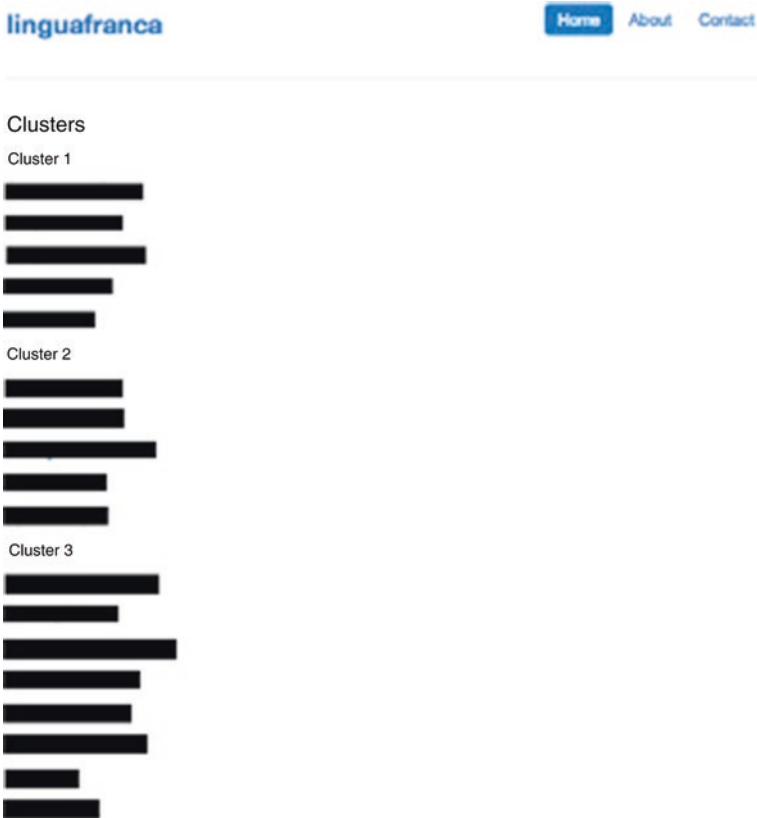


Fig. 3.3 Clustering based on multiple variables

id	first_name	last_name	facebook_id	email	age	score	link	cluster	sex	cnt_lang	cnt_educ	cnt_work	pc_lit
10				@gmail.com	29	2		1	male	3	3	4	2
37				@yahoo.com	18	1	http://www.facebook.com/	3	male	2	2	0	2
11				@yahoo.gr	26	8		2	male	4	4	2	2
31				@yahoo.com	24	6	http://www.facebook.com/	1	male	2	2	1	2
32				@yahoo.com	29	7	http://www.facebook.com/	2	male	2	2	2	2
31				@gmail.com	32	0	http://www.facebook.com/	1	male	2	2	4	2
24				@catholicpages.org	39	7	http://www.facebook.com/	2	male	2	2	1	2
23				@yahoo.com	35	6	http://www.facebook.com/	2	male	3	3	2	2
25				@yahoo.com	24	5	http://www.facebook.com/	1	male	2	3	3	2
27				@yahoo.com	22	0	http://www.facebook.com/	3	male	2	2	0	2
30				@yahoo.com	23	0	http://www.facebook.com/	3	male	2	2	2	2
28				@yahoo.com	29	5	http://www.facebook.com/	1	male	2	2	3	2
29				@yahoo.com	18	6	http://www.facebook.com/	3	male	3	3	2	2
34				@gmail.com	25	7	http://www.facebook.com/	2	male	2	2	1	1
41				@gmail.com	28	2	http://www.facebook.com/	0	male	4	3	2	2
45				@yahoo.com	21	0	http://www.facebook.com/	0	male	4	3	1	2

Fig. 3.4 Snapshot of the database in application server

3.5 Evaluation of the System

Computer-assisted learning, as the act of acquiring new or modifying and reinforcing, existing knowledge through computers can be considered successful, if it is approved by both software human instructors and students. Hence, evaluation of such kind of software is a crucial phase that has to follow development at all times. In particular, formative evaluation is one of the most critical steps in the development of learning materials. Formative evaluation is a range of formal and informal assessment procedures employed by instructors during the learning process in order to modify teaching and learning activities to improve student attainment. It typically involves qualitative feedback for both student and instructor focusing on the details of content and performance. As such, it helps the designer improve the cost-effectiveness of the software and increases the chance that the final product will achieve its fixed goals. The primary goals of educational software is teaching and/or self-learning in successful contexts. Hence, many there evaluation methods have been presented in the related scientific literature, which are completely tailored to educational software. This task is quite challenging, since such kind of applications are addressed to a very large audience.

One such evaluation framework outlines three dimensions to evaluate [16, 20]:

- Context. The context determines the reason, why the educational software is adopted in the first place, namely the underlying rationale for its development and use.
- Interactions. Users’ interactions with the software reveal information about the users’ learning processes.
- Attitudes and outcomes. This stage stage examines information from a variety of sources, such as pre- and post-achievement tests, interviews and questionnaires with students and tutors.

The underlying rationale of our system involves providing educational features, since educational systems provide opportunities within professional education, curriculum education, and learning. Moreover, the context of the evaluation required an emphasis on the automatic classification for effective cooperation among Facebook users, as students. Moreover, users’ interactions with the application were

evaluated with respect to the users' learning processes. Finally, the "outcomes" stage involved pre- and post-achievement tests before and after the use of the application. In addition, it involved questions to students and instructors, which focused mainly on evaluating the use of the aforementioned framework in the application.

In view of the above, the evaluation of our system involved both instructors and students and was conducted in two different phases. At the first phase, only the instructors took place at the evaluation. The second phase concerned the evaluation of the resulting educational application and involved both instructors and students. The instructors of the second phase were exactly the same as in the first phase, so that they could have a complete experience with our language learning application. At the first and second phase, three instructors participated in the evaluation. All of them are private instructors of the English language and were asked to use our system, namely to check in detail the tutoring of the multiple languages through Facebook. All of the instructors, who participated in the experiment, were familiar with the use of computers.

When asked, all of the instructors confirmed that our language learning application had a user-friendly interface and that the automatic classification process was satisfactory. More specifically, two of them stated that they found the application very useful, while one of them stated that the application is useful. Concerning the question "How do you see the possibility of the system to automatically classify users and are you satisfied with the recommendation of the system concerning the cooperation?", the instructors stated that they were all very satisfied from the possibility of the system to automatically classify users into clusters and to propose cooperation between them.

The second phase involved in total 40 users, from which 20 users are undergraduate students and 20 users are postgraduate students of Computer Science at the University of the Philippines. The underlying rationale of the application lies on the hypothesis that these applications are more convenient and flexible to use (because of the fact that Facebook is easily accessible). At a first glance, the validity of this hypothesis might look obvious. However, there may be users, who are not familiar with educational software in general or with the Facebook as a SNS in particular, and thus they might not like such particular applications. On the other hand, there may be students, who are very familiar with computers and/or Facebook and are eager to use them for educational purposes. Hence, one important aspect of the evaluation was to find out, whether users were indeed helped by the whole learning environment.

After the completion of the use of the application, the Facebook users were asked about their experience. 45 % (18 students) found very satisfactory the automatic classification conducted by the system and, hence, they declared that the recommended cooperation was very profitable and successful. 30 % (12 students) found satisfactory the automatic classification conducted by the system and, hence, they declared that the recommended cooperation was profitable and successful. 25 % (10 students) did not find satisfactory the automatic classification conducted by the system. Moreover, the students were very pleased and enthusiastic by the opportunity that our application gave them the possibility to collaborate with their peers, namely the members of the same cluster. Figure 3.5 illustrates a pie-chart, representing this result (users' answers).

User Satisfaction from automatic classification

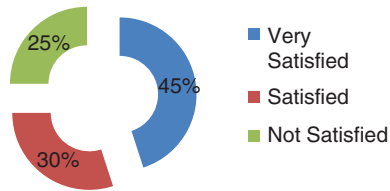


Fig. 3.5 Rate of satisfaction from automatic classification

3.6 Conclusions and Future Work

In this chapter, the automatic classification for multiple languages learning in Facebook had been presented. Our application combines the language learning with the cooperation of Facebook users. Automatic classification is conducted and the result of this procedure serves as an input for the system so that it recommends cooperation among user clusters. Furthermore, automatic classification takes as input, to initialize the process, multiple factors, namely multiple user characteristics in order to create dynamic clusters and enhance the cooperation among them. Our approach presented in this chapter exploits the fact that Facebook has a large number of users, and we use automatic reasoning mechanisms based on recognized similarities among them.

It is in our future plans to further evaluate our Facebook multilingual learning application in order to examine the degree of usefulness of the automatic classification serving for the cooperation between user clusters. Furthermore, we are planning to further evaluate the degree of usefulness of cooperation, offered by the resulting system by asking more people to use the aforementioned application and for a longer period.

References

1. Lenhart, S., Madden, M.: Teens, privacy, and online social networks. In: Pew Internet and American Life Project Report (2007)
2. Mazman, S., Usluel, Y.: Modeling educational usage of Facebook. *Comput. Educ.* **55**(2), 444–453 (2010)
3. Jones, A., Scanlon, E., Tosunoglu, C., Morris, E., Ross, S., Butcher, P., Greenberg, J.: Contexts for evaluating educational software. *Interact. Comput.* **11**(5), 499–516 (1999)
4. Social Networking Service. http://en.wikipedia.org/wiki/Social_networking_service. Accessed 08 Feb 2015

5. Piriyaasilpa, Y.: See you in Facebook: the effects of incorporating online social networking in the language classroom. *J. Global Manage. Res.* 67–80 (2011). http://www.itdl.org/journal/jan_09/article02.htm
6. Ota, F.: *A Study of Social Networking Sites for Learners of Japanese*. Japan Foundation Sydney, Sydney (2011)
7. Promnitz-Hayashi, L.: A learning success story using Facebook. *Stud. Self-Access Learn. J.* 2(4), 309–316 (2011)
8. Hiew, W.: English language teaching and learning issues in Malaysia, learners' perceptions via Facebook dialogue journal. *J. Arts, Sci. Commer. (Res World)* 3(1), 11–19 (2012)
9. Ho-Abdullah, I., Hashim, R.S., Jaludin, A., Ismail, R.: Enhancing opportunities for language use through web-based social networking. In: *International conference on social science and humanity*. IACSIT Press, Singapore, pp. 136–139 (2011)
10. Virvou, M., Troussas, C., Caro, J., Espinosa, K.J.: User modeling for language learning in Facebook. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) *Text, Speech and Dialogue*, LNCS, 7499, pp. 345–352. Springer, Berlin (2012)
11. Basile, T., Esposito, F., Ferilli, S.: Improving user stereotypes through machine learning. In: Agosti, M., Esposito, F., Meghini, C., Orio, N. (eds.) *Digital Libraries and Archives*, CCIS, 249, pp. 38–48. Springer, Berlin (2011)
12. Niño, A.: Machine translation in foreign language learning: language learners and tutors perceptions of its advantages and disadvantages. *ReCALL* 21(2), 241–258 (2009)
13. Friaiz-Martinez, E., Chen, S., Macredie, R., Liu, X.: The role of human factors in stereotyping behavior and perception of digital library users: a robust clustering approach. *User Model. User-Adap. Interact.* 13, 305–337 (2007)
14. Licchelli, O., Basile, T., Di Mauro, N., Esposito, F., Semeraro, G., Ferilli, S.: Machine learning approaches for inducing student models. In: Orchard, B., Yang, C., Ali, M. (eds.) *Innovations in Applied Artificial Intelligence*, LNAI, 3029, pp. 935–944. Springer, Berlin (2004)
15. Webb, G., Pazzani, M., Billsus, D.: Machine learning for user modeling. *User Model. User-Adap. Interact.* 11, 19–29 (2001)
16. Beck, J., Woolf, G.: High-level student modeling with machine learning. In: Gauthier, G., Frasson, C., VanLehn, K. (eds.) *Intelligent Tutoring Systems*, LNCS, 1839, pp. 584–593. Springer, Berlin (2000)
17. Virvou, M., Troussas, C., Alepis, E.: Machine learning for user modeling in a multilingual learning system. In: *International Conference on Information Society, i-Society 2012*, Article number 6284978, pp. 292–297 (2012)
18. A Tutorial on Clustering Algorithms. http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html. Accessed 08 Feb 2015
19. Data Mining Algorithms In R/Clustering/K-means. http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Clustering/K-Means. Accessed 08 Feb 2015
20. Cluster analysis. http://en.wikipedia.org/wiki/Cluster_analysis. Accessed 08 Feb 2015

Chapter 4

Improving Peer-to-Peer Communication in e-Learning by Development of an Advanced Messaging System

Marian Cristian Mihăescu, Dumitru Dan Burdescu and Mihai Mocanu

Abstract This chapter presents an advanced messaging system, whose goal is to improve the peer-to-peer communication in e-Learning. The improvement is based on the ability of the developed system to produce information that is highly related to the informational needs of the person, who accesses it. The system is an intelligent one because it integrates a classification procedure for retrieval of the messages that have a high potential of being interesting to peers. It uses as input data activity logs obtained by monitoring the communication that takes place within the e-Learning platform. The main data analysis goal is to create a user's model, for which derived classes are in close relation with specific set of messages. The outcome is in the form of a tool that allows learners to receive a set of recommended messages that is highly to be interesting for them. The tool analyzes the user's features, classifies them and according with the class label obtained set of messages. The tool also acts as a message indexing system by storing messages in correlation with labels assigned to learners. A classical classification procedure is used for obtaining a labeling. The data used to train the classifier is gathered from the on-line educational environment and contains all the necessary information (i.e., the features) regarding the activities performed by learners on the platform. The high quality of the system is based also on a text-mining module that uses stemming, annotation, and concept detection for a proper assignment of messages to learner's labels.

M.C. Mihăescu (✉) · D.D. Burdescu · M. Mocanu
Department of Computers and Information Technology, University of Craiova,
Strada Alexandru Ioan Cuza 13, 200585 Craiova, Romania
e-mail: mihaescu@software.ucv.ro

D.D. Burdescu
e-mail: burdescu@software.ucv.ro

M. Mocanu
e-mail: mocanu@software.ucv.ro

Keywords Peer-to-Peer communication · Messaging system · e-Learning · Text mining · Classification

4.1 Introduction

This chapter addresses the problem of improving peer-to-peer communication among learners by usage of intelligent retrieval of messages. The system is data driven and is based on the activity performed by the learners, when using the online educational environment. An increasing number of students are now using online educational environments. The benefits they receive include getting access to new and well-structured information, possibility to interact with other students via messages, the chance to get access to professors' knowledge and experience using the messaging system. The functionality of the system is preconditioned by the existence of an e-Learning environment that provides activity data and a custom designed data analysis system such as a trained classifier.

In our approach, learners will be placed in different classes according to several criteria such as results on tests, number of messages sent or received, time spent on the platform, etc. This process is regarded as classification or labeling. A key issue our peer-to-peer communication system is the available activity data for each learner, which must be logged by the educational environment. In order to accomplish this requirement the database and logs of Tesys [1], the e-Learning platform in used to obtain the learner's features and messages.

The data assets (e.g., database, log files, etc.) queried from the e-Learning platform must provide enough data in order to be able to create a proper data training set. The training set created from the gathered data must have enough relevant features to create a data model (i.e., the classifier). Regarding this issue there several key aspects, among which several important ones are: the number of features [2] describing a learner, the overall quantity of logged data, benchmarking data, meta-data issues, repeatability, etc. [3]. These prerequisite issues represent an invariant bottleneck, when running the data analysis system because they have direct influence over the obtained model and the quality of the results.

Once the classification model has been created, learners may improve their peer-to-peer communication by getting discussion messages that may be interesting without asking directly a peer (i.e., professor or colleague). The learner will be advised to read messages that were relevant to former learners, which were in similar activity pattern situation. All these data processes are integrated in a tool that runs along the online e-Learning platform. The tool uses the data provided by the e-Learning platform and integrates as core business logic a classification mechanism (i.e., decision tree learning).

For developing such a system an appropriate algorithm type (e.g., supervised, unsupervised, rule based, etc.) must be chosen, the algorithm itself, the features (e.g., name, meaning, type, values, etc.), and the setup of the environment in order to obtain optimal results. For building the user's model, the Decision Tree

induction algorithm [4] is used. This supervised algorithm can be used to classify new items, which in educational applications are students. In order to find them, the obtained model may be reliable and, thus, may be used to obtain interesting messages a quality evaluation and validation module is used.

Taking into consideration that most of the learner's scholar problems remain the same from generation to generation, there is a high probability that the learner's question might have already been answered and this answer can be delivered in an automatically manner.

In this chapter, the problem of improving peer-to-peer communication is tackled among learners by means of educational data mining, user modeling, and text mining. All these three activities are put together in an unifying framework that has as out the development of an advanced messaging system that implements the proposed architecture. The progress presented in this chapter regards the custom design of an effective intelligent data analysis system, which has as main beneficiary an e-Learning system, whose performance is being improved. Section 4.2 presents related work. A design of data analysis system is considered in Sect. 4.3. Experimental results are located in Sect. 4.4. Conclusions and future work are discussed in Sect. 4.5.

4.2 Related Work

In the past years, a lot of work has been done towards developing better solutions for virtual education environments. This interdisciplinary domain is represented by overlapping research areas like information retrieval, visual data analytics, psychopedagogy, artificial intelligence, data mining, and so on. Starting with user modeling, the educational data mining started to include many contributions from the areas of artificial intelligence, intelligent tutoring systems, and technology enhanced learning [5, 6].

In 2006, "Data mining in e-Learning" [7] appears as the first well documented monograph that states the fact that EDM covers the fields of Information Retrieval, Intelligent Agents, Data Mining, Data Warehouse, Text Mining, etc. Here are the first attempts in presenting an introduction to e-Learning systems, data mining and their interactions followed by several case studies experiences of applying data mining techniques in e-Learning systems.

In 2010 in "Handbook of Educational Data Mining" [8], this research area is well defined and within it, topics as visualization, data repositories, process mining, as well as a large number of case studies represent fundamental knowledge. The problem of student classification has been discussed in early 2003 by Minaei-Bidgoli [9] in a classical attempt to predict their final grade based on features extracted from logged data provided by an educational web-based system.

Similar studies were performed later in 2010 by Hamalainen and Vinni [10], in which they present a survey about the previous research on using data-driven classification for educational purposes, followed by a recall of the main principles affecting the model accuracy and finally giving several guidelines for accurate classification.

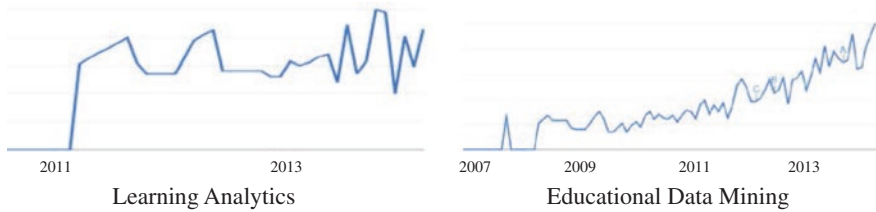


Fig. 4.1 Google trends search chart

No latter than 2012 Abu Tair and El-Halees [11] presented a case study describing how to preprocess the data and extract the features for graduate students. They also presented results and discussions about applications of data mining techniques to graduate students' dataset.

According to Google trends, one can see that that Learning Analytics started getting attention after 2007 and after 2013 it is still attracting more people as well as Educational Data Mining (Fig. 4.1).

One of the important moments that linked Data Mining and Educational Data was in 2012, when Blackboard Inc. [12] opened the field trial for learning analytics solution [13].

Another important research direction that is also discussed in this chapter regards user modeling as a main business processing mechanism for obtaining adaptivity for an e-Learning system. In this area, there are many already classical modeling techniques and corresponding evaluation techniques that led to currently existing adaptive e-Learning systems' architectures [14]. The proposed framework makes intensive use of text mining capabilities [15, 16]. In this area, Porter designed an algorithm, whose implementation is PTStemmer [17] and detailed comparative studies are presented by Sharma [18]. The framework heavily extends the functionality of an advanced messaging system [19].

The main technology that is used to implement the proposed system is Weka (Waikato Environment for Knowledge Analysis) [20], a popular suite of machine learning and data mining algorithms written in Java. The implemented algorithms are very flexible and can be used into the analyzing process of different kinds of data (from different domains). Weka has three main types of algorithms: supervised (e.g., decision trees, Bayesian networks [21], vector space classification [22]), unsupervised (e.g., partitioned, EM [23], fuzzy [24]), and rule based.

4.3 Data Analysis System Design

The chosen educational environment is Tesys, which is an online educational platform, where students, professors, and secretaries perform their responsibilities. Some of the current actions that take place in this e-Learning platform are taking tests and exams, downloading courses and communicating with professors.

All these performed activities are valuable for the data analysis process because from this platform one can gather the necessary data needed to train the classification algorithm.

The system consists of two types of processes, the server side and the client side processes. On the server part of the software system, the classification process is configured and then run, gathering data for the training set and generating the classification model. Among the utility processes, there are text processing/parsing, concept detection, and annotation. On server side, there are implemented four modules: the data gathering module, the model building module, the message indexing module, and the student classification and message retrieval module.

The first module will gather data from the database and the log files in order to generate the training files containing the values to all features that describe students. The module is responsible for actually creating the decision tree, which has in its internal nodes the features and in leaf nodes the class labels. The message indexing module inserts messages into the message index according with the class of the student, who wrote the message. The whole message is composed by the student's question along with professor's answer. The message retrieval module classifies the student and retrieves the associated messages with obtained class label. Its processing is based on a text mining capability. A stemming process similar with the ones integrated into search engines is used. Since the efficiency of the stemming for English language is somehow reduced the Information Retrieval (IR) procedure relies heavily on searching the n-grams. The Term Frequency-Inverse Document Frequency (TF-IDF) is used as a weighting mechanism for the IR and text mining activities. The current implementation sums TF-IDF value for each term in the query text. This simple approach was preferred to more sophisticated variants of this simple model. Thus, an embedded IDF factor diminishes the weight of the terms that appear with a high frequency in a document and increases the weight of the terms that rarely appear in documents. In this way, there is obtained a document indexing by extraction of terms that characterize that document. Weight assignment is accomplished with binary values: 1, when the term exists in the documents, and 0, when the term does not exist in the documents. Another implemented approach is to assign as weights to terms in the form of $TF \cdot IDF$. This measure is currently used in document collections that favor more frequent terms from relevant documents that are less frequent within the collection as a whole. As usual, the TF represents the rate frequency of meeting a term in the document and the IDF is the inverse of meeting a term in the entire collection.

From the business logic point of view, when a student sends a message to a professor and the professor answers accordingly, the message will be indexed/mapped to the class of the student. Once the message is indexed, it will be available upon request to students that have the same class label from the classifier's point of view.

For this prototype of the system, we have chosen a Decision Tree classifier and more exactly J48. The J48 [25] is the implementation of the C4.5 [26] algorithm in Weka, a data analysis algorithm, which generates a decision tree in order to classify data. Other popular decision tree algorithms are: Id3 [27], ADTree [28, 29],

BFTree [30], DecisionStump [31], NBTree [32]. In order to ensure that the right choice has maden, one can use a validation module included in Weka, which provides measures like accuracy, confusion matrix, kappa statistic, mean absolute error, etc.

