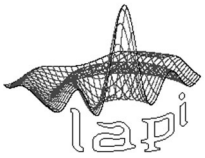


University Politehnica of Bucharest
Faculty of Electronics, Telecommunications and Information Technology

Department of Applied Electronics and Information Engineering
LAPI - The Image Processing and Analysis Laboratory

Multimedia Content Processing and Analysis for Large-Scale Information Retrieval

Habilitation thesis
PhD domain: Electronic Engineering and Telecommunications

Author : **Bogdan-Emanuel IONESCU**

,

Bucharest, 20 August 2014

Contents

Foreword

I present my habilitation thesis in view of obtaining the qualification for conducting PhD research at university level. It presents a summary of my major didactic and scientific career achievements accomplished after finalizing the doctoral studies. After more than 12 years of research, during which I have contributed to several areas of multidimensional signal processing involving image, video and multimedia processing and analysis, my personal objective is to be able to move a step forward and coordinate PhD research.

The ability of conducting PhD research plays a critical role in my strategy for the evolution and development of the professional career, as PhD research constitutes the basis research for any research institution as well as for future technological progress. Currently, I am involved with supervising of PhD research for several students at the University Politehnica of Bucharest as well as officially PhD co-advisor at the University of Trento, Italy. Obtaining the habilitation will allow me to take lead of PhD coordination and have the basis for a future research laboratory.

Before discussing in detail all these aspects, I would like to take advantage and acknowledge the significant contribution of my colleagues to all the achievements presented in this manuscript.

I would like to express my gratitude to all of my students for their hard work and constant implication in the developed projects. In particular, I thank Ionuţ Mironică, Anca-Livia Radu, Bogdan Boteanu, Cătălin Mitrea, Andrei Purică and Alexandru Marin.

I would like to thank my university colleagues, Prof. Corneliu Burileanu, Prof. Adrian Badea, Prof. Vasile Buzuloiu†, Prof. Dragoş Burileanu, Assoc. Prof. Mihai Ciuc, Prof. Constantin Vertan, for their constant support, precious advices, and guidance during my professional career.

An important role was played by all of my colleagues abroad who I warmly thank for their constant support and quality work developed together, in particular: Prof. Nicu Sebe, Prof. Fausto Giunchiglia, Prof. Patrick Lambert, Dr. Klaus Seyerlehner, Assoc. Prof. Markus Schedl, Dr. Martha Larson, Prof. Beatrice Pesquet-Popescu, Dr. Adrian Popescu, Prof. Jenny-Benois Pineau, Prof. Henning Müller, Dr. Claire-Hélène Demarty.

A special thanks to Monica, for her constant and unconditionally support and for accepting to share me with my passion for research during the last 9 years!

Abstract

This habilitation thesis presents an overview of my main research results obtained after receiving the PhD title in 2007. These results are presented in the context of my resume and professional, didactic and research career.

During my 12 years of research, I have mainly worked in the areas of multidimensional signal processing, i.e., *image*, *video* and *multimedia* processing and analysis. My contributions are grouped around three main topics, namely: *human-computer interaction*, *medical imaging* and *content-based indexing* (and information retrieval). My major research direction is however the work on content-based indexing and information retrieval. My research provided contributions to all the components of a content-based retrieval system as well as solutions to solve this paradigm in some particular application domains.

These contributions are in developing: *algorithms for video temporal segmentation* - temporal segmentation means decomposing the video into its temporal units or shots. This is commonly used as a pre-processing step for any other higher-level analysis. My major contributions were in the development of cut, fade and dissolve detection algorithms; *algorithms for video content description* - I have developed a number of techniques for representing color and action information in videos, contributed to a technique for capturing temporal variation as well as to multi-modal description and fusion (text-audio-visual); *algorithms for video summarization* - visualization of video data for video browsing requires efficient summarization tools. My contribution consisted in the development of algorithms for video summarization and video skimming; *algorithms for video genre tagging* - I have contributed to the development of a system that allows for the indexing of video materials according to video genres (e.g., movies, music, sports, etc). The research targeted the particular case of web specific video materials. I have also developed a technique for the automatic detection of the animated video genre; *algorithms for violent scenes detection* - I have contributed to the development of an indexing system for the automatic identification of violence scenes in typical Hollywood productions. Moreover, I co-coordinated an evaluation benchmarking campaign on this topic; *algorithms for search result diversification* - typical content-based retrieval focuses on the relevance of the results. However, the user is not always interested in obtaining identical replicas of the search queries. I have contributed to the development of several approaches for diversifying the search results in order to cover different topics of the query. I also initiated and organized an evaluation benchmarking campaign

on this topic; *algorithms for relevance feedback* - I have contributed to the development of several relevance feedback techniques that exploit user feedback to improve the relevance of image/video search results; and *multimedia browsing systems* - I have contributed to the development of several video and multimedia browsing interfaces that allows for a virtual representation of the datasets and interactive user experience.

My scientific results have been validated and disseminated with more than 90 scientific publications, such as book chapter contributions (e.g., Springer Lecture Notes in Computer Science, IGI Global, Springer Advances in Computer Vision and Pattern Recognition), international journal articles (e.g., SPIE Journal of Electronic Imaging - Impact Factor 1.1, Multimedia Tools and Applications - Impact Factor 1, EURASIP Journal on Image and Video Processing - Impact Factor 0.6, EURASIP Journal on Applied Signal Processing - Impact Factor 0.8, etc), international conference articles (e.g., ACM International Conference on Multimedia - MM, ACM International Conference on Multimedia Retrieval - ICMR, ACM International Conference on Multimedia Systems - MMSys, IEEE International Conference on Acoustic, Speech and Signal Processing - ICASSP, IEEE International Conference on Image Processing - ICIP, International Conference on MultiMedia Modeling - MMM, etc) as well as contributions to international benchmarking campaigns (e.g., MediaEval Benchmarking Initiative for Multimedia Evaluation).

My habilitation thesis is structured into two parts. Part 1 deals with my professional, didactic and research achievements. It presents my Curriculum Vitae (Chapter 1), a brief overview of my teaching activity (Chapter 2), my research activity in terms of scientific publications, attendance to international benchmarking, committees, conferences and international cooperation (Chapter 3), discusses the coordinated student projects (license, master as well as PhD and post-doctoral; Chapter 4), my implication with projects and strategic programmes (Chapter 5), and concludes by presenting in detail the most important scientific findings as enumerated in the previous section above (Chapter 6).