Figure 4.2 presents the general data work-flow for message indexing and retrieval. The first module is responsible for parsing the raw data and transforms it into the training data and the associated messages repository. In this particular system design, the raw data is represented by the DBMS of Tesys e-Learning system and the log files of the platform. The outcome of the parsing process is represented by an *arff* file that contains the feature/parameters numerical values for all the activities performed by students. This is a structured file that is ready to be used for training Weka’s classifiers. Along with the *arff* file, there is obtained also a message repository, in which all the exchanged messages between learners are stored.

The outcome of the user modeling process is represented by a trained classifier. Once the classifier has been obtained it is ready to be augmented with messages. In this step, all messages are distributed among class labels according with sender/receiver’s class. In this way, the augmented classifier becomes the central business logic processing engine that is responsible for learner’s classification, message indexing, and message retrieval. From a software engineering point of view the augmented classifier extends the capabilities of the baseline classifier implemented in Weka.

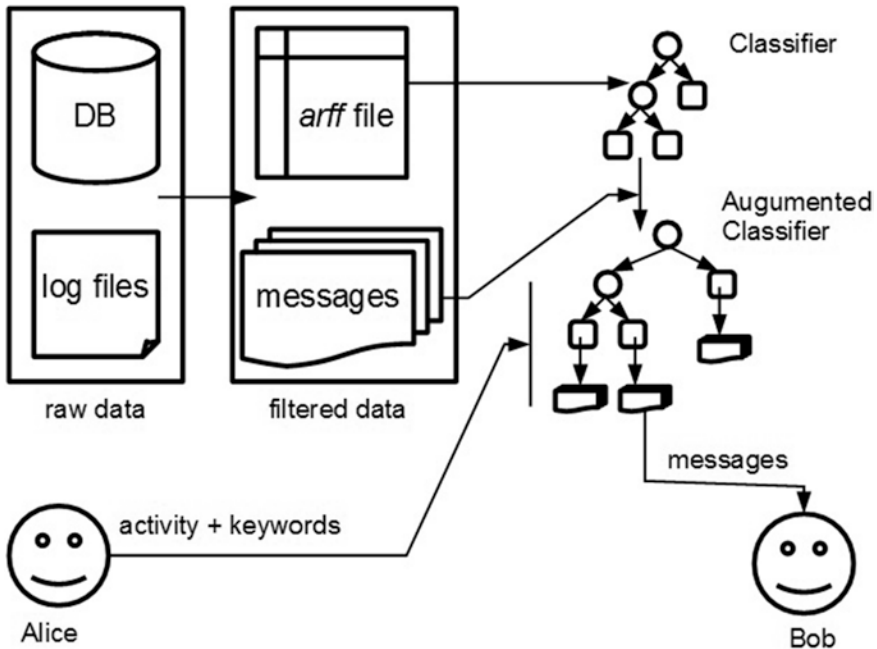


Fig. 4.2 General data workflow for message indexing and retrieval

The augmented classifier is a custom extension of J48 decision tree classifier implemented in Weka and its main capabilities regard accommodation of messages by associating them to leaves. In our particular case, each leaf represents a class of learners that has been inferred by the decision tree classifier from the training data.

The key fact, on which the business logic of the classifier relies, is related to the association between student labels/classes and exchanged messages between students from the same class. Intuitively, students belonging to the same class have a higher probability of sharing similar difficulties in the learning process and thus exchange messages discussing common issues. That is why, classification of learners is also used indirectly for classification of messages from the learner's level of knowledge point of view. This approach for message classification is, thereafter, used as the main mechanism for improvement of peer-to-peer communication between students with the same class label. This message classification system puts under the same label messages that address educational issues of similar gender and difficulty and, therefore, provides an effective way of making learners be as close as possible to their peers, when facing educational challenges.

The process of obtaining messages that may be recommended for a certain student begins with the classification of the student that is interested in obtaining messages. The student classification is performed as a classical machine learning application, in which the student is represented by several features, and the target variable is a nominal one and represents the accumulated knowledge level.

Once the learner's class is determined the associate messages that have the highest relevance may be delivered as recommendations. Those messages are the ones that were sent by students that were pretty much in the same educational situation. It means that those students have pretty much the same knowledge level and, therefore, there is a high probability of needing similar type of advice.

Another important processing activity, which is heavily used within the proposed framework, regards text mining and concept detection. Figure 4.3 presents a high level description of the main modules that implement these functionalities. The key idea is that on the corpus of all messages it is runned a text-mining and concept detection process, which will allow a proper message indexing process into the augmented classifier.

The most important step in this module is stemming. The stemming process is based on several grammatical rules used for putting aside suffixes, gender, plural, etc. One of the basic rules for correct stemming is that at least three characters make up a stem. The next step in concept detection is computing the frequency, by which each stem appear in text. The output of this step is represented by a list of stems and the number of its appearances. The TF-IDF formulas are used and the stems are arranged in non-decreasing order of TF-IDF values. The concepts along their TF-IDF values are serialized into a structured format that will be used later for improving the accuracy, by which the concepts are detected in new messages.

Putting together quiz questions, concepts, associated weights, and messages is performed by an on-line data analysis process that gathers the necessary input data and outputs the difficulty of the concept in correspondence with the number of

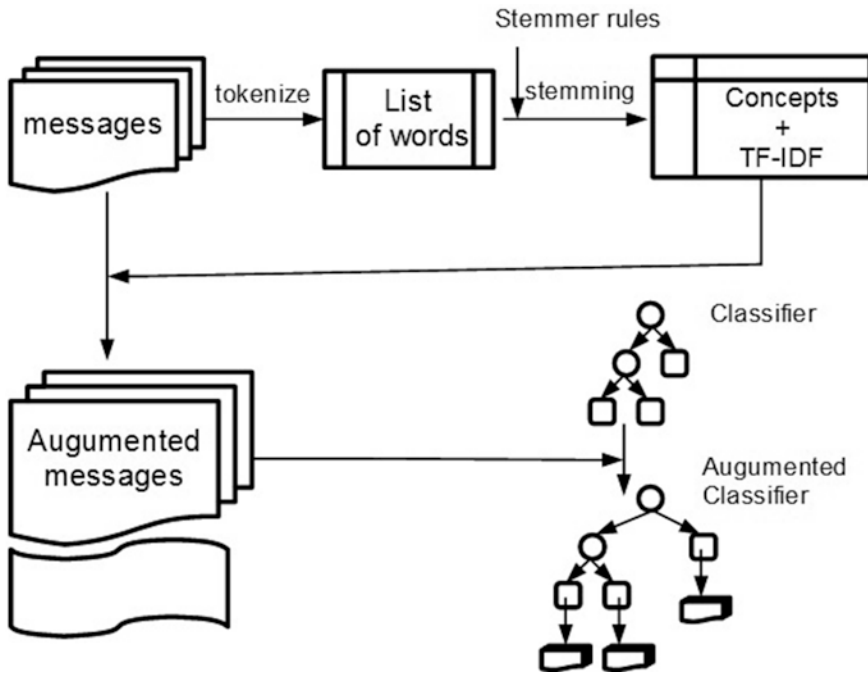


Fig. 4.3 Text stemming, concept detection and augmented classifier inference

questions, to which it is related. For each student, a table, in which coverage percentages per concept are computed according with the difficulty level derived for each concept, is obtained. The difficulty level for each question is thereafter used as an indicator variable in the classification process of learner. This classifier is, thereafter, augmented with corresponding messages.

4.4 Experimental Results

For the classification module, a data set gathered from the student’s activity repository was used. Tesys e-Learning platform is regarded as a raw data provider. The raw data (database tables, log files, etc.) residing on the server side are processed in order to obtain the benchmark data in the form of text files (i.e., arff, xml, etc. files). The output is represented by two *.arff* files: *tesysActivityData.arff* and *tesysMessages.arff*.

Attribute-Relation File Format (ARFF) is a file format of text file that describes a list of instances sharing a set of attributes in Weka. The two main text files for the data analysis are:

- **tesysActivityData.arff** file contains all the *identification*, *quiz*, *time*, and *messaging* related attributes.
- **tesysMessages.arff** contains the messages in text format and *messages* related attributes.

Identification attributes are presented in Table 4.1, in which the tuple (*learnerID*, *learningResourceID*) is primary key. Quiz related attributes and time related attributes are presented in Tables 4.2 and 4.3, respectively.

For example, if for a question the *maxAllowedTime* is 60 s and the *timeToAnswer* is 30 than *avgQuizQuestionTime* is 0.5 s.

Messaging related attributes and messages related attributes are presented in Tables 4.4 and 4.5, respectively.

In the data analysis process, all the features presented in the above Tables are regarded as parameters describing items (i.e., learners). For designing a feasible classification system, a target/class variable needs to be set from the existing features or a new one should be created. For preliminary studies the *avgMark* (i.e., the average mark of all taken tests.) feature has been discretized to *weak*, *average*, *good*, and *excellent* depending on its value.

Table 4.1 Identification attributes

Attribute name	Type	Meaning
learnerID	Integer	The learner, for which all other attributes are computed
learningResourceID	Integer	The learning resource (i.e., chapter, discipline, module), for which all other attributes are computed

Table 4.2 Quiz related attributes

Attribute name	Type	Meaning
positiveCount	Integer	The number of correctly answered questions
correctPercent	Real value from 0 to 100	The percentage of correctly answered questions from the total number of questions
totalTries	Integer	The total number of tries (answered questions)
avgTries	Integer	The average number of tries per question
avgMark	Real value from 0 to 10	The average mark of all taken tests

Table 4.3 Time related attributes

Attribute name	Type	Meaning
timeSpentOnLine	Integer	The overall time (s) that was spent on line
avgTimePerSession	Integer	The average time per session (s) that was spent online
nrOfSessions	Integer	The number of sessions (loggings)
avgQuizQuestionTime	Real value from 0 to 1	Average number of seconds per question in correspondence with maximum allowed number of seconds
totalQuizTime	Integer	Total time spent on testing (s)

Table 4.4 Messaging related attributes

Attribute name	Type	Meaning
totalNumberOfMessages	Integer	The total number of sent and received messages
totalNumberOfSentMessages	Integer	The total number of sent messages
totalNumberOfReceivedMessages	Integer	The total number of received messages
avgLengthOfSentMessages	Integer	The average length (in number of characters) of sent messages

Table 4.5 Messages related attributes

Attribute name	Type	Meaning
senderID	Integer	Identifies sender of the message
receiverID	Integer	Identifies receiver of the message
messageText	Array of chars	The raw text

Table 4.6 Concept coverage by student

Student	Concept coverage (%)				# of quest.	Avg. grade
	Node	Leaf	Traversal	Predecessor		
Alice	92	80	75	90	4	80
Bob	80	0	0	10	2	50
Carol	70	10	0	0	5	40

Table 4.7 Difficulty level of concepts

Concept	Node	Leaf	Traversal	Predecessor
Difficulty (%)	20	30	50	70

On the other hand, the messages as well as documents associated to chapters were parsed and a full set of concepts were obtained. From the design of the e-Learning environment, the professors associate each question from quizzes with a percentage of concept coverage for each quiz. The benefit of this approach allows straight forward computation of concept coverage level from the quiz activities. Therefore, once the system is running and the students are properly using it there may be obtained a general concept difficulty assignment and a student concept coverage level. Tables 4.6 and 4.7 present sample results related to the Binary Search Trees from the Algorithms and Data Structures undergraduate course.

The intuition behind the concept coverage from Table 4.6 is straight forward. Higher concept coverage represents a better/deeper knowledge level. On the other hand, a low percentage in difficulty for a concept means it has been well-understood by many students and, therefore, may be intuitively regarded as an easy one. For example, a 70 % in difficulty for *predecessor* concept is a clear indication that it is well understood only by students with a high knowledge level on the subject.

Therefore, once the system is in place and running a classifier is trained to estimate the difficulty level of a question taking into consideration the spread of concepts into that question, the average knowledge level of concepts that relate to the question, the number of concepts from the question and the number of concepts related to the question that were covered more than 50 %.

A sample input data is presented below, in *arff* file format.

```

@RELATION question Difficulty

@ATTRIBUTE concept Spread Lvl    NUMERIC
@ATTRIBUTE avg Knowledge Lvl    REAL
@ATTRIBUTE number Of Concepts   NUMERIC
@ATTRIBUTE number Easy Concepts NUMERIC
@ATTRIBUTE difficulty Lvl      {low, average, high}

@DATA
7, 8.2, 6, 3, high
3, 7.1, 5, 2, average
8, 5.4, 2, 2, low
...
    
```

The question difficulty level classifier has been obtained using inductive decision tree learner and has the following shape (Fig. 4.4).

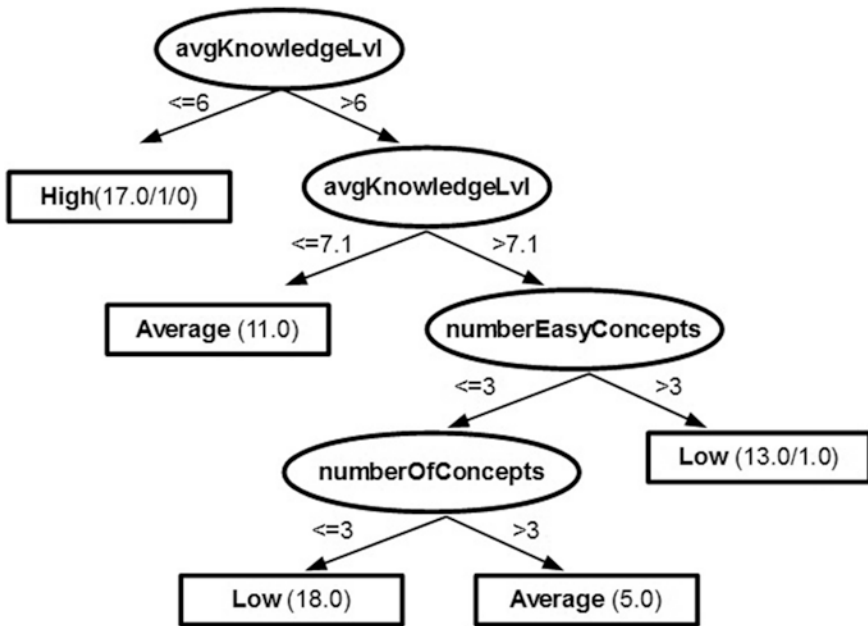


Fig. 4.4 Question difficulty level classifier

The integration of an on-line inductive learner for the estimation of a question difficulty level has direct implication into the overall classification of learners, which is the fundamental processing mechanism for messages indexing and retrieval.

The learner classifier has a similar shape and intuition as the above presented question difficulty classifier. The main differences reside in the set of attributes and in the significance of the class variable. For learner's classification, there were used all 15 attributes presented in Tables 4.1, 4.2, 4.3, 4.4 and the messages structured as in Table 4.5 were indexed into appropriate class as presented Fig. 4.3, where augmented classifier inference is presented. Once the augmented classifier is in place holding the messages, the proper retrieval is performed in a client-server manner.

4.5 Conclusions and Future Work

This chapter presents an integration of a supervised learning algorithm into an educational environment in order to develop a more efficient way of communication between professors and students.

In a traditional face-to-face education system, a professor is able to determine a student's interest in a particular subject by the continuous interaction, assessment, and behavior monitoring. In on-line educational systems, the previous approach is not possible because the professor cannot know the activity performed by students and, thus, determining the informational needs of learners becomes a critical issue. The presented peer-to-peer messaging system helps both students and professors by creating classes of students based on performed activities.

The messages exchanged between students and professors are indexed into a tree like structure (i.e., the classifier) by linking messages to student classes or labels. The goal of this chapter is to present a custom designed peer-to-peer messaging system that is implemented on top of an advanced message indexing system. This system makes intensive use of text mining and stemming capabilities that are put together into the concept detection and question difficulty estimator tools.

As future work, we can consider using additional features to the system in order to obtain more realistic and effective results. The homework mark or the time spent for testing could be among the feature to be explored.

The system should also implement a feedback relevance mechanism, in which students will mark interesting and non-interesting messages. This mechanism ensures that future requests of students will provide more relevant messages and will guide the design and development of the data analysis infrastructure towards a more effective one.

Another future improvement would be to implement this system for different types of resources like course documents or interesting articles. For example, a professor receives a question from a student and the answer in a form of a

document. The system should allow indexing of the document the same way as currently a message is indexed. Also as a future task the system may be improved by means of a module that can make an association between messages and concepts that define a chapter or a discipline.

In order to achieve better results on recommending the right messages to students, an automatic topic detection module must be integrated in the application. Using this module, the quality and relevance of the automatic answers given to students can be improved. Considering a big amount of data, this module can be used only on the classified students' messages obtained after the classification step.

When more data and/or more features for describing students are available, a challenger model may be created and the old one may be replaced, only if better classification accuracy is obtained. This should be performed in automatic way with the approval of a domain knowledge data analyst.

An extension for the currently existing infrastructure may implement automatic message sending. This approach will diminish the initial interaction for the learner and may be a proper solution for the learners, who are at the beginning of their educational endorsement, and ask themselves, what they would do next. In order to accomplish this, we may need a usability study with Human Computer Interaction (HCI) techniques, which should emphasize the effectiveness of the approach. Also from HCI perspective, employment of different visualization techniques may also improve the effectiveness of the proposed infrastructure.

From a data analysis point of view, a more fine grained classification at chapter/concept level may bring more light into the insight of the data with a higher level of understanding of learner's behavior in our attempt to model it. Mathematical modeling/measuring of the impact/effects of this approach may require further data analysis studies.

References

1. Burdescu, D.D., Mihăescu, M.C.: Tesys: e-Learning application built on a web platform. In: International Joint Conference on e-Business and Telecommunications, pp. 315–318. Setubal, Portugal (2006)
2. Hua, J., Xiong, Z., Lowey, J., Suh, E., Dougherty, E.R.: Optimal Number of Features as a Function of Sample Size for Various Classification Rules. Oxford University Press, Oxford (2004)
3. Brownlee, J.: Reproducible Machine Learning Results By Default. <http://machinelearningmastery.com/reproducible-machine-learning-results-by-default/>. Accessed 8 Feb 2015
4. Quinlan, J.R.: Induction of decision trees. *Mach. Learn.* **1**(1), 81–106 (1986)
5. Beck, J.: Proceedings of AAAI2005 Workshop on Educational Data Mining (2005)
6. Baker, R.S.J.D., Barnes, T., Beck, J.E. (eds.): Educational Data Mining 2008: 1st International Conference on Educational Data Mining, Proceedings. Montreal, Quebec, Canada (2008)
7. Romero, C., Ventura, S.: Data Mining in e-Learning. WIT Press, Ashurst (2006)
8. Romero, C., Ventura, S.: Handbook of Educational Data Mining. CRC Press, Boca Raton (2010)
9. Minaei-Bidgoli, B., Kashy, D.A., Kortemeyer, G., Punch, W.F.: Predicting student performance: an application of data mining methods with the educational web-based system LON-CAPA. In: 33th ASEE/IEEE Frontiers in Education Conference, Boulder, CO. IEEE pp. 1–6 (2003)

10. Hamalainen, W., Vinni M.: Classifiers for educational data mining. In: Handbook of Educational Data Mining (2010)
11. Abu Tair, M.M., El-Halees, A.M.: Mining educational data to improve students' performance: a case study. *Int. J. Inf. Commun. Technol. Res.* **2**(2), 140–146 (2012)
12. Blackboard Inc.: <https://www.blackboard.com>. Accessed 8 Feb 2015
13. Blackboard Inc. Press Release.: Blackboard opens field trials for learning analytics solution. <https://www.blackboard.com/About-Bb/News-Center/Press-Releases.aspx?releaseid=1646731>. Accessed 8 Feb 2015
14. Cilogluligil, B., Inceoglu, M.M.: User modeling for adaptive e-Learning systems, computational science and its applications. In: Murgante, B., Gervasi, O., Misra, S., Nedjah, N., Rocha, A.M.A.C., Taniar, D., Apduhan, B.O. (eds.) *Computational Science and Its Applications—ICCSA 2012*. LNCS, 7335, pp. 550–561. Springer, Berlin (2012)
15. Sharma, D.: Stemming algorithms: a comparative study and their analysis. *Int. J. Appl. Inf. Syst.* **4**(3), 7–12 (2012)
16. Gabriela, T.M., Cristian, M.M., Burdescu, D.D.: Building professor's mental model of student's activity in on-line educational systems. In: 3rd International Conference on Cognitronics—The Science about the Human Being in the Digital World, 16th International Multi-conference Information Society (IS'2013), pp. 472–476 (2013)
17. PTStemmer—A Stemming toolkit for the Portuguese language. <https://code.google.com/p/ptstemmer/>. Accessed 8 Feb 2015
18. Sharma, D.: Stemming algorithms: a comparative study and their analysis. *Int. J. Appl. Inf. Syst.* **4**(3), 7–12 (2012)
19. Mocanu, M., Popescu, P.-S., Burdescu, D.D., Mihaescu, M.C.: Advanced messaging system for on-line educational environments. In: Tsihrintzis, G.A., Virvou, M., Watanabe, T., Jain, L.C., Howlett, R.J. (eds.) *Intelligent Interactive Multimedia Systems and Services*, vol. 254, pp. 61–69. IOS Press, Amsterdam (2013)
20. Witten, I.H., Frank, E., Trigg, L.E., Hall, M.A., Holmes, G., Cunningham, S.J.: *Weka: Practical Machine Learning Tools and Techniques with Java Implementations*. University of Waikato, Hamilton, Department of Computer Science (1999)
21. Lowd, D., Domingos, P.: Naive Bayes models for probability estimation. In: 22nd International Conference on Machine learning, pp. 529–536. ACM Press, New York (2005)
22. Gunn, S.R.: Support vector machines for classification and regression (1998)
23. Bradley, P.S., Fayyad, U.M., Cory, A.: Reina scaling EM (expectation-maximization) clustering to large databases. Microsoft Research (1999)
24. Looney, C.G.: A fuzzy clustering and fuzzy merging algorithm. Technical Report, CS-UNR-101 (1999)
25. Loh, W.Y.: *Classification and Regression Tree Methods*. Wiley, Hoboken (2008)
26. Quinlan, J.R.: *C4.5: Programs for Machine Learning*. San Mateo, CA (1993)
27. Bahety, A.: Extension and Evaluation of ID3 Decision Tree Algorithm. University of Maryland, College Park. https://www.cs.umd.edu/sites/default/files/scholarly_papers/Bahety_1.pdf
28. Freund, Y., Mason, L.: The alternating decision tree learning algorithm. In: 6th International Conference on Machine Learning, pp. 124–133. Morgan Kaufmann Publishers Inc. San Francisco (1999)
29. Holmes, G., Pfahringer, B., Kirkby, R., Frank, E., Hall, M.: Multiclass alternating decision trees. In: 13th European Conference on Machine Learning, pp. 161–172. Springer, London (2002)
30. Shi, H.: Best-first decision tree learning. Thesis on Master Science, University of Waikato, New Zealand (2007)
31. Choy, M., Flom, P.: Building Decision Trees from Decision Stumps. SAS Global Forum (2010)
32. Fonseca, M.J., Jorge, J.A.: NB-Tree: An Indexing Structure for Content-Based Retrieval in Large Databases. <http://www.di.fc.ul.pt/~mjf/publications/2004-1999/pdf/mjf-jaj-TR-01-03.pdf> (2003)

Chapter 5

Fuzzy-Based Digital Video Stabilization in Static Scenes

Margarita Favorskaya and Vladimir Buryachenko

Abstract In recent years, a digital video stabilization improving the results of hand-held shooting or shooting from mobile platforms is the most popular approach. In this chapter, the task of digital video stabilization in static scenes is investigated. The unwanted motion caused by camera jitters or vibrations ought to be separated from the objects motion in a scene. Our contribution connects with the development of deblurring method to find and improve the blurred frames, which have strong negative influence on the following processing results. The use of fuzzy Takagi-Sugeno-Kang model for detection the best local and global motion vectors is the novelty of our approach. The quality of test videos stabilization was estimated by Peak Signal to Noise Ratio (PSNR) and Interframe Transformation Fidelity (ITF) metrics. Experimental data confirmed that the ITF average estimations increase up on 3–4 dB or 15–20 % relative to the original video sequences.

Keywords Digital video stabilization · Deblurring · Motion estimation · Fuzzy logic · Scene alignment

5.1 Introduction

The unintentional video camera motion during hand-held shooting or shooting by cameras, which are maintained on the unstable platforms, decreases a quality of video sequence that has a negative effect on the following frames processing.

M. Favorskaya (✉) · V. Buryachenko
Department of Informatics and Computer Techniques, Siberian State Aerospace University,
31 Krasnoyarsky Rabochy Avenue, Krasnoyarsk 660014, Russian Federation
e-mail: favorskaya@sibsau.ru

V. Buryachenko
e-mail: buryachenko@sibsau.ru

The use of stabilized techniques is widely spread in video surveillance [1] and video encoding [2]. Methods of videos stabilization are classified as mechanical, optical, electronic, and digital. The mechanical stabilization systems included gyroscopes, accelerometers, etc., and were used on the early stage of video cameras development [3]. The optical stabilization systems use prisms or lens of moving assembly for tuning of light length way through camera lens systems. Such technical realization is not suitable for small sizes mobile cameras. The electronic stabilization systems detect the camera jitters through their sensors, when the light hits in Charge-Coupled Device (CCD). It has the advantage against the optical stabilization by reducing of lens complexity and price.

The Digital Video Stabilization (DVS) approach is achieved by the synthesis of a new imagery based on removal of unintentional motions between key frames and the reconstruction of frame boundaries after frame stabilization. The DVS based on the algorithmic improvement became the most appropriate decision in modern compact video devices. However, the DVS algorithms ought to be robust to different scene contents, moving objects, and luminance changing. The complexity of this task connects with separation of the objects motion from camera jitters in static scenes.

Our contribution connects with the improvement of some blurred frames in original video sequence, the development of DVS method based on the Takagi-Sugeno-Kang (TSK) model for improvement of motion vectors clustering, and the application of scene alignment procedure into static scenes.

The chapter has been structured as follows. In Sect. 5.2, a brief description of the existing approaches for digital video stabilization previously in static scenes is provided. Method for automatic blurred image detection and compensation is discussed in Sect. 5.3. Section 5.4 describes the proposed method of videos stabilization based on Block-Matching Algorithm (BMA) using the fuzzy logic approach in the selected image area. Experimental numerical results of the proposed approach are presented in Sect. 5.5. Conclusions and future research are drawn in Sect. 5.6.

5.2 Related Work

Video stabilization includes three main stages: motion estimation, motion smoothness, and frame correction. Approaches of motion estimation are directed on reduction of computational cost using fast BMA [4], limited pre-determined regions [5], or feature tracking [6]. Tanakian et al. [4] proposed the integrated system of video stabilizer and video encoder based on the BMA for the Local Motion Vectors (LMVs) detection, histogram analysis for Global Motion Vector (GMV) building, and Smooth Motion Vector (SMV) calculation for intentional motion correction. Tanakian et al. suggested a low pass filtering to remove a high frequency component of the intentional motion. Equation 5.1 approximates the SMVs using a first-order auto regression function, where α is a smoothing factor, $0 \leq \alpha \leq 1$; n is a frame number.

$$|SMV_n| = \alpha |SMV_{n-1}| + (1 - \alpha) |GMV_n| \quad (5.1)$$

Also Tanakian et al. proposed the rule to chose α value ($\alpha = 0.1$ or $\alpha = 0.95$) in dependence of the SMVs and the GMVs magnitudes in previous frames. After following researches, a fuzzy system for tuning of smoothing factor α was suggested according to noise and the possible camera motion acceleration [5]. Triangular and trapezoidal membership functions were used for adaptive filtering of horizontal and vertical motion components between $(n - 3)$, $(n - 2)$, $(n - 1)$, and n frames.

Another approach was applied by Acharjee and Chaudhuri in [7] as a Three Step Search BMA, which provides the fuzzy membership values. The fuzzy membership value is calculated for each macro block according to its intensity characteristics (lighter or darker) previously in edges of an image. Such restrictions reduce the computational cost of full Three Step Search BMA, while the PSNR values are almost saved. In research [8], a fuzzy logic model to evaluate a quality of matching between a pair of Scale-Invariant Feature Transform (SIFT) descriptors was proposed. A fuzzy logic model used two error measures (Euclidean distance between expected and real points and angle between two local motion vectors) as inputs and the single quality index in the range $[0, 1]$ as output. A final decision of a quality of points' matching was estimated using a Sugeno model [9].

A fuzzy Kalman compensation of the GMV in the log-polar plane was proposed in research [10]. Due to special features of the log-polar plane, the GMV was calculated as the average value of the four LMVs. Then the GMV displacements were imported into the fuzzy Kalman system. The fuzzy system was tested with several types of Membership Functions (MFs) and different aggregation and defuzzification methods. Some original approaches may be found in the researches dedicating to video stabilization by use a principal component analysis [11], an independent component analysis [12], a probabilistic global motion estimation based on Laplacian two-bit plane matching [13], wavelet transformations [14], the calculation of statistical functions, mean and variance of pixels in each block of the BMA [15], etc.

The blurred frames have negative influence on stabilization of video sequence. The reasonable approach connects with removal such blurred frames before the application of stabilization algorithms [16]. The different deblurring methods are mentioned below:

- Deblurring methods for a single image reconstruction are applied usually the uniform deblurring core [17, 18]. Such methods cannot be applied directly for frames from video sequence in spatio-temporal domain. In recent years, 3D deblurring core was proposed to describe a spatio-changing core for a frame [19]. Gupta et al. [20] proposed to present the camera motion using motion density functions. Such methods are not reliable for real noisy and/or compressed video sequences with multiple moving objects in a scene.
- Deblurring methods for several images reconstruction are based on different approaches such as the numerical method for calculation of a blurring model [21], the use of the point spread function to restore the original images with

minimal energy [22], or the interpolation method in order to increase the sharpness all frames including the deblurred frames [23].

- Methods based on selection of “suitable” images guess the overlapping of several good input images in order to receive the best single image. However, the superposition of real frames is not a trivial task because of blurring and luminance differences of selected frames.

Consider the proposed deblurring method as a crucial issue of pre-processing in the DVS task.

5.3 Method of Frame Deblurring

The proposed method applies the anisotropic Gauss filter with adaptive automatic selection of region sizes and includes the following steps:

1. The automatic estimation of blurred frames based on gradient information.
2. The detection of textured and smoothness regions in a frame.
3. The analysis of edge information by use a Sobel filter.
4. The application of anisotropic Gauss filter with automatic mask selection in textured regions.
5. The application of unsharp mask in smoothness region.
6. The synthesis of result frame.

Let us notice that a blurring happens not always, but during high jitters, fast motion objects, or continuous camera exposition. Therefore, not all frames are blurred, and it is required to detect, which frames are blurred into the analyzed part of video sequence.

Introduce a measure of sharp estimation based on gradient information into full frame and pre-determined blurring threshold value T . A blurring degree in current frame is estimated by Eq. 5.2, where T is a blurring threshold value, $0 \leq T \leq 1$, g_n is a gradient function of frame, I_{ij} is value of intensity function of frame in a pixel with coordinates (i, j) , N and M are the frame sizes, K is a number of previous key frame.

$$\sum_{i=1}^M \sum_{j=1}^N g_n^2(I_{ij}) < T$$

$$\times \max \left\{ \sum_{i=1}^M \sum_{j=1}^N g_{n-K}^2(I_{ij}), \sum_{i=1}^M \sum_{j=1}^N g_{n-(K+1)}^2(I_{ij}), \dots, \sum_{i=1}^M \sum_{j=1}^N g_{n-1}^2(I_{ij}) \right\}$$
(5.2)

Examples of detected frames with high and low blurring degree are situated in Fig. 5.1.



Fig. 5.1 Video sequence “Sam_1.avi”: **a** frame 59 with high sharp, **b** frame 68 with low sharp, **c** the increased fragment of frame 59, **d** the increased fragment of frame 68

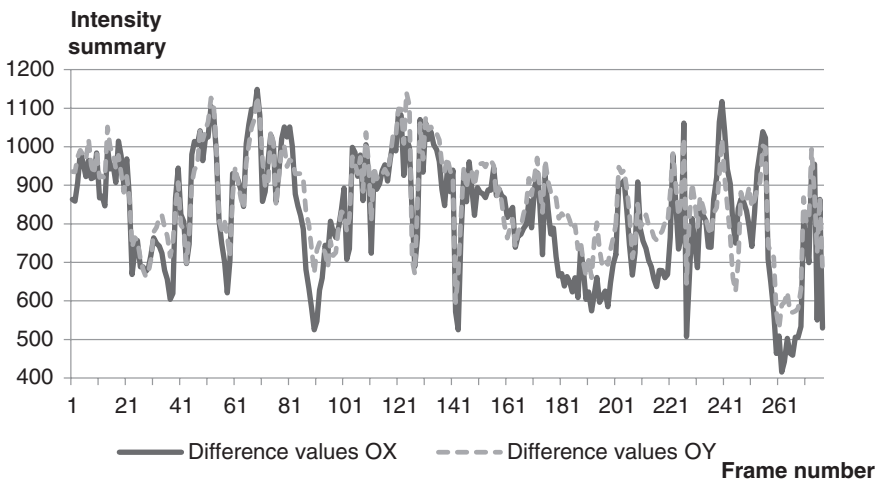


Fig. 5.2 A blurring estimation of video sequence “Sam_1.avi”

For video sequence “Sam_1.avi”, the plots of blurring degree for frames 1–265 are situated in Fig. 5.2. The maximum values mean the availability of blurred frames. This scene contains fast motion and significant jutting.

For detection of textured and smoothness regions in a frame, the frame differences are estimated in a slicing window 5×5 pixels by Eq. 5.3, where $\beta_L(x, y)$ is a blurring degree into pixel with coordinates (x, y) .

$$\beta_L(x, y) = \frac{1}{25} \sum_{i=1}^5 \sum_{j=1}^5 \left(I_{i,j}^2 - I_{(i+1),(j+1)}^2 \right) \quad (5.3)$$

According to the received values $\beta_L(x, y)$, the binary map $B_m(x, y)$ is calculated in order to estimate the textured and smoothness regions into a frame by Eq. 5.4, where T_{fl} is an automatic chosen threshold value in dependence of total g_n value provided by Eq. 5.2.

$$B_m(x, y) = \begin{cases} 0 & \text{if } \beta_L(x, y) < T_{fl} \\ 1 & \text{in other cases} \end{cases} \quad (5.4)$$

The analysis of edge information can be realized by Sobel filter or other filters for edge estimations. If a density of edges is high into a unit region, then this region is considered as a textured region. In such case, the anisotropic Gauss filter is applied. The core of this filter is adaptively tuned for pixel closed to an edge, pixel near the edge, and pixel far from an edge. This procedure is required in order to exclude the sharp edge pixels from such processing. The parameters of anisotropic Gauss filter are calculated with different scaling along axes OX and OY by Eq. 5.5, where σ_x and σ_y are standard deviations chosen in dependence of remote pixel (x, y) from an edge.

$$g(x, y; \sigma_x, \sigma_y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp \left\{ -\frac{1}{2} \left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} \right) \right\} \quad (5.5)$$

For smoothness regions, the unsharp mask, which is based on the subtraction the blurred frame from the original frame, is used. The sizes of unsharp mask are changed dynamically in dependence on analyzed region sizes in a frame. A synthesis of the deblurred frame is realized by considering the edge, improved textured, and improved smoothness information.

A method of frame deblurring was developed under the assumption that a number of blurred frames into the analyzed part of video sequence is not large, one or two from 25–30 frames. Therefore, a computational cost of this procedure is not high. However, the influence on the final stabilization result is positive.

5.4 Fuzzy-Based Video Stabilization Method

The proposed video stabilization method involves three main stages: the LMVs estimation, the GMVs smoothness, and the frames correction. The LMVs detection with following improvement by the TSK model is discussed in Sect. 5.4.1. Section 5.4.2 provides the smoothness GMVs building. In Sect. 5.4.3, the static scene alignment is considered.

5.4.1 Estimation of Local Motion Vectors

For the LMVs estimations, many approaches with various computational costs can be applied. The experiments show that the BMA provides fast motion estimations with appropriate accuracy in static scene. First, the current frame is divided in the non-crossed blocks with similar sizes (usually 16×16 pixels), which are defined by the intensity function $I_t(x, y)$, where (x, y) are coordinates, t is a discrete time moment. Second, for each block in small neighborhoods $-S_x < d_x < +S_x$ and $-S_y < d_y < +S_y$, the most similar block into the following frame $I_{t+1}(x + d_x, y + d_y)$ is searched. The similarity is determined by a minimization of the error functional e according to the used metric. Usually three metrics are applied such as a Sum of Absolute Differences (SAD), a Sum of Squared Differences (SSD), and a Mean of Squared Differences (MSD) (Eq. 5.6), where d_x and d_y are the block displacement in directions OX and OY, respectively, n is a number of analyzed surrounding blocks.

$$\begin{aligned}
 e_{SAD}(d_x, d_y) &= \sum_{x=1}^N \sum_{y=1}^N |I_{t+1}(x, y) - I_t(x + d_x, y + d_y)| \\
 e_{SSD}(d_x, d_y) &= \sum_{x=1}^N \sum_{y=1}^N (I_{t+1}(x, y) - I_t(x + d_x, y + d_y))^2 \\
 e_{MSD}(d_x, d_y) &= \frac{1}{n \times n} \sum_{x=1}^N \sum_{y=1}^N (I_{t+1}(x, y) - I_t(x + d_x, y + d_y))^2
 \end{aligned} \tag{5.6}$$

Vector $\mathbf{V}(d_x, d_y)$, for which the error functional e ($e_{SAD}(d_x, d_y)$, $e_{SSD}(d_x, d_y)$, or $e_{MSD}(d_x, d_y)$) has the minimum value, is considered as the displacement vector for the selected block. The basic BMA referred as Full Search (FS) has a disadvantage of high computer cost. Some modifications of the FS exist such as Three-Step Search (TSS), Four-Step Search (FSS), Conjugate Direction Search (CDS), Dynamic Window Search (DSW), Cross-Search Algorithm (CSA), Two-Dimensional Logarithmic Search (TDLS), etc.

After the LMVs building, it is needed to determine, which LMVs describe an unwanted camera motion and which LMVs concern to objects' motion in a scene. The proposed model is based on triangular, trapezoidal, and S -shape memberships in terms of fuzzy logic to partitioning the LMVs. The views of memberships are presented in Fig. 5.3, where parameters a and b of S -shape membership are fitted empirically. The recommendations based on the experimental results are the following: to use $a = 0.5$ and $b = 1.5$ for non-noisy videos and $a = 0.75$ and $b = 1.75$ for noisy videos. The inputs of fuzzy logic model have two error measures:

- The Euclidean distance between the expected and the real point estimated in one of the SAD, the SSD, or the MSD metrics (magnitude of vectors) $\mathbf{E}' = (e_1, e_2, \dots, e_i, \dots, e_n)$.

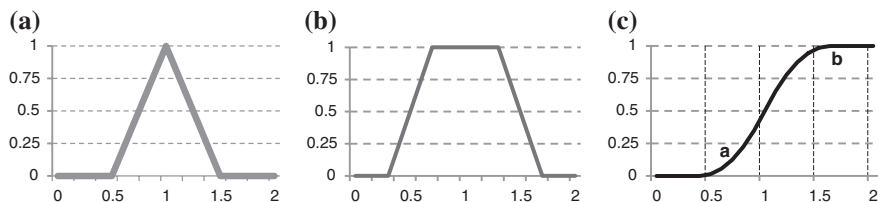


Fig. 5.3 View of memberships in fuzzy logic model: **a** triangular, **b** trapezoidal, **c** S-shape

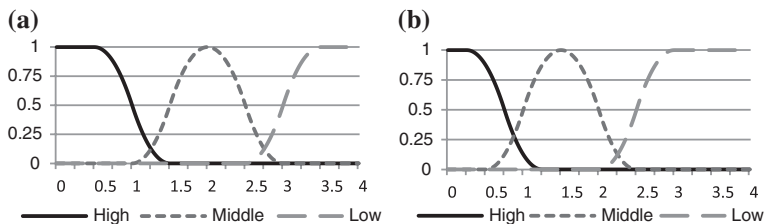


Fig. 5.4 View of S-shape memberships: **a** for non-noisy video sequence, **b** for noisy video sequence

- The angle between two local motion vectors $C' = (c_1, c_2, \dots, c_i, \dots, c_n)$, where $i = 1 \dots n$.

Equation 5.7 provides the error deviations d_i^e and d_i^c similar to research [6], where M_E and M_C are the median values of E' and C' sets, respectively.

$$d_i^e = e_i / M_E \quad d_i^c = c_i / M_C \quad (5.7)$$

Values of error deviations d_i^e and d_i^c are mapped into three different classes of accuracy: high, medium, and low. The lower values of error deviations are mapped to the best classes, and otherwise. If membership functions are overlapped, then more good definition from the input fuzzy sets is chosen.

The output of fuzzy logic model indicates a final reliability of matching quality using the TSK model [9]. Such zero-order fuzzy model infers the quality index (a value in the range [0, 1]). The quality of the points' matching is classified into four categories: excellent, good, medium, and bad. Each of these four classes is mapped into a set of constant values 1.0, 0.75, 0.5, 0.0, respectively. Views of triangular, trapezoidal, and S-shape memberships are situated in Fig. 5.3.

The recommended S-shape functions are situated in Fig. 5.4.

The output of fuzzy logic model indicates a final reliability of estimations for a quality of the matching using the TSK model. The quality index is a value in the range [0, 1]. It shows a quality of LMVs, which are clustered into four classes: excellent, good, medium, and bad. The "IF-THEN" fuzzy rules defined for two inputs (error deviations d_i^e and d_i^c) are the following:

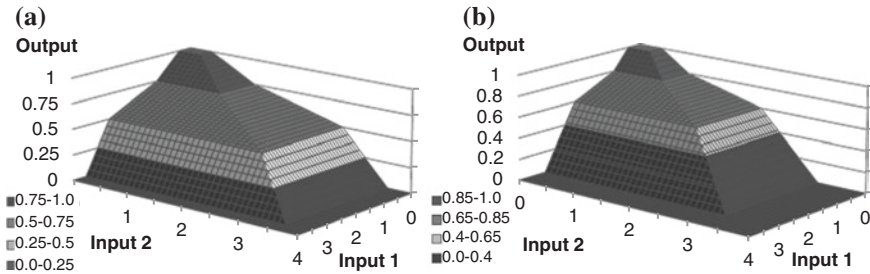


Fig. 5.5 View of TSK models: **a** for non-noisy video sequence, **b** for noisy video sequence

- IF (both inputs = “high”) THEN (quality = “excellent”).
- IF (one input = “high” AND other input = “medium”) THEN (quality = “good”).
- IF (both inputs = “medium”) THEN (quality = “medium”).
- IF (at least one input = “low”) THEN (quality = “bad”).

Each of these four classes is mapped into a set of the constant values (1.0, 0.75, 0.5, 0.0) [9]. During our experiments, the results for noisy video sequences were received with a set of the constant values (1.0, 0.85, 0.65, 0.0). The TSK models for non-noisy and noisy video sequences with the sets of constant values (1.0, 0.75, 0.5, 0.0) and (1.0, 0.85, 0.65, 0.0) are show in Fig. 5.5 [24]. The TSK model permits to discriminate the LMVs with excellent and good quality and detect the best LMVs (with excellent and good values of indexes) in order to improve a final result.

Our following researches permitted to speed the LMVs calculation for both types of video sequences in the static scenes. Introduce an initial procedure, which will put an invisible grid on each frame adaptively to the frame sizes with 30–50 cells.

The sizes of such grid are less in order to reject the boundary areas of frame, which are more stressed to artifacts of instability. For five first frames in a scene, the LMVs estimations and their improvements by the TSK model are calculated for all cells of a grid. For each cell, the information of reliable LMVs is accumulated under the condition, that 4–16 reliable LMVs are determined into a cell. According to a scene background, such several cells can be selected for following analysis. Therefore, the LMVs of unwanted motion are calculated only in the selected cells that permits to avoid the challenges of luminance changing or moving foreground objects and reduce the number of analyzing cells in 1.5–3 times. Figure 5.6 provides such adaptive and fast technique for frame number 140 from video sequence “If_juggle.avi”.

The TSK model discriminates the excellent and good results well. The selection of the best points with excellent and good values of indexes improves the final results.

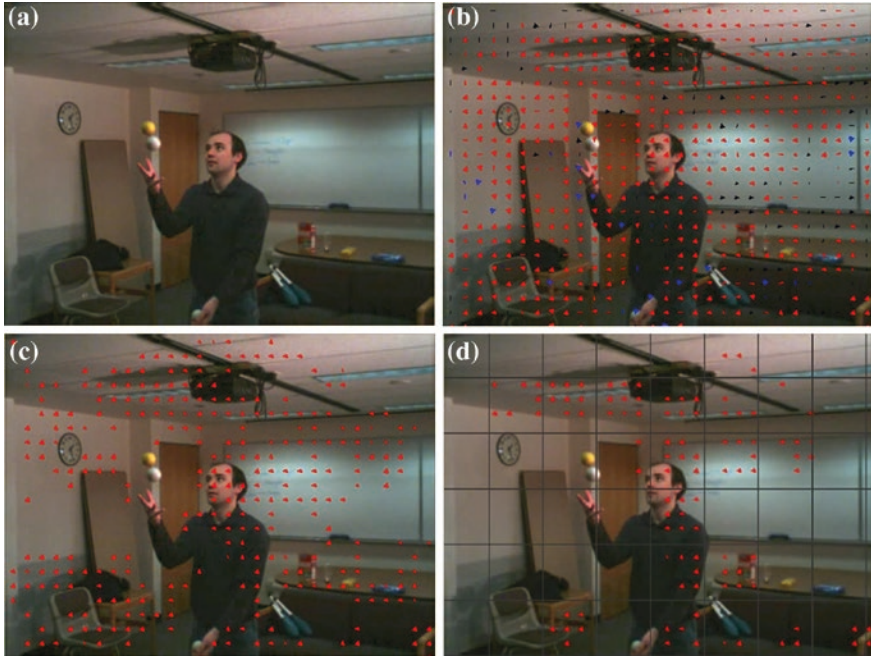


Fig. 5.6 The adaptive technique for LMVs estimation in a static scene, video sequence 'lf_juggle.avi': **a** the original frame 140; **b** all calculated LMVs; **c** the reliable LMVs in the whole frame based on TSK model; **d** the reliable LMVs in the selected cells of imposed grid

5.4.2 Smoothness of GMVs Building

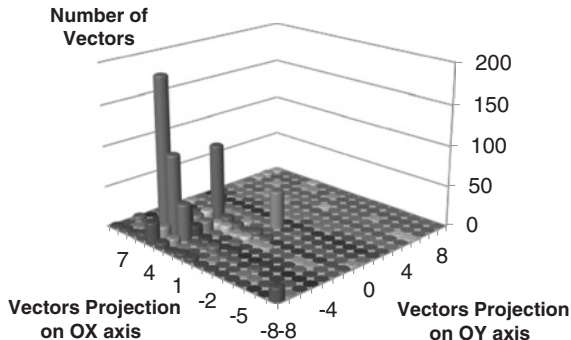
The global motion caused by camera movement is estimated for each frame by use a clustering model. The LMVs of background are very similar on magnitudes and directions but essentially different from objects' motion in foreground. The following procedure classifies the motion field into two clusters: the background and the foreground motions:

- Step 1. The histogram H is built, which includes only valid LMVs.
- Step 2. The LMVs are clustered by a similar magnitudes criterion.
- Step 3. The LMV with a maximum magnitude from background motion cluster is chosen as GMV.