Part 2 of my thesis deals with the evolution and development of my professional career and discusses my perspectives for conducting fundamental and applicative research, for increasing the visibility of my research and publications' impact, for increasing the participation to international projects and technological transfer, as well as for teaching activities and coordinating student research and research groups.

Rezumat

Prezenta lucrare de abilitare realizează o trecere în revistă a rezultatelor de cercetare obţinute după finalizarea tezei de doctorat în anul 2007. Aceste rezultate sunt prezentate în contextul carierei profesionale, didactice şi de cercetare.

Pe parcursul celor 12 ani de cercetare, am lucrat în principal în domeniul prelucrării şi analizei de semnale multidimensionale, precum imagini, video şi multimedia. Contribuţiile aduse pot fi grupate în jurul a trei direcţii principale, şi anume: interacţie om-maşină, imagistică medicală şi respectiv căutarea datelor după conţinut (căutarea informaţiei). Totuşi, direcţia principală de cercetare o constituie domeniul tehnicilor de căutare a datelor după conţinut. Cercetarea realizată a adus contribuţii la toate componentele unui astfel de sistem cât şi soluţionarea acestei paradigme pentru anumite domenii particulare de aplicaţie.

Aceste contribuţii vizează dezvoltarea de *algoritmi pentru segmentarea temporală a secvenţelor video* - segmentarea temporală reprezintă descompunerea unei secvenţe video în părţile constituente de bază numite şi plane video. Această etapă de prelucrare este unul dintre paşii preliminarii unei analize de nivel înalt. Contribuţiile principale au constat în dezvoltarea de algoritmi pentru detecţia tranziţiilor de tip "cut", "fade" şi "dissolve"; *algoritmi pentru descrierea conţinutului video* - am dezvoltat o serie de tehnici de reprezentare a informaţiei de culoare şi acţiune, am contribuit la dezvoltarea unei metodologii de reprezentare a variabilităţii temporale în secvenţe video cât şi la tehnici de descriere multimodală şi fuziune a datelor; *algoritmi de rezumare automată a secvenţelor video* - vizualizarea datelor video necesită de regulă rezumarea acestora. Contribuţia principală a constat în dezvoltarea de algoritmi de rezumare în imagini cât şi dinamică; *algoritmi pentru clasificarea automată după gen* - am contribuit la dezvoltarea unui sistem de indexare ce permite regruparea automată a documentelor video în funcţie de gen (de exemplu în filme, muzică, sport, ştiri, etc). Cercetarea a vizat în particular platformele multimedia de pe Internet. De asemenea, tot în această direcţie, am dezvoltat un algoritm de detecţie a genului de animaţie; *algoritmi pentru detectarea pasajelor de violenţă în filme* - am contribuit la dezvoltarea unui sistem de indexare ce permite identificarea automată a scenelor de violenţă în producţii tipice de la Hollywood. Mai mult, am contribuit la realizarea unei campanii de evaluare a acestui gen de tehnici; *algoritmi de diversificare a rezultatelor căutării de imagini* - în general, utilizatorul nu este interesat doar de relevanţa rezultatelor ci şi de diversitatea acestora. Am contribuit la dezvoltarea mai multor abordări pentru diversificarea rezultatelor căutării

de imagini pe platformele sociale. De asemenea, am iniţiat şi coordonat o campanie de evaluare a unor astfel de tehnici; *algoritmi de tip "relevance feedback"* - am contribuit la dezvoltarea mai multor tehnici de relevance feedback ce se folosesc de expertiza utilizatorilor pentru a îmbunătăţii rezultatele căutării de date; *sisteme de navigare multimedia* - am contribuit la realizarea mai multor sisteme software de navigare în baze de date multimedia.

Rezultatele de cercetare obţinute până în prezent au fost validate şi diseminate în mai mult de 90 de publicaţii ştiinţifice precum: contribuţii la volume colective (de exemplu Springer Lecture Notes in Computer Science, IGI Global, Springer Advances in Computer Vision and Pattern Recognition), articole în reviste internaţionale (de exemplu SPIE Journal of Electronic Imaging - Factor de Impact 1.1, Multimedia Tools and Applications - Factor de Impact 1, EURASIP Journal on Image and Video Processing - Factor de Impact 0.6, EURASIP Journal on Applied Signal Processing - Factor de Impact 0.8, etc), articole la conferinţe internaţionale (de exemplu ACM International Conference on Multimedia - MM, ACM International Conference on Multimedia Retrieval - ICMR, ACM International Conference on Multimedia Systems - MMSys, IEEE International Conference on Acoustic, Speech and Signal Processing - ICASSP, IEEE International Conference on Image Processing - ICIP, International Conference on MultiMedia Modeling - MMM, etc), precum şi participări la campanii de evaluare internaţionale (de exemplu MediaEval Benchmarking Initiative for Multimedia Evaluation).

Teza de abilitare este structurată în două părţi. Partea întâi se adresează prezentării realizărilor profesionale, didactice şi de cercetare. Aceasta prezintă parcursul profesional (Capitolul 1), o trecere în revistă a activităţii didactice (Capitolul 2), rezumarea activităţii de cercetare: publicaţiile ştiinţifice, participarea la campanii de evaluare, participarea în comitete ştiinţifice, participarea la organizarea de evenimente ştiinţifice cât şi cooperarea internaţională (Capitolul 3); prezintă proiectele studenţeşti coordonate la nivel de licenţă, master cât şi studii doctorale şi post-doctorale (Capitolul 4), implicarea în proiecte de cercetare şi strategice (Capitolul 5), şi concluzionează prezentând în detaliu cele mai importante realizări ştiinţifice aşa cum au fost enumerate în paragraful anterior (Capitolul 6).

A doua parte a tezei se adresează prezentării evoluţiei şi dezvoltării carierei profesionale. Aceasta discută perspectivele personale în ceea ce priveşte cercetarea fundamentală şi aplicativă, creşterea vizibilităţii rezultatelor de cercetare şi a impactului publicaţiilor ştiinţifice, creşterea participării la proiecte internaţionale şi a transferului tehnologic cât şi perspectivele didactice şi de coordonare a cercetării.

Part I

Professional, didactic and research achievements

Chapter 1

Curriculum Vitae

This section presents a brief introduction of my resume: education, professional experience, participation in professional committees, experience abroad, training courses, as well as the main research and didactic areas of expertise.