The example of a histogram with valid LMVs is presented in Fig. 5.7.

Any GMV includes two major components: the real motion (for example, a panning) and the unwanted motion. Usually the unwanted motion corresponds to high frequency component. Therefore, a low-frequency filtering can remove the unwanted motion. The model proposed in research [4] forms the SMV calculating by Eq. 5.1. The low-pass filter of the first order needs in low computational

Fig. 5.7 Example of a histogram with valid LMVs



resources and may be used in a real-time application. To improve these results, a similar fuzzy logic model from Sect. 5.4.1 was used for clustering the intentional and the non-intentional GMVs. The adapting tuning procedure of a smoothing factor α based on analysis of previous 25 frames was proposed. First, Eq. 5.8 calculates a Global Difference $GDiff_k$, where $|GMV_i|$ is the magnitude of global motion vector in a frame i , $k > 25$.

$$GDiff_k = \sum_{i=k-25}^k ||GMV_i| - |GMV_{i-1}|| \quad (5.8)$$

Second, α value is chosen by Eq. 5.9, where $\alpha_{max} = 0.95$ and $\alpha_{min} = 0.5$ are maximum and minimum empirical values.

$$\alpha_k = \begin{cases} \frac{GDiff_k}{|GMV_{max}|} \times \frac{\alpha_{max}}{\alpha_{min}} & \text{if } GDiff_k < |GMV_{max}| \\ \alpha_{max} & \text{if } GDiff_k \geq |GMV_{max}| \end{cases} \quad (5.9)$$

In any case, the result from Eq. 5.9 is rounded to α_{max} .

5.4.3 Static Scene Alignment

For each frame after the smooth factor α calculation, the module of smooth motion vector SMV_n using Eq. 5.1 is determined. The module of an Undesirable Motion Vector (UMV) UMV_n is calculated by Eq. 5.10.

$$|UMV_n| = |GMV_n| - |SMV_n| \quad (5.10)$$

In the development of a scene alignment, the direction of SMV_n is normalized up to 8 directions with interval of 45° . For restoration of current frame, pixels are shifted on a value of an Accumulated Motion Vector (AMV) AMV_n of unwanted motion by Eq. 5.11, where m is the number of a current key frame in video sequence.

$$AMV_n = \sum_{i=m}^n |UMV_i| \quad (5.11)$$

The stabilized location of frame is determined from previous frames beginning from the current key frame.

5.5 Experimental Results

Six video sequences received by the static camera shooting were used during experiments. The titles, URL, and snapshots of these investigated video sequences are presented in Table 5.1. All experiments were executed by the own designed software tool “DVS Analyzer”, v. 2.04. The software tool “DVS Analyzer” has two modes: the pseudo real-time stabilization of video sequences, which are broadcasted from the surveillance cameras (the simplified processing), and the unreal-time stabilization of available video sequences (the intelligent processing).

The architecture of the software tool includes the extended set of program modules, which can be developed independently each from others. The Pre-processing Module, the Motion Estimation Module, the Motion Compensation Module, the Motion Inpainting Module, the Module of Quality Estimation, the Core Module, and the Interface Module are the main components of the software tool “DVS Analyzer”. The software tool “DVS Analyzer”, v. 2.04 was designed in the Rapid Application Development Embracadero RAD Studio 2010. Some external software tools were used such as the libraries “Video For Windows” for initial processing and “AlphaControls 2010 v7.3” for enhanced user interface, and a video codec “K-Lite Codec Pack”, v. 8.0.

The experimental graphics for motion estimation and compensation of video sequences situated in Table 5.1 are represented in Fig. 5.8. The experimental plots for stabilization quality of these video sequences are represented in Fig. 5.9.

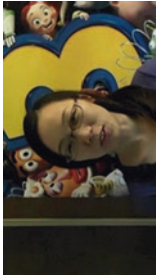
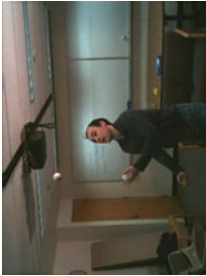


As it is shown from Fig. 5.9, the PSNR estimations of the stabilized video sequences are always higher than the PSNR estimations of the original video sequences.

The objective estimation of video stabilization quality was calculated by the PSNR metric between current frame I^{cur} and key frame I^{key} expressed in Eqs. 5.12–5.13, where MSE is a mean-square interframe error, I_{max} is a maximum of pixel intensity, m and n are sizes of frame.

$$MSE = \frac{1}{m \times n} \sum_{y=1}^m \sum_{x=1}^n \left(I^{cur}(x, y) - I^{key}(x, y) \right)^2 \quad (5.12)$$



$$PSNR = 10 \log_{10} \left(\frac{I_{max}^2}{MSE} \right) \quad (5.13)$$

Table 5.1 Description of investigated static scenes

Title, URL	Snapshot	Resolution	Frames	Motion type
“SANY0025_xvid.avi” http://cpl.cc.gatech.edu/projects/videostabilization/		640 × 360	445	Slow motion of camera, large object motion
“if_juggle.avi” http://cpl.cc.gatech.edu/projects/videostabilization/		480 × 360	460	Static camera, fast motion of small objects
“akiyo.avi” http://see.xidian.edu.cn/vips/database_Video.html		352 × 288	300	Static camera, large object motion
“EllenPage_Juggling.avi” http://www.youtube.com/watch?v=8YNUSCX_akk		1280 × 720	430	Static camera, fast motion of small objects

(continued)

Table 5.1 (continued)

Title, URL	Snapshot	Resolution	Frames	Motion type
"Butovo_synthetic.avi" http://youtu.be/0oeIZ04sXu0		640 × 480	748	Static camera, strong vertical and horizontal camera jittings
"road_cars_krasnoyarsk.avi" http://www.youtube.com/watch?v=pJ84Pwpbl_Y		852 × 480	430	Static camera, several moving objects, high deep of scene

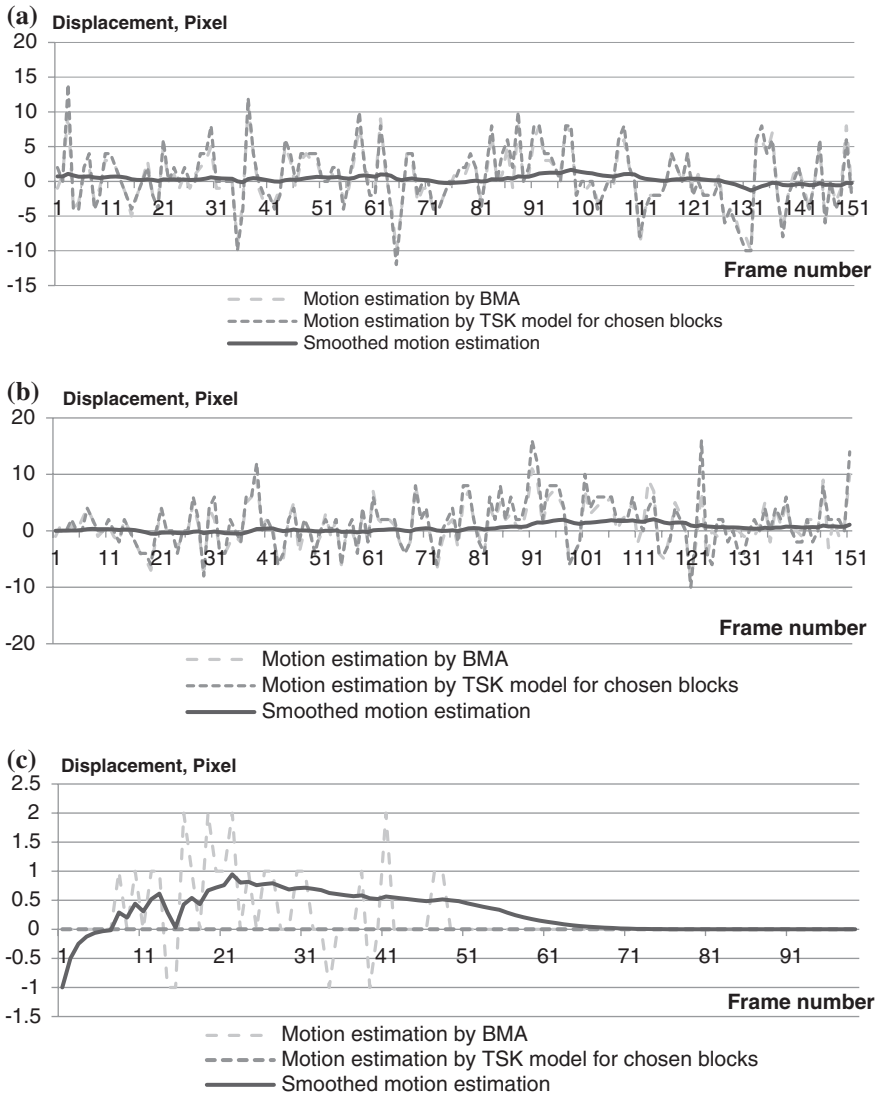


Fig. 5.8 Plots of motion estimation and compensation results in static scenes: **a** “SANY0025_xvid.avi”, **b** “If_juggle.avi”, **c** “akiyo.avi”, **d** “EllenPage_Juggling.avi”, **e** “Butovo_synthetic.avi”, **f** “road_cars_krasnoyarsk.avi”

The PSNR metric is useful for estimations between adjacent frames. A quality of the ITF metric provides the objective estimation in whole video sequence. The ITF of stabilized video sequence is higher than the ITF of original video sequence. This parameter is calculated by Eq. 5.14, where N_{fr} is a frame number in video sequence.

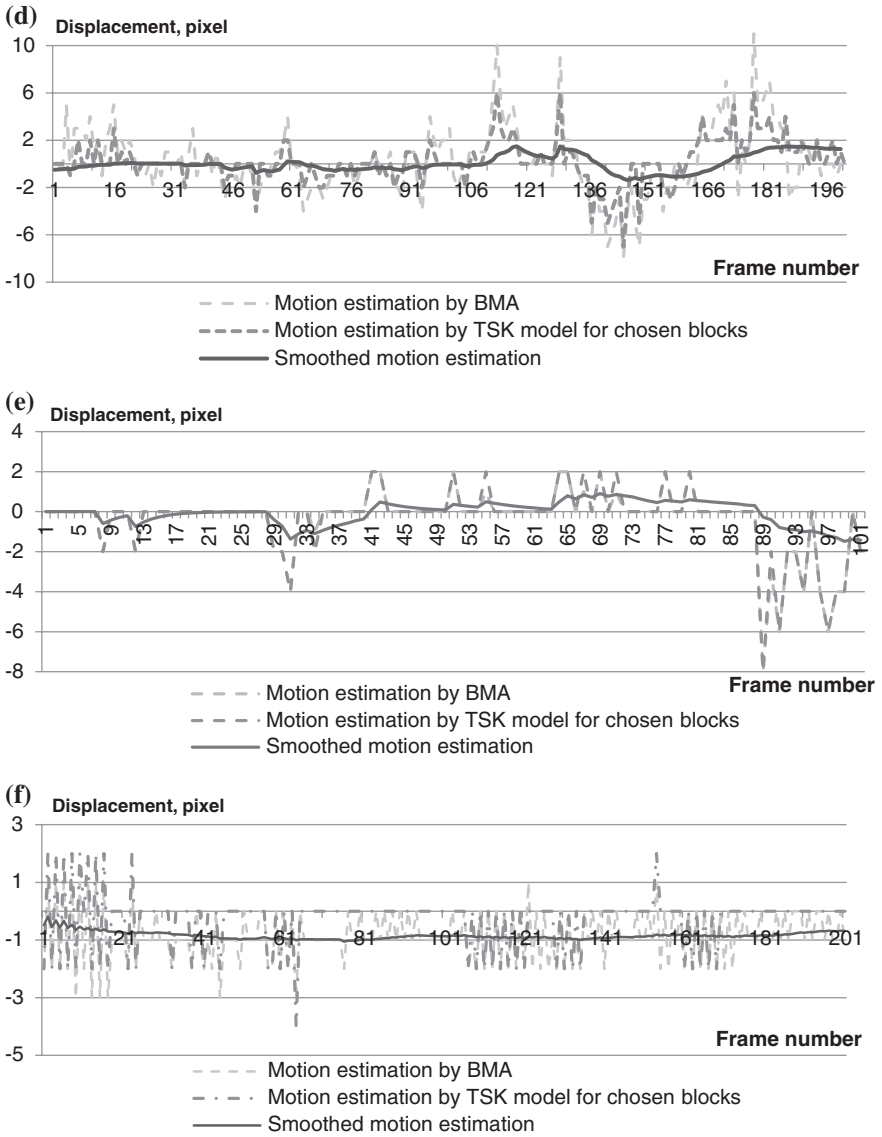


Fig. 5.8 (continued)

$$ITF = \frac{1}{N_{fr}} \sum_{k=0}^{N_{fr}} PSNR_k \tag{5.14}$$

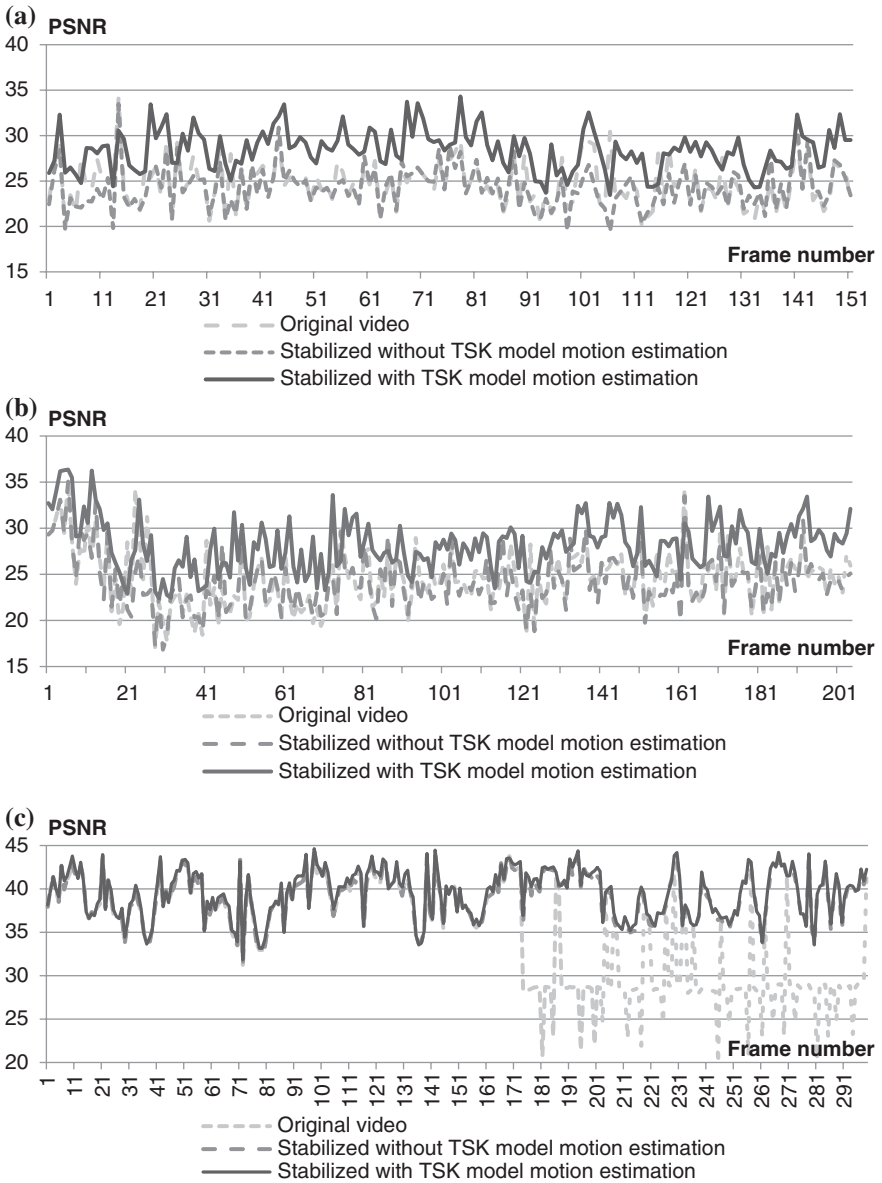


Fig. 5.9 Plots of stabilization quality in static scenes: **a** “SANY0025_xvid.avi”, **b** “lf_juggle.avi”, **c** “akiyo.avi”, **d** “EllenPage_Juggling.avi”, **e** “Butovo_synthetic.avi”, **f** “road_cars_krasnoyarsk.avi”

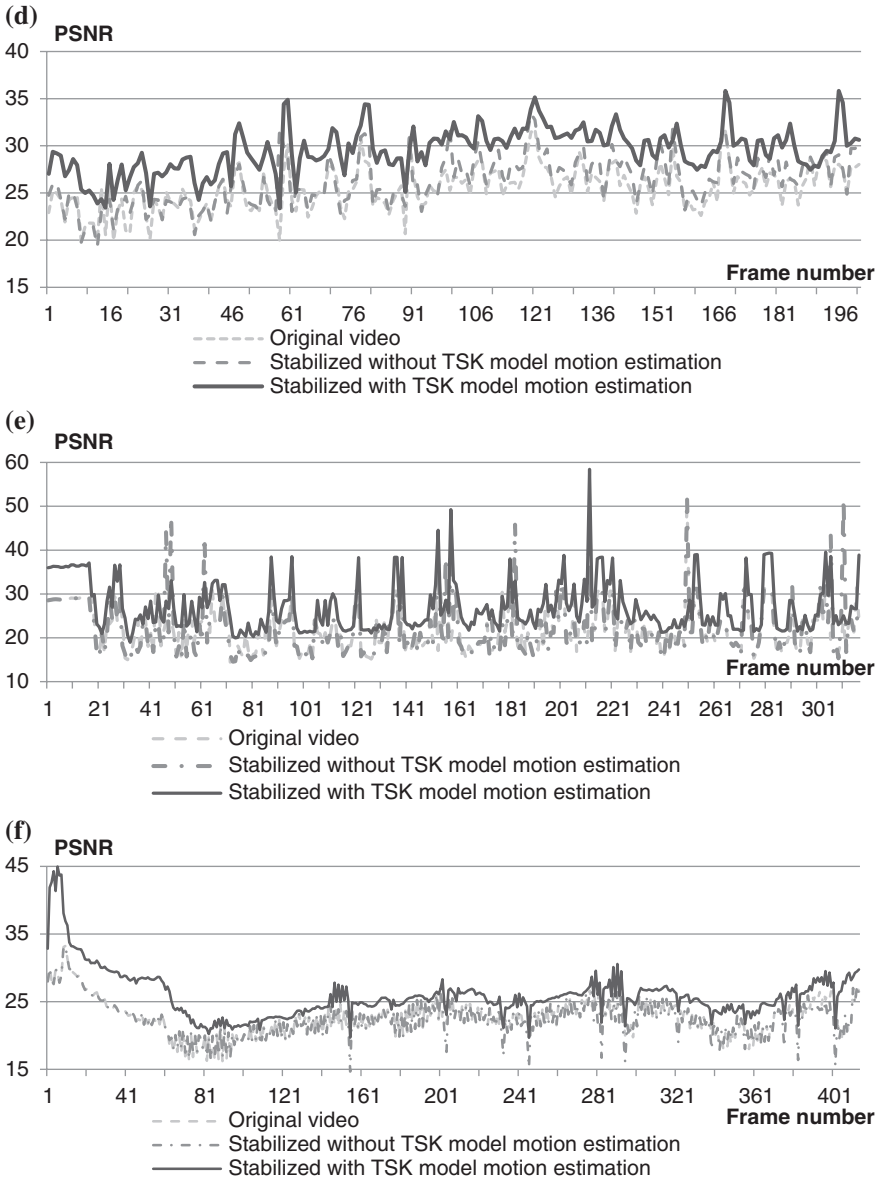


Fig. 5.9 (continued)

Table 5.2 contains the ITF estimations for video sequences: original, without and with TSK model application.

As it seems from Table 5.2, the video stabilization results are different for various video sequences because of varied foreground and background content,

Table 5.2 ITF estimations for static scenes

Video sequence	ITF estimations, dB		
	Original	Without TSK model	With TSK model
“SANY0025_xvid.avi”	20.5389	21.09076	23.79189
“lf_juggle.avi”	24.30286	24.37177	28.06012
“akiyo.avi”	35.92952	39.14661	39.53257
“EllenPage_Juggling.avi”	24.65855	25.23049	28.58255
“Butovo_synthetic.avi”	22.26415	27.19789	27.20789
“road_cars_krasnoyarsk.avi”	22.70482	22.80707	25.91258

Table 5.3 Comparison of stabilization algorithms for static scenes

Video sequence	Algorithm							
	Deshaker		Warp stabilizer		Video stabilization with robust L1 optimal camera paths		DVS analyzer	
	ITF, dB	Time, s	ITF, dB	Time, s	ITF, dB	Time, s	ITF, dB	Time, s
“SANY0025_xvid.avi”	23.53	1.33	22.7	1.87	22.74	1.34	23.79	0.17
“lf_juggle.avi”	26.65	1.22	24.41	1.64	26.15	1.18	28.06	0.15
“EllenPage_Juggling.avi”	25.61	3.53	26.68	4.53	27.33	3.17	28.58	3.54
“road_cars_krasnoyarsk.avi”	22.31	1.45	21.48	2.15	25.2	1.29	25.91	0.24

moving objects, a luminance, a noise, and the shooting condition. The use of the TSK model provides the increment of ITF estimations up on 3–4 dB or 15–20 %.

The stabilization and temporal results of video sequences from Table 5.1 by existing software tools such as “Deshaker”, “WarpStabilizer”, “Video Stabilization with Robust L1 Optimal Camera Paths”, and our “DVS Analyzer” are located in Table 5.3.