1.1 Education

- 1997: **baccalaureate degree**, "Grigore Moisil" college - Bucharest, informatics-math-physics department (overall appreciation 9.4/10), analyst programmer title delivered by the Ministry of Education and Research of Romania;

- 2002: **engineer's degree in electronics**, University Politehnica of Bucharest, Faculty of Electronics, Telecommunications and Information Technology, Department of Applied Electronics, section: Computer Vision and Artificial Intelligence (overall appreciation 9.56/10), graduation project developed jointly with LAMII, ESIA - Ecole Supérieure d'Ingénieurs d'Annecy, France;

- 2003: **master's degree in computational systems**, University Politehnica of Bucharest, Faculty of Electronics, Telecommunications and Information Technology, Department of Applied Electronics, section: Computational Systems Technology (overall appreciation 9.93/10), graduation project developed jointly with LISTIC, Polytech'Savoie, Annecy-France;

- 2003: **master's degree derogation** (DEA - Diplôme d'Études Approfondies) from Université de Savoie, France;

- 2007: **PhD degree in electronic engineering and telecommunications**, University Politehnica of Bucharest, Faculty of Electronics, Telecommunications and Information Technology, "very well" jury appreciation and "magna cum laude" award;

- 2007: **PhD degree in electronics, electrotechnics and automatics**, Université de Savoie, Polytech'Savoie, Annecy-France;

- 2010-2013: **postdoctoral researcher** with LAPI - The Image Processing and Analysis Laboratory, Faculty of Electronics, Telecommunications and Information Technology, University Politehnica of Bucharest, scholarship under European Structural Funds, POS-DRU project EXCEL, ID62557 (ranked 3rd out of 64 candidates).

1.2 Professional experience

- 2005-2006: **substitute teacher ("vacataire")**, Université de Savoie, ESIA - Ecole Supérieure d'Ingénieurs d'Annecy, France, conducting signal processing seminars and laboratories;

- 2006-2007: **temporary assistant professor** (ATER - Attaché Temporaire d'Enseignement et de la Recherche), Université de Savoie, Polytech'Savoie, Annecy-France, CNU section 61: computer science and engineering, automatics and signal processing (ranked first in Haute-Savoie region);

- 2007-2009: **senior researcher**, LAPI - The Image Processing and Analysis Laboratory, Faculty of Electronics, Telecommunications and Information Technology, University Politehnica of Bucharest; and project director of the CNCSIS - Romanian National University Research Council, Human Resources Programme, RP-2 project;

- 2008-present: **courtesy faculty member** with Université de Savoie, France;

- 2008-2014: **lecturer** with Faculty of Electronics, Telecommunications and Information Technology, University Politehnica of Bucharest;

- 2012-present: **courtesy faculty member** with University of Trento, Italy;

- 2014-present: **associate professor** with Faculty of Electronics, Telecommunications and Information Technology, University Politehnica of Bucharest.

1.3 Professional committees

- 2014-present: **administrative council member** of the Romanian National Institute for Research and Development in Optoelectronics - INOE 2000;

- 2014-present: **technical program expert** for the Horizon 2020 Programme (Romanian commission).

1.4 Experience abroad

- April 2008: **invited researcher** at LISTIC laboratory, Polytech'Savoie, Université de Savoie, France;
- September 2008: **invited researcher** at VIVA laboratory, University of Ottawa, Canada;
- June 2009: **invited researcher** at LISTIC laboratory, Polytech'Savoie, Université de Savoie and at Institut Gaspard Monge, Marne-la-Vallée, Université Paris-Est, France;
- 15 June - 31 July 2011: **invited researcher** at LISTIC laboratory, Polytech Annecy-Chambery, Université de Savoie, France;
- June 2012: **invited researcher** at LISTIC laboratory, Polytech Annecy-Chambéry, Université de Savoie, France;
- July 2012: **invited researcher** at KnowDive laboratory, University of Trento, Italy.

1.5 Training courses

- 14 - 16 May 2012: **training course** on "Evaluating the Impact of Development Programs: Turning Promises into Evidence", Paul Gertler, University of California - Berkeley, Adam Ross, HDNCE, Joost de Laat, ECSH World Bank, Bucharest, Romania;
- 27 - 28 September 2012: **training course** on "Evidence Informed Policy and Practice in Education in Europe", Institute of Education, University of London, Institute of Educational Sciences, Bucharest, Romania.

1.6 Research and didactic expertise

- **research**: video/image processing and analysis, content-based retrieval, multimedia, computer vision, machine learning, human-computer interaction;
- **didactic**: video/image processing and analysis, multimedia indexing, computer vision, computer science and software engineering.

Chapter 2

Teaching activity

My didactic activity was carried out in parallel with the research, as the two components are strongly connected. I have permanently integrated valuable research achievements with the university curricula (e.g., courses, applications) to stimulate students being actively part of the research innovation. On the other hand, my research was always coordinated with strong fundamental theoretic knowledge from curricula (e.g., signal processing fundamentals).

My didactic activity started during the preparation of my PhD thesis and had the following route. Between 2005 and 2006 I was substitute teacher ("vacataire") with ESIA - Ecole Supérieure d'Ingénieurs d'Annecy, Université de Savoie - France. Then, between 2006 and 2007, I held a temporary assistant professor position (ATER - Attaché Temporaire d'Enseignement et de la Recherche) with Polytech'Savoie, Université de Savoie, Annecy-France. In 2008 I obtained a tenured lecturer position with the Department of Applied Electronics and Information Engineering (EAII), Faculty of Electronics, Telecommunications and Information Technology (ETTI), University Politehnica of Bucharest (UPB), where in 2014 I became associate professor, position I currently held.

I was/am in charge with conducting several courses, laboratories and seminars in various areas of signal processing and software engineering, as well as with coordinating license and master thesis and doctoral and post-doctoral research. During my career, I have prepared and taught the following courses and applications:

- (2005 - 2007) "**Traitement du signal**": I was in charge with signal processing seminars and laboratories at the department of Automatique et Informatique Industrielle - AII, ESIA, Université de Savoie (second year, second cycle). Topics include deterministic signals, spectral analysis, optimal prediction, filtering of random signals, digital filters and binary random signals;

- (2005 - 2007) "**Vision par ordinateur**": I was in charge with seminars and laboratories on computer vision at the department of Automatique et Informatique Industrielle - AII, ESIA, Université de Savoie (third year, second cycle). Topics include image segmentation, region segmentation, Hough transform, pattern recognition and color gradient;

- (2006 - 2007) "**Traitement d'image**": I was in charge with seminars and laboratories on image processing at the department of Physique Appliquée et Instrumentation - PAI, Polytech Savoie, Université de Savoie (third year, second cycle). Topics include image representation (sampling, unitary transforms), smoothing and contrast enhancement, gradient operators, mathematical morphology and classification of multi-spectral images;

- (2006 - 2007) "**Programmation orientée objet**": I was in charge with laboratories on object-oriented programming at the department of Génie Logiciel et Services - GLS, Polytech Savoie, Université de Savoie (second year, second cycle). I taught Java language with event management, graphical user interfaces and UML modeling (Unified Modeling Language);

- (2008 - 2009) "**Analyse et traitement d'image**": I was in charge with laboratories on image processing at the Faculty of Engineering in Foreign Languages (FILS), French language department, University Politehnica of Bucharest. Topics include punctual operations, neighborhood processing, integral processing and mathematical morphology;

- (2008-present) "**Programarea calculatoarelor**" (Computer Programming): I am in charge with the C language computer programming course at EAII, ETTI-UPB (first year, license cycle). In 2008 I have conceived the course and developed the support materials[1]. The course covers the introduction of computational systems and computer languages, fundamental C programming (operators, conditional and repetitive structures, structural data types), functions and recurrence, working with pointers and data files. I am also in charge with the laboratories whose topics follows closely the course;

- (2008-2010) "**Arhitecturi pentru semnale multidimensionale**" (Architectures for Multi-Dimensional Signal Processing): I was in charge with the course and laboratories for multi-dimensional signal processing, department EAII, ETTI-UPB, master programme on "Imagini, Forme şi Inteligenţă Artificială" - IFIA (Images, Pattern Recognition and Artificial Intelligence). Topics include image acquisition and image, video and audio coding systems;

- (2008-2013) "**Tehnici avansate de prelucrarea şi analiza imaginilor**" (Advanced Image Processing and Analysis): I was in charge with the image

[1]materials are available to students at the following link: `http://imag.pub.ro/~bionescu/index_files/Page328.htm`.

processing and analysis course, department EAII, ETTI-UPB, master programme on "Sisteme Inteligente şi Vedere Artificială" - SIVA (Intelligent Systems and Computer Vision). I have conceived the class support materials[2]. Topics include color representation systems, geometric transforms, punctual operations, linear and non-linear filtering, mathematical morphology and unitary transforms;

- (2009-present) "**Indexarea conţinutului vizual**" (Indexing of Visual Contents): I am in charge with the module on video indexing techniques, department EAII, ETTI-UPB, master programme on "Tehnici Avansate de Imagistică Digitală" - TAID (Advanced Digital Imagining). I have conceived the course materials [26, 9][2]. I also conduct the laboratories whose topics include video information, color analysis, video temporal structure and motion analysis;

- (2010-present) "**Prelucrarea şi analiza imaginilor color**" (Color Image Analysis and Processing): I am in charge with the module on video analysis and processing, department EAII, ETTI-UPB, master programme on "Tehnici Avansate de Imagistică Digitală" - TAID (Advanced Digital Imaging) and "Tehnologii Multimedia în Aplicaţii de Biometrie şi Securitatea Informaţiei" - BIOSINF (Multimedia Technologies for Applications in Biometrics and Information Security). I have conceived the course materials[2]. I am also in charge with the laboratories of this module;

- (2011-present) "**Interfaţare vizuală om-maşină**" (Human-Computer Interaction): I am in charge with the module on hand gesture recognition, department EAII, ETTI-UPB, master programme on "Tehnici Avansate de Imagistică Digitală" - TAID (Advanced Digital Imagining) and "Tehnologii Multimedia în Aplicaţii de Biometrie şi Securitatea Informaţiei" - BIOSINF (Multimedia Technologies for Applications in Biometrics and Information Security). I have conceived the course module and support materials. The topics include hand gesture modeling and computer vision techniques for gesture recognition;

- (2011-present) "**Computer vision. Analiza informaţiei vizuale**" (Computer Vision. Analysis of Visual Information): I am in charge with the module on geometric camera calibration, department EAII, ETTI-UPB, master programme on "Tehnici de Analiză, Modelare şi Simulare pentru Imagistică,

[2]materials are available to students at the following link: `http://imag.pub.ro/~bionescu/index_files/Page328.htm`.

Bioinformatică şi Sisteme Complexe" - ITEMS (Analysis, Modeling and Simulation for Imagining, Bioinformatics and Complex Systems). I have conceived the course module and didactic materials [10];

- (2011-present) "**Proiect de cercetare şi documentare în prelucrarea imaginilor**" (Research project on the review of the state-of-the-art in image processing): I am in charge with the semester project on image processing literature overview, department EAII, ETTI-UPB, master programme on "Tehnologii Multimedia în Aplicaţii de Biometrie şi Securitatea Informaţiei" - BIOSINF (Multimedia Technologies for Applications in Biometrics and Information Security). I have conceived the curricula whose topics include student familiarization with accessing research information, conceiving consistent research reports, conceiving and writing of scientific articles and familiarization with dedicated editing tools such as the Latex environment.

Chapter 3

Research activity

My research activity started early back in high school (1997) when I have developed for my accreditation project on informatics a software tool for medical image processing. It integrated a broad range of techniques for computer tomograph (CT) processing and analysis. The passion for research continued with my license degree thesis (2002) followed by my master degree thesis (2003), developed with Université de Savoie, France. After graduating, I have enrolled a joint doctoral programme between University Politehnica of Bucharest and Université de Savoie which I have successfully completed in 2007. Following my PhD, my research continued during a post doctoral position at the university. At present time, my research is conducted as part of my tenured associate professor position.

This section presents the impact of my research activity in terms of scientific publications, attendance to international benchmarking campaigns, attendance to international committees, organizing international conferences and scientific events and international cooperation. These are in my opinion highly important assets for the ability of leading competitive research in the field. A detailed description of my most relevant research work is presented in Chapter 6.