The ITF estimations of the proposed software tool “DVS Analyzer” provides better results (at average 1–3 dB or 5–15 %) with the lower processing time relatively the existing software tools.

5.6 Conclusion

The proposed approach for video stabilization of static scenes includes the automatic detection and improvement of blurred frames as well as the LMVs and GMVs estimations using the TSK model in order to separate a camera motion from a motion of moving objects and provide a scene alignment. The development of deblurring method with applied fuzzy logic rules for better motion estimations is discussed in this chapter. All methods and algorithms were realized by the

designed software tool “DVS Analyzer”, v. 2.04. During experiments, the PSNR and the ITF estimations were received for six video sequences with static camera shooting. The ITF estimations increase up on 3–4 dB or 15–20 % relative to the original video sequences.

The development of advanced motion inpainting methods and algorithms for the DVS task and also fast realization of algorithms without essential accuracy fall for pseudo real-time application are the subjects of interest in future researches.

References

1. Marcenaro, L., Vernazza, G., Regazzoni, C.S.: Image stabilization algorithms for video surveillance applications. *Int. Conf. on Image Process.* **1**, 349–352 (2001). Thessaloniki, Greece
2. Peng, Y.C., Liang, C.K., Chang, H.A., Chen, H.H., Kao, C.J.: Integration of image stabilizer with video codec for digital video cameras. In: *International Symposium on Circuits and Systems*, pp. 4781–4784. Kobe (2005)
3. Rawat, P., Singhai, J.: Review of motion estimation and video stabilization techniques for hand held mobile video. *Int. J. Sig. Image Process.* **2**(2), 159–168 (2011)
4. Tanakian, M.J., Rezaei, M., Mohanna, F.: Digital video stabilization system by adaptive motion vector validation and filtering. In: *International Conference on Communications Engineering*, pp. 165–183. Zahedan (2010)
5. Tanakian, M.J., Rezaei, M., Mohanna, F.: Digital video stabilization system by adaptive fuzzy filtering. In: *Proceedings of the 19th European Signal Processing Conference*, pp. 318–322. Barcelona (2011)
6. Battiato, S., Gallo, G., Puglisi, G., Scellato, S.: Fuzzy-based motion estimation for video stabilization using SIFT interest points. In: *SPIE Electronic Imaging 2009—System Analysis for Digital Photography V EI-7250*, pp. 1–8 (2009)
7. Acharjee, S., Chaudhuri, S.S.: Fuzzy logic based three step search algorithm for motion vector estimation. *Int. J. Image Graph. Sig. Process.* **2**, 37–43 (2012)
8. Gullu, M.K., Erturk, S.: Membership function adaptive fuzzy filter for image sequence stabilization. *IEEE Trans. Consum. Electron.* **50**(1), 1–7 (2004)
9. Sugeno, M.: *Industrial applications of fuzzy control*. Elsevier Science Inc., New York (1985)
10. Kyriakoulis, N., Gasteratos, A.A.: Recursive fuzzy system for efficient digital image stabilization. *Advan. in Fuzzy Syst.* **2008**, 1–8 (2008)
11. Shen, Y., Guturu, P., Damarla, T., Buckles, B.P., Namuduri, K.R.: Video stabilization using principal component analysis and scale invariant feature transform in particle filter framework. *IEEE Trans. on Consum. Electron.* **55**(3), 1714–1721 (2009)
12. Tsai, D., Lai, S.: Defect detection in periodically patterned surfaces using independent component analysis. *Pattern Recogn.* **41**(9), 2812–2832 (2008)
13. Kim, N., Lee, H., Lee, J.: Probabilistic global motion estimation based on Laplacian two-bit plane matching for fast digital image stabilization. *EURASIP J. Adv. Sig. Process.* pp. 1–10 (2008)
14. Pun, C.M., Lee, M.C.: Log-polar wavelet energy signatures for rotation and scale invariant texture classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(5), 590–603 (2003)
15. Shakoor, M.H., Moattari, M.: Statistical digital image stabilization. *J. Eng. Technol. Res.* **3**(5), 161–167 (2011)
16. Cho, S., Wang, J.: Video deblurring for hand-held cameras using patch-based synthesis. *ACM Trans. Graph. (SIGGRAPH 2012)* **31**(4), 1–64 (2012)
17. Cho, S., Lee, S.: Fast motion deblurring. *ACM Trans. Graph.* **28**(5), 1–8 (2009)
18. Shan, Q., Jia, J., Agarwala, A.: High-quality motion deblurring from a single image. *ACM Trans. Graph.* **27**(3), 1–10 (2008)

19. Whyte, O., Sivic, J., Zisserman, A., Ponce, J.: Non-uniform deblurring for shaken images. *Int. J. Comput. Vis.* **98**(2), 168–186 (2012)
20. Gupta A., Joshi N., Zitnick C.L., Cohen M., Curless B.: Single image deblurring using motion density functions. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *Computer Vision—ECCV 2010*, LNCS 6311, Part 1, pp. 171–184. Springer, Heidelberg (2010)
21. Cai, J., Walker, R.: Robust video stabilization algorithm using feature point selection and delta optical flow. *IET Comput. Vis.* **3**(4), 176–188 (2009)
22. Cho S., Matsushita Y., Lee S.: Removing non-uniform motion blur from images. In: 11th International Conference on Computer Vision, pp. 1–8. Rio de Janeiro (2007)
23. Matsushita, Y., Ofek, E., Ge, W., Tang, X., Shum, H.Y.: Full-frame video stabilization with motion inpainting. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**(7), 1150–1163 (2006)
24. Favorskaya, M., Buryachenko, V.: Video stabilization of static scenes based on robust detectors and fuzzy logic. *Front. Artif. Intell. Appl.* **254**, 11–20 (2013)

Chapter 6

Development of Architecture, Information Archive and Multimedia Formats for Digital e-Libraries

Margarita Favorskaya and Mikhail Damov

Abstract The efficient management of Digital e-Library Warehouse (DLW) and Library Information Archive (LIA) provides the advanced facilities for customers, who use stationary and non-stationary working places. The chapter involves a description of library standards and document formats, and also a multi-level content-dependent architecture of the DLW. The Electronic PUBLication (EPUB) format was extended by the original page markup and better multimedia data representation especially for the end users, who work in groups. The usability of the DLW is based on skill acquisition with collaboration of advanced software/hardware technologies. The experimental tests on vulnerabilities and efficiency estimations permitted to choice the content management system “1C-Bitrix” as a middleware in the proposed multilevel architecture.

Keywords Digital e-Library warehouse · Library information archive · Multimedia formats · e-Book navigation

6.1 Introduction

The first serious attempts of digital library design are concerned to 1990–2000 thanks to the network developments (the internet and intranet architectures) and advances in software technologies (the client-server and the middleware techniques) [1]. Large repositories, retrieval and indexing techniques were implemented in the infrastructures of the Stanford Digital Library Infobus [2] and the

M. Favorskaya (✉) · M. Damov
Department of Informatics and Computer Techniques, Siberian State Aerospace University,
31 Krasnoyarsky Rabochy av., Krasnoyarsk 660014, Russian Federation
e-mail: favorskaya@sibsau.ru

M. Damov
e-mail: damov_mv@sibsau.ru

Patron-Augmented Digital Library project to support a digital scholarship [3]. In the Infobus, the documents stored in the remote information sources as objects in the hierarchical metadata model. The Personal ADaptable Digital Library Environment (PADDLE) architecture was designed as a personalization environment for knowledge end users based on flat data models [4].

The advantages of digital library technologies lay on a way of effective medium and rapid tuning of information services. A traditional knowledge organization in libraries is based on a manual indexing and classification. In digital e-libraries, a data knowledge organization provides better access to multimedia data content in knowledge systems. At present, a great variety of modern end user devices such as tablets PC and smartphones (androids, iOSes, etc.) is available. Such devices are involved in the educational process anywhere and anytime. The development of library information services is a priority direction not only for individual end users but also for users, who work in groups and need in the real time synchronization of educational tasks or business cooperation.

One of the crucial issues is a quality of knowledge locating in a Library Information Archive (LIA), which ought to be evaluated before being added to the knowledge base and then screened to end user. On the one hand, the LIA is an open resource. On the other hand, it ought to be detected from unproved and irrelevant information and also protect the author' rights in virtual space as possible.

Brief descriptions and analysis of digital libraries architectures are situated in Sect. 6.2, and the overview of existing standards and document formats is located in Sect. 6.3. Section 6.4 presents the requirements and the objectives. The proposed multilevel content-dependent architecture including server and web client parts is discussed in Sect. 6.5. The extensions of the EPUB format for multimedia data representation are introduced in Sect. 6.6. The client software design and experimental researches of vulnerabilities are situated in Sect. 6.7 with the followed conclusion in Sect. 6.8.

6.2 Related Work

The PADDLE architecture was proposed by Hicks and Tochtermann [4] as a personalization environment for knowledge workers with distributed information needs. A personalization in their context refers to the ability of user or group of users to customize or modify information objects. The PADDLE architecture was decentralized and supported available metadata services. Individual descriptors into a digital catalogue created for each user that provided a detailed level of metadata granularity. The primary functional component of this network architecture was the Customization Metadata Manager (CMDM), which positioned between client applications and the information resources. The CMDM is a server process that performs necessary functions in response to the client application requests. The customization metadata storage was structured according to content of individual users or user groups. A client application was implemented and integrated

into the environment that enabled the access to information objects such as images, HTML documents, etc. from the remote sources. The user could define a new metadata field for the descriptions of the documents according to a new criterion, hide or delete one or more fields in the document descriptions, if they were not relevant to the current task.

The CMDM has been implemented in Java and is based upon the Netscape Fastrack Web server. A Microsoft Access database provided the functionality of the customization metadata storage. The communication between the CMDM and the external or remote information systems was flexibly defined using abstract Java classes as a basis of different protocols application. Three different information systems had been integrated in the prototype environment:

- The first system contained a collection of over 2,000 Microsoft Office documents.
- The second system consisted of over 100,000 HTML documents.
- The third system included the electronic thesis archive of Aalborg University, Denmark.

Such components as data repository, metadata manager, search and retrieval components were included in Multimedia dIgital Library Object Server (MILOS) based on a Multimedia Content Management System (MCMS) [5]. The MILOS MCMS had been developed using the Web Service technology (.NET, EJB, CORBA, etc.), which provided the complex support for standard operations owing to any DBMS (authentication, encryption, replication, distribution, data loading, etc.). Three basic MCMS functions were connected with the management of different documents, the operation with heterogeneous metadata, and the support of views' applications based on the metadata schema. The metadata storage and retrieval component supported the native XML databases and XML encoded metadata standards such as MPEG-7. Such functionalities permitted the automatic extraction of color histograms, textures, shapes, etc. from visual documents. This software provides an automatic classification of XML documents and feature similarity search. The service of digital library applications with heterogeneous data is the main advantage of the MILOS.

The Greenstone software of digital library supports various data formats, frequent version updates, and an easy installation under different operating systems such as Windows, Linux, UNIX, and Macintosh [6]. According to the author opinion, the Greenstone system is one of the leading applications for digital libraries' creation. It can handle documents in multiple languages including Arabic text, pictures, audio information, and videos. Additionally, the Greenstone system supports various file formats (MS-Word, PDF, HTML, PostScript, JPEG, and GIF). The indexing provides the access to the defined sections within a document (title, chapter, or paragraph) or to the whole document. Stemming and case-sensitive searching are also available. The Greenstone software has a built-in access control mechanism, which allows the work with documents and collections by using a password protection scheme. Highly enriched annotated metadata describe attributions, contents, formats, usage conditions, and rights. This system may be concerned to highly interoperable software. Its server serves any collection by the Open Archives Protocol for Metadata Harvesting (OAI-PMH) and Z39.50 protocol [7].

A framework for wireless multimedia digital library using Grid and Peer-to-Peer (P2P) technology under the Linux and FreeBSD operating systems was proposed in [8, 9]. The digital data are stored in the cluster, and users can access these data from anywhere and anytime securely. The proposed system architecture contains five layers:

- Layer 1. The network fabric layer for communication between heterogeneous peers or nodes.
- Layer 2. The P2P middleware layer, which controls the peers connected.
- Layer 3. The grid middleware layer, which transfers files and supports the data storage.
- Layer 4. The components of proposed system.
- Layer 5. The end user nodes with such services as sharing, browsing, searching, and downloading of multimedia data.

The system has a set of sample scenarios for different end users. The increment a number of nodes trends to the bandwidth reduction and the increase of retrieval time. The authors of research [9] have proposed to incorporate their framework in a cloud computing technology.

The knowledge management in the academic libraries is a multi-dimensional field of study and practice due to the knowledge access by various ways:

- The knowledge links or networking with other libraries and institutions.
- The buying of training programs.
- Symposiums, conferences, and workshops.
- The subscription to listservs and online or virtual communities.
- The buying manuals, reports, and research reports, etc.

Some interesting ideas of knowledge management in digital e-libraries are discussed in researches [10–15].

6.3 Overview of Standards and Document Formats

The main standards including e-catalog standards, access standards, and meta data standards are situated in Fig. 6.1. Consider the most popular from them.

Formats for storage and exchange of book catalog cards are based on MARC21, UNIMARC, MARCXML standards, or national standards (for example, RUSMARC for Russian Federation). The machine-readable format MARC was proposed for the USA Congress Library in 1960 as a format for cataloging, classification, and propagation of bibliographic information. The enhanced format MARC21 was developed later.¹ However, each country designed its national format, which did not possess compatibility. Thus, the UNIMARC format was

¹MARC Standards, Available online, <http://www.loc.gov/marc/marcdocz.html>.

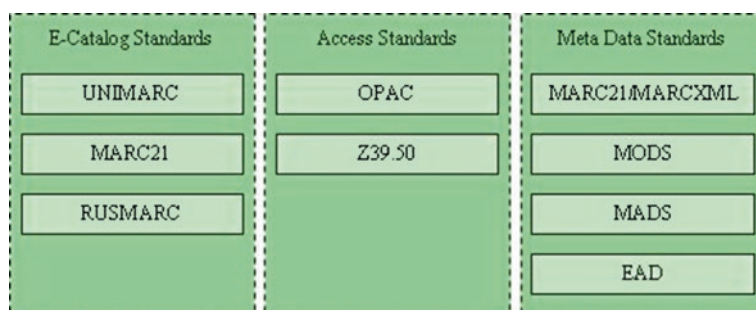


Fig. 6.1 The main standards for digital e-libraries

designed to remove the accumulative mismatches. This format classifies many information objects such as office documents, modern books, ancient books, book series, music documents, graphic materials on opaque basis, audio materials, video materials, electronic resources, maps, etc.²

There are two field categories in the UNIMARC format: common and special. The common fields describe the any type document. The special formats are used for documents of defined types as block constructions 0XX, 1XX, and 2XX. Block 0XX includes fields for storage of unique international identification numbers such as ISBN, ISSN, ISMN, etc. Fields of block 1XX encode the types of media: books (105), series (110), videos (115), graphic materials (116), and electronic resources (135). Fields in block 2XX specify the detailed information; for example, field 230 contains the specific information about electronic resources. The main advantages of the MARC formats are their facilities and a possibility of widely distributed bibliographic materials. However, the MARC formats require the special machine tools. This disadvantage was removed by the MARCXML format, which uses the usual tools.

The Z39.50 protocol is a widely spread protocol used in the DLW for information exchange in the distributed non-uniform environment of the client-server architecture.³ Such protocol is used for gateway organization between geographical distributed libraries systems.

The e-books may be represented in a printed version (in this case scan formats are used) and/or an electronic version. The common formats are Word Document (DOC), Portable Document Format (PDF),⁴ DjVu,⁵ Hypertext (HTML),⁶

²UNIMARC formats and related documentation, <http://www.ifa.org/publications/unimarc-formats-and-related-documentation>.

³Z39.50—Supports information retrieval among different information systems, <http://www.loc.gov/z3950/agency/>.

⁴PDF Reference and Adobe Extensions to the PDF Specification, http://www.adobe.com/devnet/pdf/pdf_reference.html.

⁵What is DjVu, <http://www.djvu.org/resources/whatisdjvu.php>.

⁶HTML 5.1 Nightly, <http://www.w3.org/html/wg/drafts/html/master/>.

Microsoft Help (CHM),⁷ Fiction Book (FB2), and Electronic Publication (EPUB).⁸ Also hardware-dependent formats exist for the special reader devices of e-books.

The DOC format was one of historically first formats (the beginning of 1980s) proposed by corporation “Microsoft”. Software editor tools (commercial and free) support the high possibilities for text formatting, equations, and image embedded functions as Object Linking and Embedded (OLE) objects. The poor defense of text editing and absence of full possibilities in the non-fee versions are the main disadvantages of DOC format.

The PDF format was designed by company “Adobe” in 1993. Primarily it was applied for storage and mapping only printing production; then this format became support the text, forms, bitmapped and vector graphics, and multimedia elements. Also this format has a mean of document defense—the e-signature. An accurate transition from printed version to e-version on the assumption of necessary embedded fonts into document is a main advantage; otherwise critical errors are displayed. For these purposes, commercial and free software tools exist (the last one is in a view of virtual printer). However, the software tools for PDF documents reading are free.

The DJVU format was developed by company “AT&T Research” in 1998. This is a format for storage images with losses and scanned books or other documents with the text layer extension. Free software tools are used for browse and create documents in such format.

The HTML format was development by consortium “W3C” in 1995 as a format for internet sites design. At the early stages, the editing functions were restricted by font tuning parameters, paragraph formatting, image embedding, etc. At present, HTML 5.1 version with a set of CSS format styles possesses the full formatting possibilities of text and the limited support of multimedia content. The HTML format is a usual text. Therefore, one can work without special software tools but defense from unauthorized access is absent. Also it is oriented on hardware, and the camera-ready documents will depend from the end devices (monitor, tabletPC, smartphone, e-book reader, etc.). The integrity of the HTML document is a problematic because any such document is a set of files: a main file, style tables, images, objects’ files, etc. Any HTML document can be created using usual text editor or special visual HTML editor. The internet browsers are applied for their mapping.

Corporation “Microsoft” designed the CHM format (as a help-format) for operating system Windows in 1996. The CHM format includes an encoded HTML document and attached files. It has all advantages of the HTML format (with smaller sizes of file) but is documented and oriented only to the operating system Windows. Some converter tools exist for other software platforms.

The FB2 format is a XML-file with restricted set of text logic formatting (authors, title, paragraph, image, etc.). Therefore, the external view of document

⁷Microsoft HTML Help, <http://msdn.microsoft.com/en-us/library/ms670169.aspx>.

⁸International Digital Publishing Forum. EPUB 3 Overview, <http://idpf.org/epub/30/spec/epub30-overview.html>.

depends from selected browser. The FB2 format is a format for storage, and distribution of fiction e-books, and cannot be used for educational and scientific books because some necessary functions are absent (for example, the equations editing is not supported).

The EPUB format was proposed by International Digital Publishing Forum in 2007. This is an encoding container (zip), which includes main files of e-books in the XHTML format, attached files (styles files, embedding multimedia files, etc.), and file of objects' description. A large set of software tools for view and creation of documents, support of embedded equations and multimedia are the advantages of this format. A restricted support of multimedia for educational and scientific purposes and an absence of original book copy are the main disadvantages of the EPUB format. However, this format is the opened and the enhanced format and provides a support of own XML-elements markup. Thus, the EPUB format is a more suitable format for DLW as an educational and scientific resource. For compatibility, it is needed to support PDF and DJVU formats.

6.4 Requirements and Objectives

The requirements of the DLW include the conventional functions connecting with collection, processing, dissemination, storage, and utilization of learning and scientific actual information:

- The work under Linux, Apache, Mysql, PHP (LAMP-server).
- A large set of standard editing functions; a usability of functions.
- The API functions with high functionality.
- The multi-functional authorization system.
- The embedded tools of multimedia content management.
- The embedded tools of security, virus attack protection.
- The embedded tools of component diagnostics.
- The embedded tools of backup management.
- The embedded tools of software and hardware scaling management.
- A fast access to documents.
- The technical support service.
- The training of on-line users and handbook preparation.
- The absence of critical vulnerability into software versions.

The knowledge management deals with knowledge creation, acquisition, packaging, and application or reuse of knowledge. This process basically involves four steps: knowledge collection, organization, data protection and presentation, and dissemination of knowledge information. The main objectives of knowledge management are the following:

- The collection, processing, storage, and distribution of knowledge.
- The creation of knowledge repositories and management of knowledge.

- The support of multimedia data.
- The organization of effective and fast research.
- The protection of the intellectual property rights.
- A relationship between e-library and separate end users.
- A relationship between e-library and groups of end users.

6.5 Proposed Architecture of Digital e-Library Warehouse

The collaboration of knowledge and data based technologies requires the common use of knowledge by many applications, the independence of knowledge performance during information editing, and the knowledge distribution in geographical distributed database. The achievements of geographical information systems, drawing databases, multimedia information systems, and document information search systems can be included in the DLW as necessary functionalities. The traditional multimedia system processes conventional data types (character, numeric, data/time as in any relational database), images and collections of images, video sequences, audio sequences, other graphic types, and embedded data from text processors or electronic tables. The multimedia data require not only high memory volumes but need in high bandwidth networks especially for global distributed environment. That is why, a Wide Area Network supports an asynchronous transfer mode of data or switched multimegabit data service. The management of multimedia information is often based on temporal and spatial properties of data (for example, video data conjunction from several sources requires the correct order of frames, or a multimedia journal includes the different data types even on one page). The coding standards such as Joint Photographic Expert Group (JPEG), Moving Picture Coding Expert Group (MPEG), and Multimedia and Hypermedia information coding Expert Group (MHEG) are also ought to be integrated into DLW technologies.

The proposed architecture of the DLW is a simplified version of traditional warehouse (data shop-case) [16]. It is represented in Fig. 6.2.

The proposed architecture includes three levels: Warehouse Level, Web Database Level, and End User Level. The Data Sources provide information from various sources (jointed network, digital libraries, internet, and intranet). The Warehouse contains various types of data. However, they are divided as data types and knowledge types. The Partial Joint Data include text files, the Deep Jointed Data are used for complex query execution. The Metadata contain the data descriptions needed for Web database. The Knowledge and knowledge rules accomplish the intelligent indexing and the retrieval in the DLW.

The Web Database Manager has own functions of authorization, security, Web applications, and device drivers. The Web Database Query Manager is a middle-ware between the end users and the Warehouse. At present, more interesting aspect is a management of mobile devices connecting with the DLW. Also the work in groups requires following investigations and software tools design.