3.1 Scientific publications

The scientific research developed so far has been disseminated with more than 90 scientific publications (the most significant ones are presented in the bibliography section), which can be grouped around the following categories:

- **3 annotated datesets** were made publicly available to the community to support reproducible research (e.g., MediaEval Benchmarking Initiative for Multimedia Evaluation data: search result diversification dataset [1], violent scene detection dataset [2]; and multiple-instance based video surveillance retrieval dataset [3]);

- **3 edited publications**: a book volume (Springer Advances in Computer Vision and Pattern Recognition), a conference proceedings (MediaEval Mul-

Figure 3.1: Citations of my work according to Google Scholar (accessed on 12-08-2014).

timedia Benchmark Workshop, ACM Multimedia) and an upcoming special issue (Elsevier Image and Vision Computing - Impact Factor 2.3);

- **3 books** (e.g., "Editura Tehnică" and "MatrixRom" publishing houses);
- **8 book chapters** (e.g., Springer Lecture Notes in Computer Science, IGI Global, Springer Advances in Computer Vision and Pattern Recognition);
- **16 international journal articles** (e.g., SPIE Journal of Electronic Imaging - Impact Factor 1.1, Multimedia Tools and Applications - Impact Factor 1, Eurasip Journal on Image and Video Processing - Impact Factor 0.6, Eurasip Journal on Applied Signal Processing - Impact Factor 0.8, Academy Publisher Journal of Multimedia, etc);
- **60 international conference articles** (e.g., ACM International Conference on Multimedia - MM, ACM International Conference on Multimedia Retrieval - ICMR, ACM International Conference on Multimedia Systems - MMSys, IEEE International Conference on Acoustic, Speech and Signal Processing - ICASSP, IEEE International Conference on Image Processing - ICIP, International Conference on MultiMedia Modeling - MMM, International Conference on Information Fusion - Fusion, International Workshop on Content-Based Multimedia Indexing - CBMI, International Workshop on Adaptive Multimedia Retrieval - AMR, IEEE International Conference on Intelligent Computer Communication and Processing - ICCP, IEEE International Symposium on Signals, Circuits and Systems - ISSCS, etc);

A quantification of my research impact in the community is the number of citations, which according to Google Scholar[1] is above 340, whereas the h-index[2] is 9 (last accessed on 12-08-2014). Some statistics are presented in Figure 3.1.

[1] `http://scholar.google.ro/citations?user=11cvotAAAAAJ&hl=en`.
[2] `http://en.wikipedia.org/wiki/H-index`.

3.2 International benchmarking

Some of the developed algorithms have been tested in several international benchmarking campaigns to position them with respect to existing state-of-the-art. I have attended to the following campaigns:

- 2006: **ARGOS** Evaluation Campaign for Surveillance Tools of Video Content (algorithms for video temporal segmentation);
- 2008: **Trecvid** - BBC Rushes Summarization Campaign (algorithms for automatic video summarization);
- 2011: **MediaEval** Benchmarking Initiative for Multimedia Evaluation - Genre Tagging Task (algorithms for automatic video genre tagging of Web videos);
- 2012: **MediaEval** Benchmarking Initiative for Multimedia Evaluation: Genre Tagging Task (algorithms for automatic video genre tagging of Web videos - best run ranked 2nd out of 29 runs, best system non-organizer connected); Affect Task: Violent Scenes Detection (algorithms for automatic violence detection - best system, best run ranked 1st out of 35 runs); and Spoken Web Search Task (algorithms for automatic speech retrieval);
- 2013: **MediaEval** Benchmarking Initiative for Multimedia Evaluation: Affect Task: Violent Scenes Detection (algorithms for automatic violence detection); Retrieving Diverse Social Images Task (algorithms for search result diversification); and Spoken Web Search Task (algorithms for automatic speech retrieval);
- 2014: **MediaEval** Benchmarking Initiative for Multimedia Evaluation: Affect in Multimedia: Violent Scenes Detection (algorithms for automatic violence detection) and Retrieving Diverse Social Images Task (algorithms for search result diversification).

3.3 International committees

The recognition of my scientific activity allowed me to join some of the major scientific communities in the field, namely:

- **associate editor** for the Academy Publisher Journal of Multimedia;
- **reviewer** for:

- Scientific Bulletin of University Politehnica of Bucharest (SCOPUS);
- SPIE Journal of Electronic Imaging (ISI, Impact Factor 1.1);
- IS&T Journal of Imaging Science and Technology (ISI);
- IEEE Transactions on Image Processing (ISI, Impact Factor 3.2);
- IEEE Transactions on Circuits and Systems for Video Technology (ISI, Impact Factor 1.8);
- IEEE Transactions on Multimedia (ISI, Impact Factor 1.8);
- Multimedia Tools and Applications (ISI, Impact Factor 1);
- IEEE Signal Processing Letters (ISI, Impact Factor 1.7);
- IET Computer Vision (ISI, Impact Factor 0.6);
- Journal on Evolving Systems (SCOPUS);
- SPIE Optical Engineering (ISI, Impact Factor 0.9).

• **technical program committee** for:

- European Signal Processing Conference - EUSIPCO 2012, 2013 and 2014;
- International Workshop on Content-Based Multimedia Indexing - CBMI 2012, 2013 and 2014;
- ACM International Conference on Multimedia - MM 2012 and 2014;
- Workshop on Information Fusion in Computer Vision for Concept Recognition, European Conference on Computer Vision - ECCV 2012;
- European Workshop on Visual Information Processing - EUVIP 2013 and 2014;
- MediaEval Benchmarking Initiative for Multimedia Evaluation 2013 and 2014;
- IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP 2014;
- International Workshop on Social Media Retrieval and Analysis, ACM Special Interest Group on Information Retrieval - SIGIR 2014;
- IEEE International Conference on Image Processing, Theory, Tools and Applications - IPTA 2014;
- Pacific-Rim Conference on Multimedia - PCM 2014.

3.4 International conferences and events

During my career, I have been invited to be involved with organizing of several major conferences and events in the field, namely:

- 2010: **co-chair and organizer** of the workshop "Electronics, Telecommunications and Information Technology in the World and in Romania", Diaspora Conference on Higher Education and Scientific Research, University Politehnica of Bucharest;
- 2009-2013: **chair** for doctoral symposiums, PhD POS-DRU funded programmes, University Politehnica of Bucharest;
- 2012: **local arrangements chair** for the 20th European Signal Processing Conference - EUSIPCO, Bucharest-Romania;
- 2012: **chair** for "Biomedical Image Analysis" and "Remote Sensing and Geographic Signal Processing" sessions at EUSIPCO, Bucharest-Romania;
- 2012: **publication co-chair** for the 10th International Workshop on Content-Based Multimedia Indexing - CBMI, Annecy-France;
- 2012: **co-organizer and co-chair** of the special session "Automatic Indexing of Internet Multimedia" at CBMI, Annecy-France;
- 2012: **co-organizer** of the workshop "Information Fusion in Computer Vision for Concept Recognition" at the European Conference on Computer Vision - ECCV, Firenze-Italy;
- 2013: **proceedings co-chair** for ACM International Conference on Multimedia - MM, Barcelona-Spain;
- 2013-2014: **co-organizer** of the Affect Task: Violent Scenes Detection @ MediaEval, Barcelona-Spain;
- 2013-2014: **lead organizer** of the Diversity in Social Photo Search Task @ MediaEval, Barcelona-Spain;
- 2013: **co-organizer** of the Workshop on Event-based Media Integration and Processing, co-located with ACM Multimedia, Barcelona-Spain;
- 2013: **chair** for "Image Processing II" session at IEEE ISSCS, Iaşi-Romania;

- 2013: **awards and sponsorship co-chair** for MediaEval 2013 Multimedia Benchmark Workshop, co-located with ACM Multimedia, Barcelona-Spain;
- 2013: **coordinator** of the panel session on "Future trends in events for media" at the Workshop on Event-based Media Integration and Processing, co-located with ACM Multimedia, Barcelona-Spain;
- 2014: **local arrangements co-chair** for IEEE International Conference on Image Processing - ICIP, Paris-France;
- 2014: **co-organizer** of the Workshop on Human-Centered Event Understanding from Multimedia, ACM Multimedia, Orlando-USA.

3.5 International cooperation

My research activity so far allowed me to establish a vast network of collaborators around the world, e.g., Université de Savoie - France (Prof. Patrick Lambert), Télécom ParisTech - France (Prof. Beatrice Pesquet-Popescu), Université de Bordeaux - France (Prof. Jenny-Benois Pineau), LIG Grenoble - France (Research director at CNRS George Quénot), TU Delft - The Netherlands (Senior researcher Martha Larson), Queen Mary University of London - UK (Senior researcher Tomas Piatrik), Dublin City University - Ireland (Prof. Gareth Jones), University of Trento - Italy (Prof. Nicu Sebe, Prof. Fausto Giunchiglia, Prof. Francesco De Natale), University of Applied Sciences Western Switzerland (Prof. Henning Müller), Johannes Kepler University - Austria (Assoc. Prof. Markus Schedl), Technicolor France (Senior researcher Claire-Hélène Demarty), University of Texas at San Antonio - USA (Prof. Qi Tian), University of Edinburgh - UK (Assoc. Prof. Subramanian Ramamoorthy), University of Houston - USA (Prof. Ioannis Kakadiaris), CERTH Greece (Senior researcher Yiannis Kompatsiaris), Kiel University - Germany (Prof. Ansgar Scherp), etc.

This is in particular important for keeping a permanent contact with the international community and accessing a competitive research environment (e.g., for students).

Chapter 4

Student coordination

In what concerns my ability of coordinating didactic and research activities, I have successfully coordinated/currently coordinating more than 3 license degree thesis, 17 master degree thesis, PhD research for 7 students as well as a postdoctoral researcher. 17 of these thesis were developed jointly with universities abroad, e.g., Université de Savoie, Université Paris-Est, University of Trento, Télécom ParisTech, the students benefiting from local scholarhips as well as Socrates/ERASMUS mobility grants (I was in charge for 6 Socrates/ERASMUS mobility grants). A detailed list of the developed projects is presented in the following sections.

4.1 License degree thesis

- (2007-2008) **Alexandra Păcureanu**: "Symbolic Description of the Action Content in Animated Movies", developed jointly with LISTIC, Polytech'Savoie, Annecy-France (Prof. Patrick Lambert);
- (2008-2009) **Alexandru Marin**: "Investigations on Applying Stein's Principle for SENSE pMRI Reconstruction in the Wavelet Domain", developed jointly with LI, University Paris-Est, "Gaspard Monge" Institute, France (Prof. Jean-Christophe Pesquet, C.R. CNRS dr. eng. Caroline Chaux);
- (2008-2009) **Cosmin Ion-Petrescu**: "Motion Estimation Techniques in Video Indexing", LAPI, University Politehnica of Bucharest, Romania.

4.2 Master degree thesis

- (2009-2010) **Diana Grama**: "People Detection using Contour Information in Video Surveillance Videos", LAPI, University Politehnica of Bucharest, Romania;
- (2010-2011) **Vlad Dima**: "A 3D System for Navigating through Multimedia Databases", LAPI, University Politehnica of Bucharest, Romania;

- (2010-2011) **George Zaharia**: "Video Quality Assessment", LAPI, University Politehnica of Bucharest, Romania;

- (2011-2012) **Bogdan Alecu**: "Robust Background Substraction", developed jointly with LISTIC, Polytech Annecy-Chambery, Annecy-France (Prof. Patrick Lambert, Dr. Michel Cintract) and in cooperation with Eboo Solutions (private company developing video surveillance solutions). Socrates-ERASMUS mobility;

- (2011-2012) **Iulia Cazan**: "Video Genre Classification", developed jointly with LISTIC, Polytech Annecy-Chambery, Annecy-France (Prof. Patrick Lambert). Socrates-ERASMUS mobility;

- (2011-2012) **Anca-Livia Radu**: "Crawling Media for Tagging Locations on Large Scale", developed jointly with Department of Information Engineering and Computer Science, University of Trento, Italy (SR Julian Stöttinger, Prof. Fausto Giunchiglia);

- (2012-2013) **Calotă Anamaria - Mihaela**: "Automatic Multimedia Categorization", LAPI, University Politehnica of Bucharest, Romania;

- (2012-2013) **Irina-Emilia Nicolae**: "Versatile Layered Video Coding based on Scalable Distributed Video Coding", developed jointly with the Multimedia Group, Télécom ParisTech, France (Prof. Beatrice Pesquet-Popescu);

- (2012-2013) **Corina Macovei**: "Multilateral Layered Video Coding Based on Distributed Video Coding and Scalable Distributed Video Coding", developed jointly with the Multimedia Group, Télécom ParisTech, France (Prof. Frederic Dufaux);

- (2012-2013) **Andrei Purică**: "View Synthesis Techniques in Multiview Video plus Depth Coding", developed jointly with the Multimedia Group, Télécom ParisTech, France (Prof. Beatrice Pesquet-Popescu);

- (2012-2013) **Ionuţ Duţă**: "Detection and Localization of Objects in Images", developed jointly with the MHUG group, University of Trento, Italy (SR Jasper Uijlings, Prof. Nicu Sebe);

- (2013-2014) **Bogdan Boteanu**: "Enforcing Diversity in Photo Retrieval", LAPI, University Politehnica of Bucharest, Romania;

- (2013-2014) **Andrei Filip**: "Large-Scale Concept Detection in Videos", developed jointly with LISTIC, Polytech'Savoie, Annecy-France (Prof. Patrick Lambert). Socrates-ERASMUS mobility;
- (2013-2014) **Mădălina Tiţă**: "Mining Events from Social Media using Visual Content, Contextual Information and Event Saliency Models", developed jointly with the MultiMedia Signal Processing and Understanding Lab, University of Trento, Italy (Prof. Francesco de Natale, Assoc.Prof. Giulia Boato);
- (2013-2014) **Mihai Puşcaş**: "Logo/Commercial Detection", developed jointly with the MHUG group, University of Trento, Italy (Prof. Nicu Sebe);
- (2013-2014) **Vlad Ruxandu**: "A Study on Holographic Compression Techniques", developed jointly with the Multimedia Group, Télécom ParisTech, France (Prof. Beatrice Pesquet-Popescu). Socrates-ERASMUS mobility;
- (2013-2014) **Ioan Chera**: "Multimedia Browsing Interfaces", LAPI, University Politehnica of Bucharest, Romania.