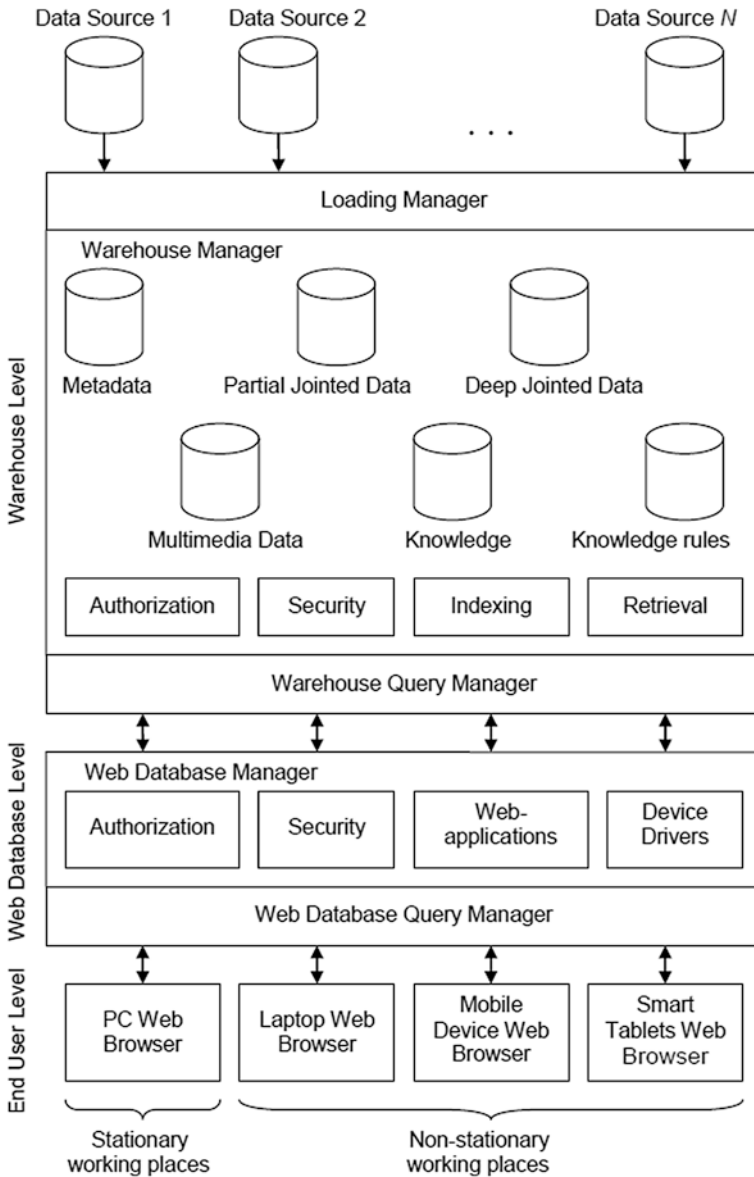


Fig. 6.2 Architecture of the DLW

6.6 Proposed EPUB Format Extensions

Two novel possibilities in the EPUB format were proposed: the original page markup (tag <PAGE>) and the work with audio and video fragments. A block tag <PAGE> in page content </PAGE> was introduced into the structure of the XML/XHTML-documents, for example:

```

<?xml version="1.0" encoding="UTF-8"?>
<html>
<body>
<page>
<p>Lorem ipsum dolor sit amet, consectetur adipiscing elit. Maecenas port-
titor congue massa. Fusce posuere, magna sed pulvinar ultricies, purus lectus
malesuada libero, sit amet commodo magna eros quis urna. Nunc viverra imper-
diet enim. Fusce est.</p>
<p>Vivamus a tellus. Pellentesque habitant morbi tristique senectus et netus et
malesuada fames ac turpis egestas. Proin pharetra nonummy pede. Mauris et orci.
Aenean nec lorem.</p>
</page>
<page>
<p>In porttitor. Donec laoreet nonummy augue. Suspendisse dui purus, scler-
isque at, vulputate vitae, pretium mattis, nunc. Mauris eget neque at sem venenatis
eleifend. Ut nonummy.</p>
<p>Fusce aliquet pede non pede. Suspendisse dapibus lorem pellentesque
magna. Integer nulla. Donec blandit feugiat ligula. Donec hendrerit, felis et imper-
diet euismod, purus ipsum pretium metus, in lacinia nulla nisl eget sapien.</p>
</page>
</body>
</html>

```

If a reading device can not support these additional tags, then the tags will be ignored. The interpretation of e-book text can be looked variously (Fig. 6.3).

There are two variants of text view: using parameters of device screen or considering the sizes of printed page. The first variant is better because physical sizes of page are supported by applying the Cascading Style Sheets (CSS). According to the user choice, the e-book can be drawn by using parameters of device screen or by physical page markup. The last possibility is very useful during the work in groups. This permits to organize references uniformly (on the physical pages). Such reference is pointed conventionally, for example <http://example.com/test.epub#page1>.

At present, a multimedia content requires more detailed possibilities than other formats. The tags <VIDEO> and <AUDIO> of markup language HTML5 are supported by standard EPUB 3 but finally are not asserted. The proposed extension of this standard by video file indexing, navigation, and scene description is represented in the following manner:

```

<video src="test.mp4" poster="test.jpg">This is fallback content to display if
the browser does not support the video element.
<scene start="01:04" name="Scene 01">
<p>In porttitor. Donec laoreet nonummy augue. Suspendisse dui purus, scler-
isque at, vulputate vitae, pretium mattis, nunc. Mauris eget neque at sem venenatis
eleifend. Ut nonummy.</p>
</scene>
<scene start="05:10" name="Scene 02">

```

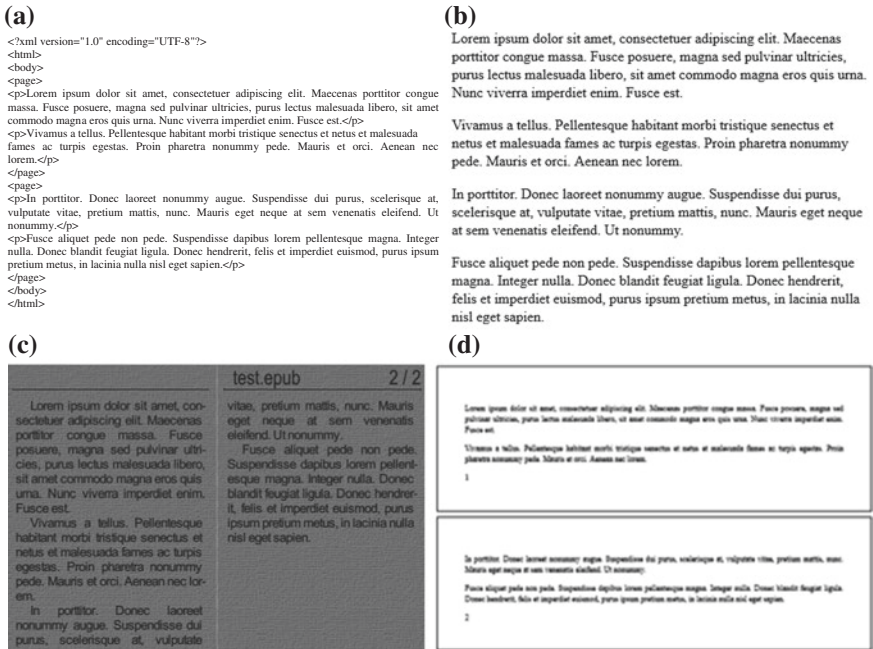


Fig. 6.3 Text view representation. a XML markup of book. b Firefox 16 browser. c CoolReader program 3. d Text view in browse with original pages

```

<p>In porttitor. Donec laoreet nonummy augue. Suspendisse dui purus, scelerisque at, vulputate vitae, pretium mattis, nunc. Mauris eget neque at sem venenatis eleifend. Ut nonummy.</p>
</scene>
</video>

```

The processing algorithm includes the following steps:

- Step 1. If a browser finds a tag <VIDEO> in page, then each content source (the attribute SRC) is checked.
- Step 2. The optimal source of a multimedia content reproduction is selected.
- Step 3. If rendering is possible, then there is a jump to Step 4, otherwise there is a jump to Step 6.
- Step 4. A control element is drawn for a multimedia content reproduction according to the user query. A control element can be fulfilled by image from the attribute POSTER or any frame of video sequence, if this attribute is absent or empty.
- Step 5. In tag <VIDEO>, a search of tags <SCENE> is accomplished in a cycle. For each tag <SCENE>, a list element is formed, which contains a time mark, a scene title, and a scene description.
- Step 6. The output of tag <VIDEO> content ignores the tags <SCENE>.

Such algorithm adds not only a multimedia content but also a scene navigation of embedded video sequence with extended scene descriptions and audio-fragment navigation into the e-books. The proposed standard modifications will not influence on application execution without modifications of the EPUB3 and the HTML5 standards. The results of algorithm work are situated in Fig. 6.4.

6.7 Client Software Design and Researches of Vulnerability

The client software involves the functionalities mentioned below:

- The database management of a multimedia LIA: creation, editing, and removal of electronic records of resources.
- The support of users' access to a multimedia LIA according to roles and privileges.
- The citation and search into a multimedia LIA.

The main screens of user's interface are located in Fig. 6.5.

For security investigations, open databases^{9,10} were used for receiving data about common rates (Table 6.1 presents these results) and types of vulnerabilities at last years and in 2013 (the results are situated in Table 6.2) for some Content Management Systems (CMS). All vulnerabilities were divided on three types according to parameter of a Common Vulnerability Scoring System (CVSS): high ($CVSS \geq 7.4$), medium ($4.7 \leq CVSS < 7.4$), and low ($CVSS < 4.7$).

As it seems from Tables 6.1 and 6.2, the majority of systems are not available for DLW design according to the security requirements. The "UMI CMS" and the "1C-Bitrix" have approximately closed functions. The "1C-Bitrix" uses PHP language and the "UMI CMS" has own script language that is a disadvantage for user training and project support. Therefore, the CMS "1C-Bitrix" is the most appropriate systems providing necessary functionality and project security.¹¹

The efficiency estimations of DLW based on the "1C-Bitrix" are presented in Table 6.3. These data were provided by Automatic Diagnostic Master of the "1C-Bitrix" (server with processor i7 975, 12 GB RAM, 3×1.5 GB soft RAID 5 SATA HDD). The test results show the high efficiency of designed system.

The experimental graphics of time and number of queries are represented in Fig. 6.6.

⁹Search the Secunia Advisory and Vulnerability Database, <http://secunia.com/community/advisories/search/>.

¹⁰Security Lab by Positive Technologies. National Vulnerability Database, <http://en.securitylab.ru/nvd/>.

¹¹Bitrix Site Manager. Product Features, <http://www.bitrixsoft.com/products/cms/index.php>.



00:00:00 Episode 01

6-grader Kolya Gerasimov discovers a time machine in a basement of an old house in Moscow and gets transferred into the 21st century.

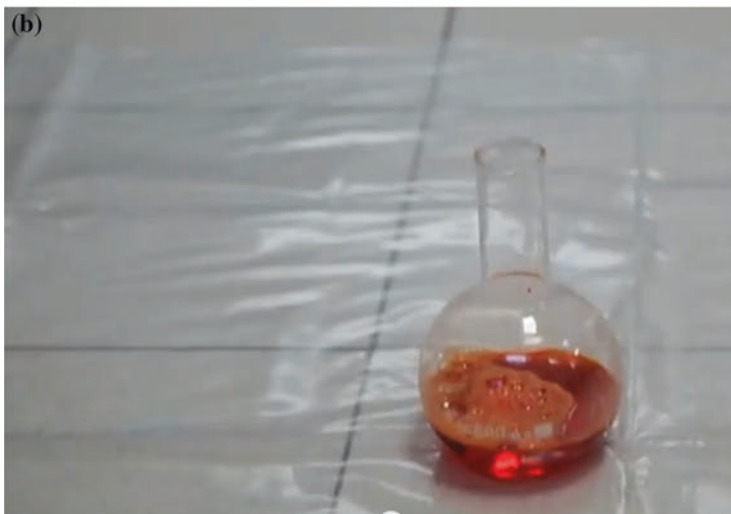
01:15:10 Episode 02

Two space pirates try to steal a device called a "Mielophone" (which can read thoughts) from Alisa Seleznyova - a girl that performs experiments with this device and animals.

02:30:46 Episode 03

Kolya brings device back to the 20th century.

1



00:00:00 Scene 01

Begin title

01:15:10 Scene 02

Chemical experiment

1

Fig. 6.4 Examples of tag <VIDEO> modification, **a** video sequence "Guest from the Future (Episode 1)", **b** video sequence "Chemistry show"

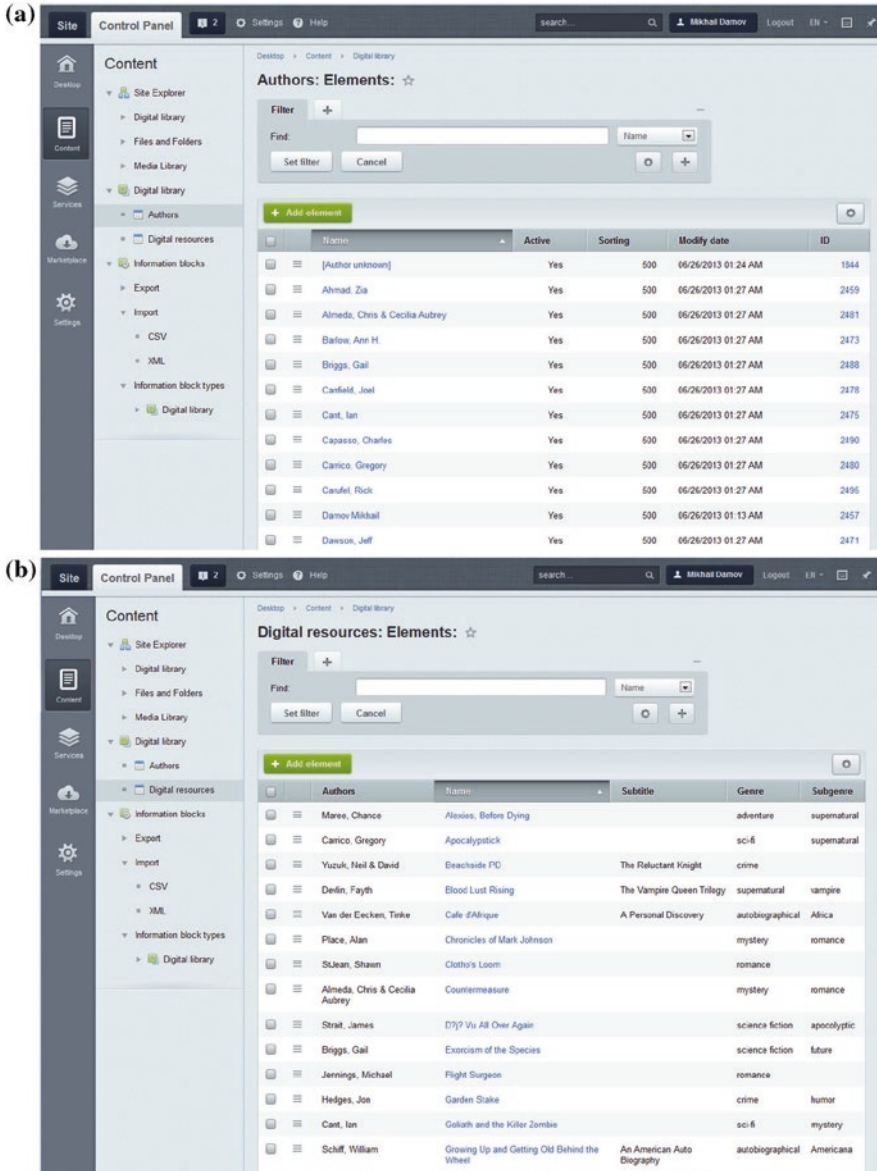


Fig. 6.5 Screens of user's interface, a an authors' list, b a books list, c the information blocks, d the edit block

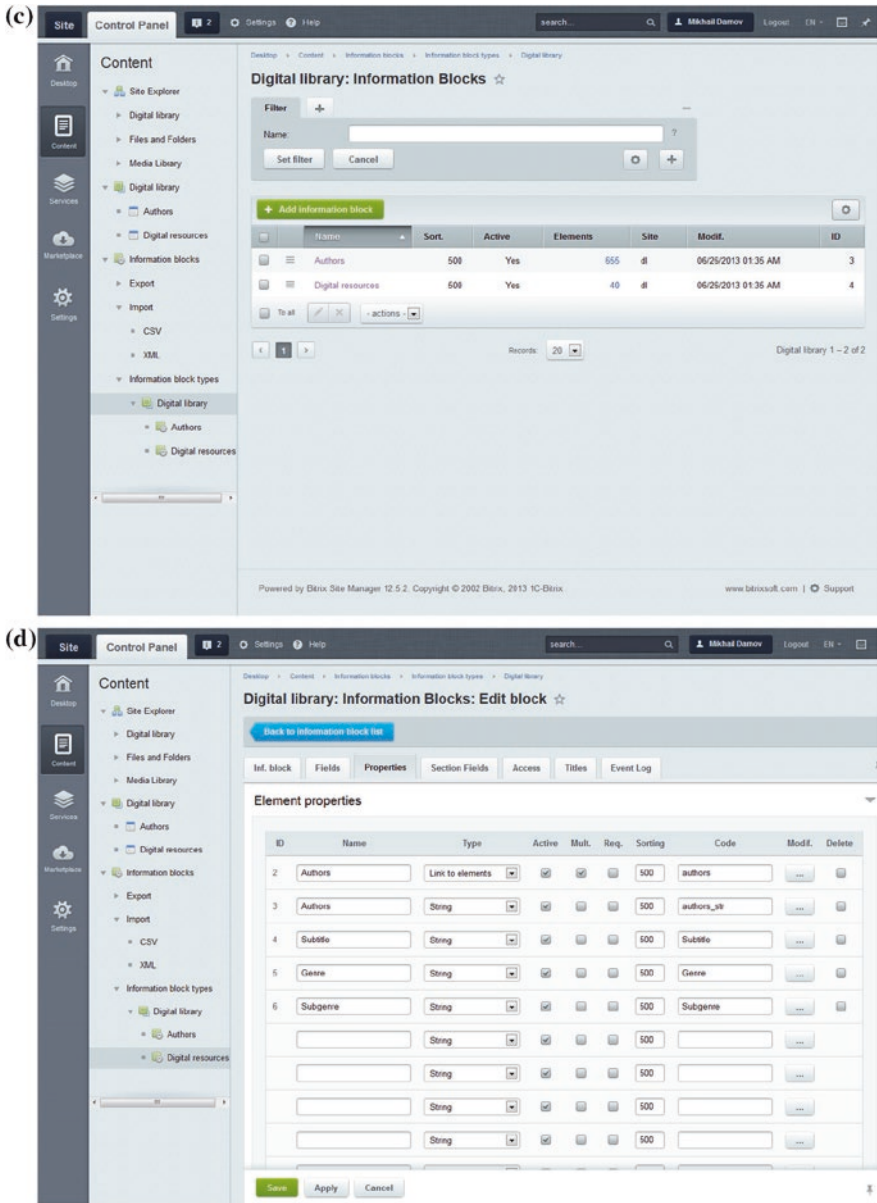


Fig. 6.5 (continued)

Table 6.1 Common rates of vulnerabilities

CMS	Rates by Secunia (Europe)	Rates by positive technologies (Russian Federation)
“1C-Bitrix”	4	7
“UMI CMS”	2	1
“Drupal”	709	576
“Joomla”	582	744
“Wordpress”	922	393
“ModX”	15	17

Table 6.2 Rates of vulnerabilities by types (recent years/in 2013)

CMS	High	Medium	Low
“1C-Bitrix”	1/0	2/0	4/0
“UMI CMS”	0/0	0/0	1/0
“Drupal”	75/2	284/>20	180/10
“Joomla”	427/1	242/12	75/0
“Wordpress”	100/7	227/>20	66/9
“ModX”	2/0	14/0	1/0

Table 6.3 Efficiency estimations of DLW based on “1C-Bitrix”

Criterion	Result value	Etalon value
Mean time of response, s	0.0185	0.033
Processor efficiency, million operations/s, 1 core	14.2	9.0
HDD and file system efficiency, operations/s	7281	10,000
MySQL, write, query/s	2381	5600
MySQL, update, query/s	3250	5800
MySQL, read, query/s	9120	7800
Total efficiency estimations of “1C-Bitrix”, scores	54.15	30.0

One can see from graphics represented in Fig. 6.6 that the system can process up to 180 queries/s for 70 synchronous interconnections without the efficiency decrease. Therefore, the theoretical efficiency achieves up to 15 million of queries per day.

The received results satisfy to criteria and recommendations of productivity except parameters queries/s, write and queries/s, update. However, this fact is not essential because more that a half time the Data Base Management System (DBMS) works in a read mode (according to experimental data, this parameter is near 60.9 % for DBMS MySQL).

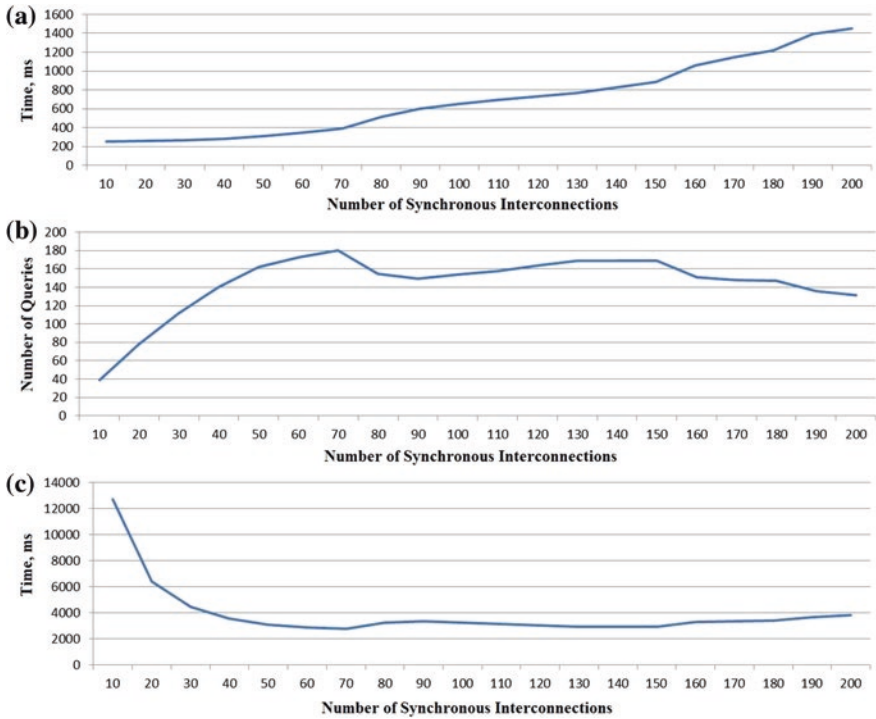


Fig. 6.6 Experimental graphics, **a** time for one query execution, **b** number of queries per unit, **c** total test time

6.8 Conclusion

Two aspects of the DLW were investigated: the architecture building and the enhanced formats of multimedia content. The modern necessary functionalities of educational and scientific resources are well realized in a three-level architecture including Warehouse Level, Web Database Level, and End User Level. The last one supports the access from various mobile devices. The analysis of existing formats permitted to choose the EPUB format as more appropriate format for multimedia content management.

The novel possibilities of the EPUB format were proposed: the original page markup (tag <PAGE>) and the work with audio and video fragments from sequences. The first possibility is important for a group work of end users, when the current pages are uniformly assigned. The second possibility is useful for more accurate delivery of video and audio resources. Video sequences are divided into the scenes or music fragments with corresponding descriptions. Also common rates and types of vulnerabilities of the proposed DLW were analyzed. The received results satisfy to criteria and recommendations of productivity of similar software tools based on the non-expensive middleware.