4.3 Doctoral research

- (2012 - 2013) **Ionuţ Mironică**: "Fisher Kernel Paradigm in the Context of Image Retrieval", research stage at the Department of Information Engineering and Computer Science, University of Trento, Italy (SR Jasper Uijlings, Prof. Nicu Sebe). Socrates-ERASMUS mobility. PhD supervisor Prof. Radu Dogaru;
- (2012 - 2015) **Cătălin Mitrea**: "Content-based Multimedia Retrieval using Cellular Neural Networks", University Politehnica of Bucharest. PhD supervisor Prof. Radu Dogaru;
- (2012 - 2015) **Anca-Livia Radu**: "Large Scale Media Analysis via Media Fusion and Crowdsourcing", co-tutelle thesis between University Politehnica of Bucharest, supervisor Prof. Corneliu Burileanu; and University of Trento, supervisor Prof. Fausto Giunchiglia and co-advisor Assoc. Prof. Bogdan Ionescu;
- (2013 - 2016) **Andrei Purică**: "Semantic Video Coding", co-tutelle thesis between University Politehnica of Bucharest, supervisor Prof. Corneliu Burileanu; and Télécom ParisTech, supervisor Prof. Frederic Dufaux and Prof. Beatrice Pesquet-Popescu;

- (2013 - 2016) **Ionuţ Duţă**: "Crossmedia Knowledge Extraction", University of Trento, supervisor Prof. Nicu Sebe and co-advisor Assoc. Prof. Bogdan Ionescu;
- (2012 - 2015) **Ioana Dumitrache**: "Natural Computing for Efficient Pattern Recognition in Medical Imaging", University Politehnica of Bucharest. PhD supervisor Prof. Radu Dogaru;
- (2013 - 2016) **Camelia Moldovan**: "Natural Computing with Applications in Bioinformatics", University Politehnica of Bucharest. PhD supervisor Prof. Radu Dogaru.

In particular, I am officially co-advisor of two PhD students, Anca-Livia Radu and Ionuţ Duţă, at the University of Trento, Italy.

4.4 Postdoctoral research

- (2014 - 2015) **Ionuţ Mironică**: "Large Scale Content-Based Search of Multimedia Information", LAPI - Video Processing Group, University Politehnica of Bucharest. Scholarship under POS-DRU InnoRESEARCH, grant ID132395. Supervisor Assoc. Prof. Bogdan Ionescu.

Chapter 5

Projects and strategic programmes

My research activity after finalizing the PhD was conducted as part of several research and strategic projects/programs, namely:

5.1 Medical imaging and systems

- 2006-2008: **researcher**, CEEX research grant "Digital Radiologic Image Analysis for the Monitorization of Orthopaedic Prosthesis", owner LAPI-University Politehnica of Bucharest;

- 2007-2009: **researcher**, research grant "Using Dog Cancer Model as Clinical Pretrial for Human Oncology", PRECLIN, owner The Oncologic Institute from Bucharest-Romania;

- 2008-2011: **in-charge with research** at UPB, CNMP research grant "Integrated System for Real-Time Assistance During the Surgical Act Through 3D Image Reconstruction and Visual Synchronization with Voice Commands", SIATRACH, owner University of Medicine and Pharmaceuticals "Carol Davila" from Bucharest;

- 2012-2015: **researcher**, UEFISCDI research grant "SRSPIRIM - New Innovative System for Radiation Safety of Patients Investigated by Radiological Imaging Methods, based on Smart Cards and PKI Infrastructures", owner ETTI - University Politehnica of Bucharest with partners: S.C. CertSIGN S.A. and "Carol Davila" Military Hospital, Bucharest.

5.2 Content-Based Indexing of Multimedia Documents

- 2007-2009: **principal investigator**, CNCSIS research grant "Content-Based Semantic Retrieval of Video Documents, Application to Navigation, Research

and Automatic Content Abstraction", RP-2, owner LAPI, University Politehnica of Bucharest;

- 2013-2015: **principal investigator**, UEFISCDI Innovation axis, product-oriented grant, "Smart Video Surveillance System with Automatic Localization of Target Events" SCOUTER, owner UTI Grup, partner LAPI, University Politehnica of Bucharest.

5.3 Networking

- 2006-2008: **researcher**, CEEX excellences project "The Software Infrastructure for the Data Acquisition System of the ATLAS Experiment", owner "Horia Hulubei" Romanian National Institute for Nuclear Physics, partners CERN-Geneve and LAPI-Bucharest.

5.4 Software development

- 2010-2013: **management team** - project assistant, European Structural Funds, POS-CCE RECOLAND, ID519, owner University Politehnica of Bucharest.

5.5 Strategic programmes

- 2008-2011: **IT expert**, project "Knowledge-Based Competitive Training of PhD Students in the Main Domains of Our Society", European Structural Funds POS-DRU project, owner University Politehnica of Bucharest, PhD Scholarships Programme, ID7713;

- 2009-2012: **IT expert**, project ProDOC, "Competitiveness and Performance in Scientific Research through High Quality PhD Programme", European Structural Funds POS-DRU project, owner University Politehnica of Bucharest, PhD Scholarships Programme, ID61178;

- 2010-2013: **implementing expert**, project PROMISE, "An Integrated Master's Degree Programme in the Fields of Sound, Image and Multimedia Engineering", European Structural Funds POS-DRU project, owner University Politehnica of Bucharest, Master Programme, ID61178;

- 2010-2013: **short-term implementing expert**, project ITEMS, "Analysis, Modeling and Simulation Techniques for Imaging, Bioinformatics and Sys-

tems", European Structural Funds POS-DRU project, owner University Politehnica of Bucharest, Master Programme, ID61756;

- 2012-2015: **management team**, project CAMPUS, "University Politehnica of Bucharest - Center for Advanced Research on Materials, Products and Innovative Processes", European Structural Funds POS-CCE project, owner University Politehnica of Bucharest, ID956;

Chapter 6

Relevant research results

During my 12 years of research, I have mainly worked in the areas of multidimensional signal processing, i.e., image, video and multimedia processing and analysis. My contributions can be grouped around three main topics, namely:

- *human-computer interaction*,
- *medical imaging*,
- *content-based indexing* (and information retrieval).