Acknowledgments This study was supported by The Ministry of education and science of Russian Federation, project 14.B37.21.0457 “Development of Internet-platform of broadband access to the library multimedia resources”, 2012–2013.

References

1. Yin, L.T.: Handbook of Research on Digital Libraries: Design, Development, and Impact. Hershey, New York (2009)
2. Paepcke, A., Wang Baldonado, M.Q., Chang, C.-C.K., Cousins, S., Garcia-Molina, H.: Using distributed objects to build the stanford digital libraries infobus. *IEEE Comput.* **32**(2), 80–87 (1999)
3. Goh, D., Leggett, J.: Patron Augmented Digital Libraries. In: The 5th ACM Conference on Digital Libraries, pp. 153–163. ACM, New York (2000)
4. Hicks, D., Tochtermann, K.: Personal digital libraries and knowledge management. *J. Univ. Comput. Sci.* **7**(7), 550–565 (2001)
5. Amato, G., Gennaro, C., Rabitti, F., Savino, P.: Milos: A Multimedia Content Management System for Digital Library Applications. In: The 8th European Conference on Digital Libraries (ECDL), pp. 14–25. Springer, Berlin (2004)
6. Elaïess, R.: Greenstone Open Source Digital Library Software in the Context of Arabic Content. *Int. J. Digit. Inf. Wirel. Commun. (IJDIWC)* **2**(2), 181–196 (2012)
7. Goutam, B.: An evaluative study on the open source digital library software for institutional repository: special reference to Dspace and greenstone digital library. *Int. J. Libr. Inf. Sci.* **2**(1), 1–10 (2010)
8. Renuga, A.R., Sudhasadasivam, G.: P2P information retrieval framework for digital library system. *J. Appl. Theor. Inf. Technol.* **5**(3), 301–306 (2009)
9. Arulanandam, S., Jaganathan, S., Avula, D.: P2P and grid computing: opportunity for building next generation wireless multimedia digital library. *EURASIP J. Wirel. Commun. Network.* **165**, 1–16 (2012)
10. Sharma, C.K., Gupta, S.: Knowledge management: its application in research in social science. *J. Libr. Inf. Technol.* **3**(2), 1–5 (2007)
11. Roknuzzaman, Md, Umemoto, K.: Knowledge management’s relevance to library and information science: an interdisciplinary approach. *J. Inf. Knowl. Manag.* **7**(4), 279–290 (2008)
12. Luczak-Rsch, M., Heese, R., Paschke, A.: Future content authoring. *Mag. Semant. Web* **11**, 17–18 (2010)
13. Rajurkar, M.U.: Knowledge management in academic libraries. *Int. J. Parallel Distrib. Syst.* **1**(1), 5–8 (2011)
14. Grassi, M., Morbidoni, C., Nucci, M.: Semantic web techniques application for video fragment annotation and management. In: International Conference Analysis of Verbal and Nonverbal Communication and Enactment, pp. 95–103. Springer, Berlin (2011)
15. Grassia, M., Morbidoni, C., Nucci, M., Fonda, S., Ledda, G.: Pundit: Semantically Structured Annotations for Web Contents and Digital Libraries. In: 2nd International Workshop on Semantic Digital Archives (SDA 2012), pp. 49–60. Springer, Berlin (2012)
16. Favorskaya, M., Damov, M.: Architecture and formats of digital e-Library warehouse. *Front. Artif. Intell. Appl.* **254**, 21–30 (2013)

Chapter 7

Layered Ontological Image for Intelligent Interaction to Extend User Capabilities on Multimedia Systems in a Folksonomy Driven Environment

Massimiliano Dal Mas

Abstract This chapter describes a method for enabling an ontological interaction on video clip shown on ubiquitous systems as a computer monitor, mobile, or tablet. We use a layered representation based on semantic texton forests to obtain spatiotemporal object attributes. The interface is created by extracting object information from the video with a human based computation to obtain a richer semantics of attribute to bridge the semantic gap between words describing an image and its visual features. Users can navigate and manipulate objects displayed on video by associating semantic attributes and comments evaluated by the data and sentiment extraction. Folksonomy tags are extracted from users' comments to be used in a dynamical driven system (Folksodriven). This chapter documents the advantages on a case application for advertisement inside the objects displayed on a video and argues how the method proposed may be preferred to the more traditional video advertisement.

7.1 Introduction

Interactive systems can encourage people to perform chores that they ordinarily consider boring such as completing surveys, shopping, filling out tax forms, or reading web sites. Taking advantage of humans' psychological predisposition to engage in gaming it's possible to extend user capabilities in new ways. We explore a different system approach to enrich user interactions on video contexts towards ubiquitous Human Computer Interaction (HCI) on systems as computer monitor, mobile, or tablet.

Nowadays towards a preprocessing of the video it's possible to detect objects, tracking associated regions and constructing a constant background image. The proposed "Layered Ontological Image Interaction" in video contexts defines a

M. Dal Mas (✉)
via M. Gioia 137, Milan, Italy
e-mail: me@maxdalmas.com
URL: <http://www.maxdalmas.com>

model for an interface built on object of interest displayed in videos. Our approach for object representation in videos allows a user to directly access and process objects attributes as basic video components. Making technology in videos more engaging, it is possible to encourage users to engage in desired behaviors showing them a path to mastery and autonomy to contribute in defining value data for the object description in videos.

After having extracted the underlying data structures contributed by users towards folksonomy tags, we correlate those with the source and time living. Considering those we define a structure called *Folksodrive* defined in [1–6] to bridge the semantic gap between words describing an image and its visual features that could be extremely helpful, when training data is limited. Those attributes do not help in classification, but are great for descriptions and finding unexpected attributes. The goal on working with detected object is not to find, but to describe an object learning new categories by sharing higher-level traits. The benefit from that approach are on the value data obtained that can be used for different tasks as shown in the application examples: advertisement inside the video, object of interest video navigation, or mask layer on an object of interest.

The background of related technologies is briefly described in the next section. Then, a system overview of the “dynamic learning ontology structure” is presented, focusing on the definition of the “Layered Representation” of the object of interest inside videos. That is correlated with the “Semantic Attributes” and the “Attributes extraction and Sentiment analysis” to define the “Folksodrive Bounding Box Notation”. Possible use of the proposed method is described in the “Application Examples”. Lastly, some conclusions are inferred.

7.2 Human Based Computation

Human Based Computation (HBC) is a class of hybrid techniques, where a computational process performs its function via outsourcing certain steps to a large number of human participants [7]. Towards HBC ways humans and computers can work together to accomplish a task.

7.2.1 Motivation of Human Contribution

The system proposed in this chapter is based on the human contribution to refine and enrich the recognition of objects. Therefore, it is important to motivate the participation of users during the insertion of data and any correction of those and to report abuse on the system.

In “mature community” with a strong ethical motivation or participation, e.g. “Fansubber”—community of fans that translate and subtitled video, users participate without the need for special reasons. High or low levels of contribution

are critical for sustaining collaborative user-generated content as analyzed in [8] respect the volunteers' motivations. From [8] we can deduce that a mature community is mostly based by "Intrinsic Motivation". Their contribution is based on the "Ideology and Values" (users are doing something they trust on).

While in "less mature community" that we are going to consider it will be important to leverage gamification as "social" recognition in the community or even reward schemes (towards a rewording program) net of contributions of little or no value (abuse treatment/background noise). That community is mostly based by "Extrinsic Motivation" [9].

As regards the abuse treatment, the system can provide both an automatic action via stoplist and management of discrepancies (a tag is accepted, if identified as discordant, only in case of indication by multiple users). If for instance an object is defined as "car", the system does not propose the tag "souvenir" but it could be proposed in case a significant percentage of users used that tag. That can be realized through human intervention pruning report abuse or through the presence of significant negative feedbacks (e.g. asking to the users, if the definition proposed is useful for the interpretation with a yes/no answer).

7.3 Background of Related Work

The proposed method is based on concepts and/or technical components that are well-known facts within the scope of.

An object detection system recognizes objects from an image/videos of the world performing supervised learning or unsupervised learning.

Unsupervised learning: techniques don't require explicit input of human operators to learn to recognize objects.

Google has recently registered a patent on "Automatic large scale video object recognition" [10], where the object classification stores cluster of feature vectors associated with the received visual content item. The patent describes a method, by which the company's technology will identify objects in a video without the need for the user's input. Instead of asking the creator to label objects every time, Google proposes to use a database of "feature vectors" such as color, movement, shape, and texture to automatically identify subjects in the frame through their common traits. In order to determine what different objects are—both famous, like the Eiffel Tower, and seemingly inconsequential, like a chair—Google will have an "object name repository." This object recognition technology works in a similar way to other face recognition techniques, but it takes things a step further. It can actually recognize the difference between varieties of objects, not just human faces. After recognizing an object, this is labeled with certain tags. A database repository would hold at least 50,000 object names, information, and shapes that would allow an easy identification. Using that large-scale repository, the object recognition system compares a feature vector to other feature vectors, and calculates a score based on the comparisons to determine similar and different objects.

An object name in the repository is associated to textual descriptions and the visual content items in video system. The system would still allow for manual tagging of objects and animals, but would do a lot of the work itself through an automated process. Essentially this is just another way to tag and better track videos for things like SEO.

Google has been offering up a host of additions to YouTube lately. Back in March 2012, for example, the company added one-click video editing, allowing users to automatically modify their uploaded clips. Ultimately that system makes it easier for YouTube and Google to tag and track videos without any user intervention, but it also could make it easier for us to search for them.

Supervised learning: techniques require training sets of images that rely on database of images classified with conventional attribute by direct human input.

On image processing the Spatial Pyramid Matching (SPM) [11] is a common spatial pooling method for object classification been frequently used for expressing computationally efficient approximations to scale-space representation. However, the SPM results in inconsistent image features, when computing multi-scale image features from real-world image data. When the object of interest (as a car) appears in different location within images, the SPM learning is more difficult.

The Object-centric Spatial Pooling (OCP) presented in [12] is a method that localizes the object of interest and considers the object features independently from background features.

The OCP method uses weakly supervised object detection to identify the main objects present in images, not to specify the location of objects, to improve image classification. As training it was used the hand-labeled ImageNet dataset (10,000,000 labeled images depicting 10,000 + object categories) organized according to the WordNet hierarchy for the nouns. “In the ImageNet database each node of the hierarchy is depicted by hundreds and thousands of images” (<http://www.image-net.org/explore>).

The image annotations of object bounding boxes are saved in XML files in PASCAL VOC format (<http://pascallin2.ecs.soton.ac.uk>).

7.3.1 Object Tracking

After detecting the object of interest, the object position and size are tracked in every frame. In computer vision literature, different kind of object tracking algorithms are proposed [13]. Among those we use a hybrid-tracker as described in [14].

We aim to define an object-level interaction using a region tracking, which does not improve much from dense motion information. This tracking method uses a Kalman filtering to perform a simple connected-component tracking. When objects in a video are not well separated and Kalman filter’s prediction suggests a possible overlap of objects in next frame, a reliable Mean-shift tracker is used [15].

7.4 Dynamic Learning Ontology Structure

7.4.1 Richer Semantics of Attributes

One of the criticisms addressed to unsupervised learning algorithms is that they have many hyper-parameters and variants and that exploring their configurations and architectures is an art. Unsupervised learning refers to learning problems, in which there is a small amount of labeled data and a large amount of unlabelled data. The system will not represent true picture of data, not because of data quality but because of unidentified attributes—e.g. quantitative attributes are selected as descriptive as well or vice versa. These problems are very natural, especially in domains, where collecting data can be cheap (i.e. the internet) but labeling it can be expensive or time consuming.

One problem is that pulling up on the “feature vectors” of unobserved points in high dimensional spaces is often very difficult and even intractable. One key question in unsupervised learning is how the data distribution from the unlabelled data should influence the supervised learning problem [16].

The unsupervised algorithms generally suffer from the problem of combinatorial explosion, when dealing with realistically large patterns. The problem of over fitting generally occurs for algorithms that select a single structure with highest score out of exponentially many, when having too many parameters relative to the number of observations.

A model could be excessively complex considering useless matching, pairing in shape of similarity and dissimilarity matrix. While data redundancy can cause low query performance. This loss of information becomes a problem, when the information that is lost is necessary to successfully complete the task at hand such as object classification. For example, color is often a very discriminative feature in object classification tasks. Losing color information through feature pooling would result in significantly poorer classification performance.

While on supervised learning the datasets are much less restricted and do not guarantee good correspondence, with often huge variations between annotated bounding box instances.

One solution to the problem of information loss that would fit within the feature subspace paradigm, is to consider many overlapping pools of features based on the same low-level feature set. Such structure would have the potential to learn a redundant set of invariant features that may not cause significant loss of information. However, it is not obvious, what learning principle could be applied that can ensure that the features are invariant while maintaining as much information as possible.

One possible approach to classify an object in a video is focused on tracking interesting/salient objects in the video streams based on visual attention model [17] with a semi-supervised learning on it (see Object on Layered Representation) then selecting the most robust sentiment analysis (see Attributes extraction and Sentiment analysis) to obtain richer semantic attributes (see Folksodriven Bounding Box Notation) to bridge the semantic gap between words describing an image and its visual features.



Fig. 7.1 Foreground layer (FL) is defined by the moving objects (pedestrians, bicycle, cars, and trams) while the Textural layer (TL) determines the group of objects considered (the trams here colored in *orange*) (color figure online)

7.4.2 Object on Layered Representation

Layers are used to separate different elements of an image or video. We adopt a layer representation correlated to the Object Detection (OD)—see definition on section above—for its simplicity and suitability, respect more complicated layered representations [18, 19].

We use layer representation to identify the characteristics of the objects belonging to the same layer. A layer is defined by the same kinds of objects belonging to a same object class in a frame of the video. That could be used to correlate different objects at the same layer through analysis of aggregation (e.g.: presence of cars in a video of a road), a possible application is depicted in the following section “Mask layer on object of interest” (Fig. 7.1).

The structure adopted for the automated method for individual layers detection is composed by the background and the foreground layers:

- *Background Layer* (BL) is defined to be the sensed environment without any moving object (e.g., walking person, moving vehicle). Background layer determines the “scene representation” [20] that enable deeper understanding on the contest of the video. To determine, which attributes are most relevant for describing scenes, we use the extensive Scene UNDERstanding (SUN) database that contains 899 categories and 130,519 images performing open-ended image description tasks on Amazon Mechanical Turk (AMT) [20].
- *Foreground Layer* (FL) is defined to be the sensed environment with moving objects (e.g., walking person, moving vehicle). The same kind of moving objects in a video define a layer—e.g.: all the cars define a common layer for the cars; all the pedestrians define a common layer for the pedestrians.

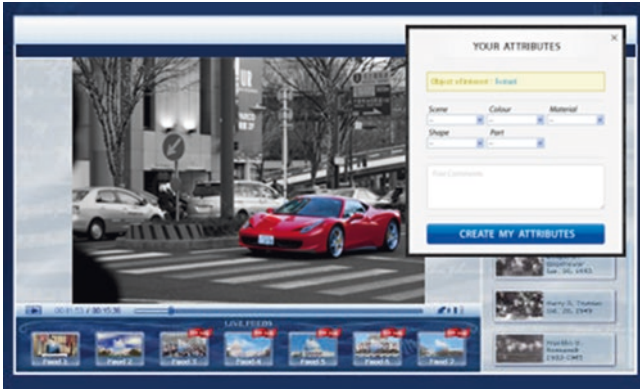
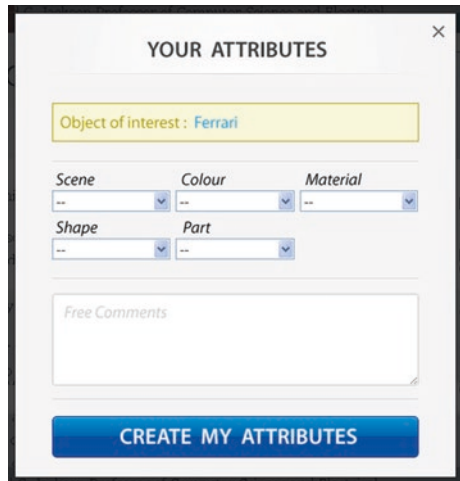


Fig. 7.2 The user is asked to fill the form with the semantic attribute and a free text comments on the object of interest chosen and automatically recognized by the system (as a “Ferrari” car in the video)

Fig. 7.3 The interface of the contribution form with the semantic attributes and free text comments for the attribute extraction and sentiment analysis



The tracking of moving objects is varying in both spatial and temporal domain. Different objects can move in the same time with different spatial and temporal domains in the same layers. Taking pedestrians—belonging of the same layer—as an example, each person can walk with different direction (spatial domain) and speed (temporal domain). The “Layered Representation” approach for object inside a video allows a user to directly access and process objects attributes as basic video components to build proactive engagement (see Figs. 7.2 and 7.3).

Temporal constraints of motion, shape and appearance for visible imagery hold for image recognition. The position of the object *O* is represented by the Object Position (*P*) depicted by the triple $P = (X, Y, T)$ to figure out its relative position $(X, Y) \in \Omega = [1, X] \times [1, Y]$ at time *T*.

$$O := (OD, FL, BL, TL, N) \quad (7.1)$$

An object O can be represented by a tuple (7.1) respect: the Object Detection (OD) defined by (7.2), the Foreground Layer (FL) belonged to the same kind of object classification, the Background Layer (BL), the Textural Layer (TL) of the single object (or group of objects) considered and the Noise on the image (N).

The Object Detection (OD) can be described by the layer representation (7.2).

$$OD = (FL - BL)TL - N \quad (7.2)$$

Textural layer (TL) of the single object considered is targeted at a maximum likelihood extraction with the background since it carries critical information about moving objects. The discriminative model described by [21] exploits texture-layout filters, novel features based on textons, which jointly model patterns of texture and their spatial layout. Semantic texton forests are ensembles of decision trees for semi-supervised learning that act directly on image pixels for the image categorization.

They are extremely fast for train and test respect k-means clustering and nearest-neighbor algorithms, and they do not need the expensive computation of filter-bank responses or local descriptors [22]. For our scope that model is used to achieve an accurate segmentation, in order to automatically determine a “mask layer” on an object, with a possible application as depicted in the following section “Mask layer on object of interest”.

Noise (N) can be suppressed in background extraction by adaptive temporal filtering while in moving objects noise interference is minimized on principal components.

7.4.3 Semantic Attributes

The Semantic Attributes are chosen by user for the object of interest from predefined values (see Interface on Object of Interest).

In this chapter, we will use five groups of semantic attributes to describe the object of interest on the video: ‘scene’, ‘colour’, ‘part’, ‘shape’, and ‘material’. From these attributes, a set of independent classifiers, we can define a semantic image descriptor.

The attributes describing the defined object are used respect the contextual information of the video because that often helps in recognizing objects.

Semantic attributes do not help in classification, but are great for descriptions and finding unexpected attributes. The goal on working with detected object (towards the bounding box) is not to find, but to describe an object.

The attributes describing the defined object are used respect the contextual information of the video because that often helps in recognizing objects.

Attributes bridge the semantic gap between words describing an image and its visual features and could be extremely helpful, when training data is limited.

Giving sufficient training data is not useful for the traditional naming task. Semantic attributes allow learning new categories by sharing higher-level traits been useful for hierarchical descriptions of objects.

7.4.4 Attribute Bounding Box Position

To further improve a classification performance on the objects displayed in video, we can incorporate bounding box annotations during training for the video or some frames of it.

$$B := (P, O, A, I) \quad (7.3)$$

To better understand a video we can determine the object position on it by utilizing location information to consider the spatial relationships among the objects, which is very important in image or video representation.

For an object represented on a video we can define an Attribute Bounding Box Position tag defined as a tuple (7.3) composed by:

- Object Position (P) is a triple $P := (X, Y; T)$ defined by spatial (X, Y) and temporal (T) variables (as defined before)
- Object represented in a video (O) is defined by an own URI;
- Semantic Attributes (A) of the object (O)
- I is defined by the relation $I = P \times O \times A$ in a Minkowski vector space [23] delimited by the vectors P, O and A (Fig. 7.3).

7.4.5 Attributes Extraction and Sentiment Analysis

Combining saliency detectors and attribute classifiers with a Sentiment extraction we can describe the significant “feeling” aspects of an object according to the users. The sentiment extraction is based on the “free text” submitted by the user for the object of interest (see Interface on Object of Interest)

Sentiment orientation analysis has been developed recently on three major directions:

- simple statistics to obtain the whole tendency of texts for orientation values [24];
- machine learning to generate classification model through the training of numerous labeled corpuses and classifying the test texts using generated model [25];
- attribute-level to mine knowledge from consumer product reviews by utilizing data mining and information retrieval technology based on a ranking mechanism taking temporal opinion quality and relevance into account [26].

The system here proposed consists of two parts:

- Natural Language Processing (NLP) pre-processing to determine attributes not classified in the five groups of semantic attributes considered before (scene, colour, part, shape, and material);
- Attribute-level analysis on those determined by the NLP.

A personal opinion has a variety of expressions and attribute or “sentiment words” with different meanings on different domains. We considered a domain-based multi-dictionary, which is divided as: attribute domain, degree words, and negative

words. In each dictionary, a variety of synonymous expression words for a sentiment is defined. A statistical analysis is performed through the users' opinions and summarized to determine: the attributes of commentary object; the sentiment words or phrase and the attribute-polarity (positive/negative) and make sentiment analysis.

7.4.6 *Folksodrive*n Bounding Box Notation

A network structure of “Folksodrive

n tags”—Folksodriven Structure Network (FSN)—was thought as a “Folksonomy tags suggestions” for the user on a dataset built on chosen objects [1–4]. The term “folksonomy” indicates the emergent labeling of lots of things by people in a social context, typically flat name-spaces. Folksonomy consists of disconnected and loosely related keywords, which ideal typically exist in a coordinated horizontal universe, only connected by associative relations. The emergent data from the actions of millions of ordinary untrained folks can be used for describing and finding unexpected attributed of an object in a video. Towards “Attributes extraction” and “Sentiment analysis” (see previous section) we can derive the Folksonomy tags correlated to the considered object to obtain a Folksodriven Bounding Box Notation (7.4) for the object of interest in the video.

$$FB := (B, E, S, X) \tag{7.4}$$

For an object represented on a video we can define a Folksodrive

n tag as a tuple defined by finite sets composed by:

- Bounding Box (B) is a tuple $B := (P, O, A, I)$ defined above;
- Time Exposition (E) is the clickthrough rate (CTR) as the number of clicks on an Object (O) divided by the number of times that the Object (O) is displayed (impressions);
- Sentiment Attributes (S) is defined above;
- X is defined by the relation $X = B \times E \times S$ in a Minkowski vector space [22] delimited by the vectors B, E and S.

We can use these attribute classifiers (7.4) to make soft predictions, of which attributes apply to an object of interest of a frame video, and use those predictions as features for object classification as shown after in Advertisement inside the video of the Application Examples.

7.5 Image Analysis and Feature Selection

In this section we see through the implementation of the proposed method.

The user is provided with the option to click over an object of interest displayed on the single scenes. Upon so doing the video is stopped and over the chosen object of interest is displayed a choice of possible folksonomy tags to correlate

from the ImageNet video database (see Semantic Attributes) and a form to add a new folksonomy tag reach the semantic attributes (see Attribute extraction and Sentiment analysis).

While viewing the video, the user can click controls in the control panel to stop the video and go back or forward as in normal use of video players. The provider of the video can also measure the amount of the time the user watched the video and are able to control the beginning and the ending images displayed before and after the video plays.

Thereafter, the merged file with the Folksodriven Bounding Box Notation (7.4) is stored as a video format file on a computer system or network server and accessible over a network or on the Internet. The stored file should be in a form that reduces storage capacity, and is preferably compressed. The JavaScript code for the input form interaction resides in the same file as the object code for the video. The result is that the object is written and the video is queued up but it will not play if it is not used so as not to disrupt the video playing.

The ontological display requires that the code is written on a separate webpage in order for the control panel to work.

This is done using an inline frame, also known as `iframe`, to load in the new page in a hidden space above the video displayed. The `iframe` is placed into a layer with settings that enables the `iframe` to remain hidden from the user and not disrupt the fullscreen display of the video clip.

The interface was developed using CSS3, HTML5 and JavaScript with GWT and Backbone.js.

7.5.1 Object Position Detection

A JavaScript function detects the position of the mouse pointer while it is over the video. The position of the mouse pointer is detected from its top and left position, which is its position relative to the top/left point of the screen, which is a signed coordinates 0, 0.

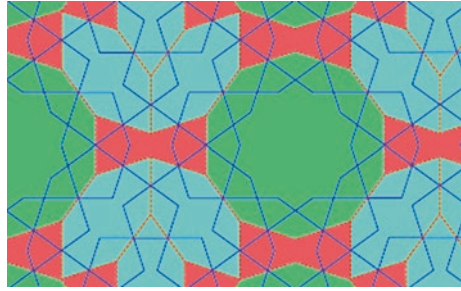
This information corresponds to the position object of interest while creating the form panel graphic over the video (see Fig. 7.2).

JavaScript is used to detect the position of the mouse pointer once the user clicks the button. This is done with the following code:

```
<SCRIPT LANGUAGE="JScrip" for="km_player"
event="MouseCown(nButton, nShiftState, fX, fY)">
```

As a further feature, a tracking program monitors the users and tracks the video that has been played by the user. For instance, the tracking may track the meta description of the video and the number of seconds or in milliseconds of the video that users watch. The tracking information is sent via http to a tracking server. This information can provide feedback for providers of the present method (see A case study: In-Video Advertisement).

Fig. 7.4 Penrose “aperiodic” tiling patterns



7.6 Previsions on Ontology Structure

FSNs (Folksodriven Structure Network) have regular elements, like normal networks. But these elements fit together in ways, which never properly repeat themselves [1]. The two-dimensional equivalent is known as Penrose tiling (Fig. 7.4), after Sir Roger Penrose, a British mathematician who put this form of geometry on a formal footing.

Penrose created “aperiodic” tiling patterns (Fig. 7.4) that never repeated themselves; work that he suspects was inspired by Kepler’s drawings. A three-dimensional Penrose tile equivalent is used in chemistry for the quasicrystals [27–29] having revolutionized materials science.

The paper [1] discuss the deformation exhibiting linear behavior on FSN based on folksonomy tags chosen by different user on web site resources, this is a topic which has not been well studied so far. The discussion shows that the linear elastic constitutive equation possesses some leaning for the investigation. A constitutive law on FSN [1–3] is investigated towards a systematic mathematical analysis on stress analysis and equations of motion for an evolving ontology matching on an environment defined by the users’ folksonomy choice. The adaptive ontology matching and the elastodynamics are merged to obtain, what we can call the “elasto-adaptive-dynamics methodology” of the FSN.

FSN deals mostly with how something evolves over time considering n-dimensions, virtually infinite dimensions, depending on the network connection.

Space or distance can take the place of time in many instances. FSN happens only in deterministic, nonlinear, and dynamical systems. For the conventional mathematical models the nonlinear behavior means mainly plasticity. In the study on the classical plasticity, there are two different theories, one is the macroscopic plasticity theory and the other is based on the mechanism of motion of dislocation.

The so-called FSN plasticity theory developed in [1] is based on classical theory [30] considering the mechanism of motion of dislocation, and in some extent can be seen as a “microscopic” theory.

The variations on the lattice FSN using parameters defined on the FD structure was observed [1]. The plasticity was considered as the capacity of FSN lattice to vary in developmental pattern, by the interaction of the Folksodriven Bounding

Box Notation (FB) and the web source environment, or in user behavior according to varying user choice of the FB tags.

The important connection between plasticity and the structural defects in FSN was observed in experiments, where plastic deformation of a Folksodiven Structure Network (FSN) was induced by motion of dislocations in the network. Salient structural features are presented in the FSN.

An example of experimental data for stress-strain curves for FSN news websites is depicted in [1], it presents plastic properties.

7.7 A Case Study: In-Video Advertisement

Nowadays almost 60 % of advertisements on YouTube videos are “skippable ads” that pay for only when people watched them (<http://gigaom.com/2012/01/19/more-than-half-of-all-youtube-video-ads-are-skippable>). That was a radical change from the pay-per-click model of Google for advertisements and from the more conventional model of paying for “impressions”—that doesn’t take into account if advertisements are actually viewed. The criticism on using “skippable ads” was that people might not watch the ads at all. But YouTube has found that viewership is 40 % higher with skippable inside videos while only 10 % of viewers always skip ads. As a result, video advertisements revenue per hour for video viewed on YouTube results higher respect cable TV [31]. With skippable-ads users were encouraged to engage in desired behaviors by letting them a path to mastery and autonomy to increase the value for those advertisements that people chose to watch. Skippable-ads represent an application of a “gamification” techniques: taking advantage of humans’ psychological predisposition to engage in gaming to encourage people to watch advertisement [32].

The “Layered Ontological Image Interaction method” proposed in this paper finds particular use for advertising feature to the user using “gamification” techniques. It is particularly well suited for providing advertisements correlated to different objects displayed in video and can provide an incentive to create better advertising to increase their value.

For the ontological objects in video defined with the FD tags and according to the plasticity we have a change on the ontology matching correlating the related ontologies relative to an object defined in videos with the defined Folksodiven Bounding Box Notation (FB)

According to the “elasto-adaptive-dynamics methodology” of the FSN defined in [1, 2] the system can chose the relative FB tags that is correlated to the classified in-video ads.

In this way the in-video ads are targeted not only on the keywords chosen for it but even for the interests of the users that are determined by the ontology matching on the FB created by the users’ folksonomy “driven” by the system—that constitute the FB systems.

7.7.1 *In-Video Advertisement Functionality*

Our system allows, through the identification of individual objects in a video can be used to create a new advertisement format (that we call In-Video Ads) from the extension of the In-Text Ads philosophy.

The results of the In-Text Ads, and the contextual advertising, are very interesting (see: <http://www.dblp.org/search/#query=contextual%20advertising>) and should be preserved, if not increased in the In-Video Ads due to a greater emotion generated by multimedia.

In-Video Ads can have a major impact on expand awareness for a brand associated on an object displayed in a video (see Fig. 7.5) increasing the consumer's recall and awareness of the primary message for the campaign.

With contextual advertising the ads are more targeted, it is most likely to be clicked, thus generating revenue to the website owner. On the other hand, it allows business to expand awareness for their brand. Although it poses issues like third-party hyperlinking, it could still be a potential way of promoting a product or service on the Internet.

The In-Video Ads, thought as a contextual advertising system, scans a video content for the Folksodrive associated, and then it will return advertisements based on the video attributes (see Richer Semantics of attributes).

This form of contextual advertising can work using the mouseover on the objects comparing the recognition of the objects proposed in this work with the tags connected or the keywords related to the product to advertise. So it will be possible to show the advertising message as a tooltip/layer on the video itself, or on one side of the space dedicated to video.

Context advertisement service providers will provide with JavaScript code inserted to web pages, and then it will display relevant ads. A script included in the code can provide the relation of ad copies with the attributes of the video content, so that it will show a relevant ad for the user.

The In-Video Ads can be displayed over a video as an over layer iframe, a separate ad unit on the page, as a pop-up, or fully on the screen as opposed to inside the video.

The objects displayed on the video (like a car in Fig. 7.5) works like a button and when clicked will call a JavaScript program that will display the relative advertisement. The advertisement keeps visualizing, from where it was, when the button was clicked. The result is an in-video page advertisement in the Internet (that we call In-Video Ads), which allows the end user, at their discretion to select a full screen over the video clip. Once in the full screen mode, this present method enables the user to seamlessly close out of the full screen mode.

7.7.2 *Web GRP*

We consider here how to correlate the scalar advertisement metrics to the vector representation of the Folksodrive Bounding Box Notation (FB). To perform that



Fig. 7.5 In-video advertisement of the car selected by the user on the video, the ad can be personalized according to the object of interest

we define before a Web Gross Rating Point and the Target Rating Points (TRP) to compare different media more easily (i.e.: web, mobile, tablet, ...).

In a highly competitive market it is crucial to create online tools for different campaigns. For each of those different specific measurement, systems are customized in line with the brand’s multimedia strategy to be used in one or more campaigns. However, these indicators cannot be standardized and compared across the market. An initial but by no means exhaustive list of measurement tools available to advertisers is presented in [33, 34], which gives an idea of the range available. The same indicator can often be suitable for measuring different objectives.

Moreover, the online strategy is not considered in isolation, but in terms of its contribution to the overall marketing and sales objectives. The Web’s potential as a branding medium depends on the ability to compare different media in terms of coverage and repetition.

In order to compare different media more easily, for the past few years media agencies have been referring to the Web Gross Rating Point (Web GRP). This is when the Gross Rating Point (GRP) is applied to the internet, and is defined as follows:

$$\text{Web GRP} = \text{Reach} \times \text{Frequency} \tag{7.5}$$

- Reach is the percentage of a target audience reached by an advertisement.
- Frequency is the average amount of exposures, so the message’s average repetition per targeted individual.

From a general Daily GRP’s (7.6), we can derive the Web Daily GRP’s as in (7.7) considering the Bounding Box (B) as campaign tool and the total web users as total market population

$$\text{Daily GRP's} = \left(\frac{\# \text{ of campaign} \times \text{campaign CTR}}{\text{total market population}} \right) \times 100 \tag{7.6}$$

$$\text{Web Daily GRP's} = \left(\frac{\#B \times \text{CTR}}{\text{total web users}} \right) \times 100 \quad (7.7)$$

From the measuring exposure to advertising [29], we can obtain the Frequency and the Reach respect the Web Daily GRP's

$$\text{Frequency} = \left(\frac{\text{Target Audience Weights*} \times \text{Daily GRP's} \times \text{days in campaign}}{100} \right) \quad (7.8)$$

$$\text{Reach} = \frac{(\text{Target Audience Weights*} \times \text{Daily GRP's} \times \text{days in campaign})}{\text{Frequency}} \quad (7.9)$$

*Target Audience Weight for Adults 18+ is 1.00

To combine the GRP's, Reach and Frequency for different media (i.e.: Web and Mobile as formats X and Y), we assume a constant Reach across a display period according to the Gallop math models—an industry accepted arithmetic formula that has been in use for over 40 years.

$$\text{Combined GRP's} = X \text{ GRP} + Y \text{ GRP} \quad (7.10)$$

$$\text{Combined Reach} = (X \text{ Reach} + Y \text{ reach}) - (0.01 \times X \text{ reach} \times Y \text{ Reach}) \quad (7.11)$$

$$\text{Combined Frequency} = \frac{\text{Combined GRP's}}{\text{Combined Reach}} \quad (7.12)$$

Target Rating Points (TRP) is the criterion that indicates the popularity of a video. Target Rating Points (TRP) figures the sum of the audiences of all the devices insertions of each individual Web GRP multiplied by the estimated target audience in the gross audience.

$$\text{Web Daily TRP} = \sum \text{Web Daily GRP} \times \text{Target Audience} \quad (7.13)$$

7.7.3 *Folksodriven Ontology Prediction for Advertisement*

The limitation in using this indicator on the Web essentially concerns the quality of the method used to gauge the Web audience, attributable to the large number of publishers (for example, compared with the number of TV channels). Some of the advertisers use the Web GRP, when preparing their media strategy, often alongside a “recognition beta”, which measures the recognition rate for different advertising media for a given target audience. This basis is then used to determine the memorized coverage. Measuring the effectiveness of online advertising is a key issue in managing campaigns and justifying the media. Selection, which types of media to use, largely depends on coverage indicators and planning tools.

For determining the marketing effectiveness of a Folksodriven Bounding Box Notation (FB) for an advertisement on an object of interest, depicted in (7.2), we can correlate that to the Web GRP and Web TRP metrics.

As example of application, we consider that correlation only on mobile devices, avoiding to calculate the combined values. In the case of a video advertisement that in one day is scheduled on a Folksodriven Bounding Box (FB) of a “Ferrari” with a Frequency of 10 times with a Reach of 50 % of the gross audience—as the audience for all the devices—we have a Web GRPs of 500 ($= 10 \times 50$). Considering a Target Audience of 60 % out of the Web GRP of 500, we obtain a TRPs of 300 ($= 500 \times 0.6$).

The Folksodriven Prediction (FP) tuple for a “Ferrari advertisement campaign” is defined as in (12): $\text{FerrariCampaign} := (\text{Ferrari}, 500, 300)$

Towards the Folksodriven Bounding Box Notation (FB)—defined in a Minkowski vector space as in (7.2)—for the “Ferrari” FB we can deduce the correlation among the WG and WT predictions with the clickthrough rate (CTR)—expressed by the Time Exposition (E)—and the users Sentiment Attributes (S).

7.7.4 In-Video Advertisement Validation

The validation for the In-Video Advertisement was done respect a defined reference based on skippable video advertisements that is delivered on defined tagged video. We considered an observation on 200 different videos during six months. The videos were classified with different tags correlated to the subject on them, according to those classified tags the correlated skippable video advertisements has been delivered. The same skippable video advertisements were correlated even to the FB tags defining the object on the videos. Using the same videos and advertisements we could compare the static system to correlate advertisements in video with the dynamic system defined by the Folksodriven Bounding Box Notation (FB tags).

An In-Video Advertisement shown their visual impact and the targeting on object in video with a dynamic correlation among the ontologies defined by the FB tags.

Figure 7.6 is compared the tendency on time (six months) of the impact factor of the video ads (CTRs/Impressions) for the generic “predefined tagged video” and for the “FB tags on objects in a video”—defined by the dynamics FB tags correlations.

We can see how on evolving time the CTRs/Impressions increase on time having targeting the advertisements with the objects in a video according to the “definition” used by the users on those.

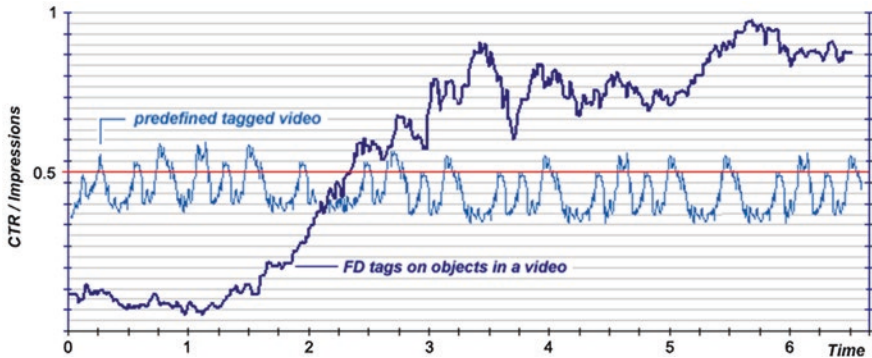


Fig. 7.6 Comparison among in-video advertisement and skippable video on 200 videos during 6 months, measuring the CTR respect the impressions of the advertisement

7.8 Relevant Resources

The task of “Object Recognition”—recognize and analyze items in photos or video sequence—is still a challenge for computer vision systems.

Many approaches have been developed to the challenges of Object Recognition such as algorithms from image processing, pattern recognition, computer vision and machine learning (see Background of related work and [10, 18] for introductions to those algorithms).

The approach described in this work combines multiple approaches to the challenges of Object Recognition—such as algorithms from image processing, pattern recognition, computer vision and machine learning [11–21]—with a layered representation [19] based on semantic texton forests [22] to obtain spatiotemporal object attributes based on Folksonomy tags extracted from users’ comments to be used in a dynamical driven system (Folksodriven) [1–6].

An updated place for all sorts of information about the method presented is the Web site: <http://www.folksodriven.com/multimedia-systems-in-practice>

It is possible to see a video describing the features on the system proposed and the video recording of the presentation at the 3rd International Conference on Web Intelligence, Mining and Semantics (WIMS’13).

7.9 Conclusion

We had in this work an overview of the field of the object detection on a video from the perspective of human computer interaction to build proactive engagement on users.

We explored a different system approach to enrich user interactions on video contexts towards ubiquitous Human Computer Interaction (HCI) and Human

Based Computation (HBC) on systems as computer monitor, mobile, or tablet. The proposed “Layered Ontological Image Interaction” can be motivated on the value of data obtained filling the gap between words describing an object in a video and its visual features. The contribution on working with detected object is to find unexpected attributes for an object of interest learning new categories by sharing higher-level traits. We briefly examined the benefit of this approach for the interactive systems for the case study of the “In-Video” Advertisement respect a skippable advertisement.

References

1. Dal Mas, M.: Elastic adaptive dynamics methodology on ontology matching on evolving folksonomy driven environment. *Evol. Syst. J.* **5**(1), 33–48 (<http://dx.doi.org/10.1007/s12530-013-9086-5>) (2014)
2. Dal Mas, M.: Folksodriven structure network. In: *Ontology Matching Workshop (OM-2011) Collocated with the 10th International Semantic Web Conference (ISWC-2011)*, CEUR WS, vol. 814. (<http://ceur-ws.org/Vol-814>) (2011)
3. Dal Mas, M.: Elastic adaptive ontology matching on evolving folksonomy driven environment. In: *Proceedings of IEEE Conference on Evolving and Adaptive Intelligent System (EAIS 2012)*, Madrid, Spain, 35–40. IEEE, New York. (<http://dx.doi.org/10.1109/AIS.2012.6232801>) (2012)
4. Dal Mas, M.: Intelligent interface architectures for folksonomy driven structure network. In: *Proceedings of the 5th International Workshop on Intelligent Interfaces for Human-Computer Interaction (CISIS-2012)*, Palermo, Italy, 519–525. IEEE, New York, (<http://dx.doi.org/10.1109/CISIS.2012.158>) (2012)
5. Dal Mas, M.: Elasticity on ontology matching of folksodriven structure network. Accepted for the 4th Asian Conference on Intelligent Information and Database Systems (ACIIDS 2012)—Kaohsiung Taiwan R.O.C., CORR—Arxiv, US (<http://arxiv.org/abs/1201.3900>) (2012)
6. Dal Mas, M.: Cluster analysis for a scale-free folksodriven structure network. Accepted for the International Conference on Social Computing and its Applications (SCA 2011), CORR—Arxiv, US (<http://arxiv.org/abs/1112.4456>) (2011)
7. Sebe, N.: *Human-centered Computing Handbook of Ambient Intelligence and Smart Environments 2010*, Part IV, 349–370 (2010)
8. Nov, O.: What motivates Wikipedians? *Commun. ACM* **50**(11), 60–64 (2007)
9. Ryan, R.M., Deci, E.L.: Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemp. Educ. Psychol.* **25**:54–67 (2000)
10. Zhao, M., Yagnik, J.: Automatic large scale video object recognition United States Patent. 8,254,699 (2012)
11. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: *CVPR 2006* (2006)
12. Russakovsky, O., Lin, Y., Yu, Fei-Fei, K.L.: Object-centric spatial pooling for image classification. In: *European Conference on Computer Vision (ECCV 2012)* (2012)
13. Yilmaz, A., Javed, O., Shah, M.: Object tracking:survey. *ACM Comput. Surv.*, **38** (2006)
14. Chen, T.P., Haussecker, H., Bovyrin, A., Belenov, R., Rodyushkin, K., Kuranov, A., Eruhimov, V.: Computer vision workload analysis: case study of video surveillance systems. *Intel Technol. J.* **9**(2), 109–118 (2005)
15. Comaniciu, D., Ramesh, V., Meer, P.: Real-time tracking of non-rigid objects using mean shift. In: *Proceedings IEEE CVPR 2000*, vol. 2 (2000)

16. Seeger, M.: Learning with labeled and unlabeled data. Technical report, University of Edinburgh (2001)
17. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**, 1254–1259. IEEE, New York (1998)
18. Forsyth, D., Ponce, J.: *Computer Vision: a Modern Approach*. Prentice-Hall, Englewood Cliffs, NJ (2002)
19. Tao, H.S.H., Kumar, R.: Object tracking with bayesian estimation of dynamic layer representations. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 75–89 (2002)
20. Xiao, J., Haysy, J., Ehinger, K.A., Oliva, A., Torralba, A.: SUN database: large-scale scene recognition from abbey to zoo. In: *IEEE Conference Computer Vision and Pattern Recognition (CVPR 2010)* (2010)
21. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost for image understanding: multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *Int. J. Comput. Vision Arch.* **81**(1), 2–23 (2009)
22. Criminisi, A., Shotton, J., Konukoglu, E.: Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Found. Trends® Comput. Graph. Vision* **7**(2–3), 81–227. NOW Publishers, Boston (2012)
23. Courant, R., Hilbert, D.: *Methods of Mathematical Physics, vol II*. Interscience. Wiley, New York (2008)
24. Nasukawa, T., Yi, J.: Sentiment analysis: capturing favorability using natural language processing In: *Proceedings of the 2nd International Conference on Knowledge Capture* (2003)
25. Mehrotra, S.: *Technology Review* September/October 2012, p 58 (2012)
26. Deterding, S., Dixon, D., Khaled, R., Nacke, L.: From game design elements to gamefulness: defining gamification. In: *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, pp 9–15 (2011)
27. Mussel, L.: *Structure of bio-polymer*. Publication of the Chemical Science Department—Padua University Press (1964)
28. Born, M., Huang, K.: *Dynamic Theory of Crystal Lattices*. Clarendon Press, Oxford (1954)
29. Calliard, D.: Dislocation mechanism and plasticity of quasicrystals: TEM observations in icosahedral Al-Pd-Mn. *Mater. Sci. Forum* **509**(1), 49–56 (2006)
30. Chakrabarty, J.: *Theory of Plasticity*. Butterworth-Heinemann, London (2009)
31. Chang, H.-S., Kinnucan, H.W.: Measuring exposure to advertising: a look at gross rating points. *Agribusiness* **8**, 413–423 (1992)
32. Nielsen online GRPs to launch with Facebook as partner (<http://www.research-live.com/news/analytics/nielsen-online-grps-to-launch-with-facebook-as-partner/4005762.article>)
33. Kotler, P.: *Marketing Management*. Prentice Hall, New Jersey (2000)
34. Farris, P.W., Bendle, N.T., Pfeifer, P.E., Reibstein, D.J.: *Marketing Metrics: The Definitive Guide to Measuring Marketing Performance*. Pearson Education, New Jersey (2010)