My major research direction is however the work on *content-based indexing* and *information retrieval* (known also as data indexing).

This chapter details the most important achievements on this topic. The presentation will focus mainly on the results obtained after my PhD. However, due to the fact that some of them rely on some previous research, for the sake of readability, I will briefly introduce those as well.

To situate my contributions, the following part consists of a brief introduction of the indexing concept.

6.1 Data indexing formalization

In general, the concept of data indexing is related to the notion of *data annotation*. Data annotation is the process of describing data content with additional information, significantly reduced in size compared to the initial data, but however enough representative to preserve the key specificity of that data. These information are known as *descriptors* or indexes.

Within a large collection of data, data without associated content descriptions is basically unaccessible for the common user, without the possibility of being included in the search mechanism. Let's take for example a common file system of any personal computer where information is preserved on some storage devices, e.g., hard drive, external memory, etc. Without the file indexing system that is integrated in the operating system, data is basically a collection of 0 and 1 with no specific meaning for the user. Although the information is there, it is basically unaccessible

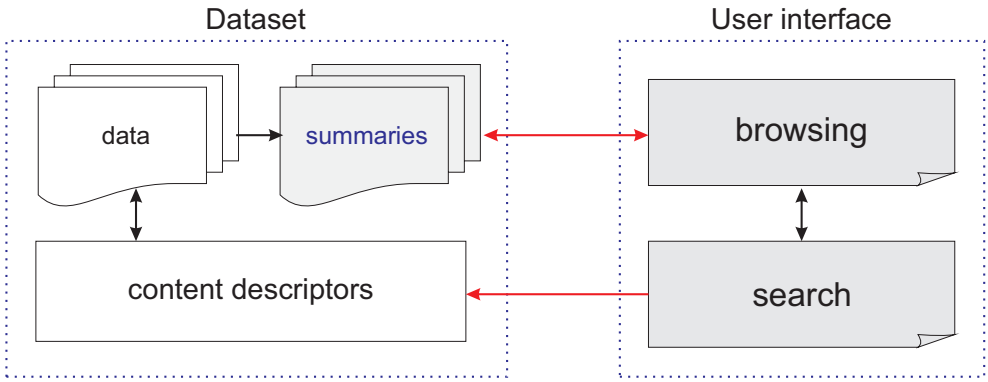


Figure 6.1: The components of a general purpose content-based indexing system (arrows depict the interaction between different blocks) [9].

for a (common) user. Enhancing this information with content indexes, such as appurtenance to a certain file, folder, information about content type, date, etc will allow further for data indexing and content-based search of the needed information.

Data annotation is critical for the indexing mechanism but is however not sufficient. The annotations themselves are only additional data, typically low-level, i.e., numbers. To access the actual information, the user needs to have some additional tools that convert those data into a mechanism for visualizing and searching. The other two components of an indexing system are therefore an information *browsing system* for visualizing data and a *searching mechanism*.

The diagram of a general purpose content-based indexing system is depicted in Figure 6.1. The typical content-based retrieval mechanism works as follows:

- **content description**: in a first stage, content descriptors are computed for the entire dataset. Descriptors are typically low-level information about the properties of the data, e.g., color-texture-shape information for images, audio parameters for sound, motion information for video, etc; or even high-level descriptors such as textual descriptions. This process is not necessarily time restricted and can be performed off-line before the system is started for the first time. Also, depending on the data, e.g., for video, content summarization can be performed. This will provide the possibility of quickly visualizing contents for browsing;

- **user query**: the user specifies the needed information via the formulation of a search query. There are various approaches to formulate a query, e.g., with textual descriptions of the requested data (which is one of the most popular), by providing examples, by gestures, miming, etc;

- **query conversion to descriptors**: the system converts the query in content descriptors using the same mechanism as for the dataset descriptors;

- **search**: the actual search is carried out by exhaustively comparing the descriptors of the query with the descriptors of the dataset. This is performed using a certain similarity measure, e.g., most commonly a distance metric. The results yielding the lowest distance score are provided in an ascending order to the user as being the most representatives for the search;

- **user interaction**: search results are provided to the user via the browsing system that uses an intuitive data visualization interface. Optional, search results can be refined and improved using users' feedback about their relevance, technique known as relevance feedback.

My research provided contributions to all these components of a content-based retrieval system as well as solutions to solve this paradigm in some particular application domains. These contributions are:

- **algorithms for video temporal segmentation**: temporal segmentation means decomposing the video into its temporal units or shots. This is commonly used as a pre-processing step for any other higher-level analysis. My major contributions were in the development of cut, fade and dissolve detection algorithms (presented in Section 6.2);

- **algorithms for video content description**: I have developed a number of techniques for representing color and action information in videos, contributed to a technique for capturing temporal variation as well as to multi-modal description and fusion (text-audio-visual; presented in Section 6.3);

- **algorithms for video summarization**: visualization of video data for video browsing requires efficient summarization tools. My contribution consisted in the development of algorithms for video summarization and video skimming (presented in Section 6.4);

- **algorithms for video genre tagging**: I have contributed to the development of a system that allows for the indexing of video materials according to video genres (e.g., movies, music, sports, etc). The research targeted the particular case of web specific video materials. I have also developed a technique for the automatic detection of the animated video genre (presented in Section 6.3.1);

- **algorithms for violent scenes detection**: I have contributed to the development of an indexing system for the automatic identification of violence

scenes in typical Hollywood productions. Moreover, I co-coordinated an evaluation benchmarking campaign on this topic (presented in Section 6.5);

- **algorithms for search result diversification**: typical content-based retrieval focuses on the relevance of the results. However, the user is not always interested in obtaining identical replicas of the search queries. I have contributed to the development of several approaches for diversifying the search results in order to cover different topics of the query. I also initiated and organized an evaluation benchmarking campaign on this topic (presented in Section 6.6);
- **algorithms for relevance feedback**: I have contributed to the development of several relevance feedback techniques that exploit user feedback to improve the relevance of image/video search results (presented in Section 6.7);
- **multimedia browsing systems**: I have contributed to the development of several video and multimedia browsing interfaces that allows for a virtual representation of the datasets and interactive user experience (presented in Section 6.8).

6.2 Video temporal segmentation

Most of the existing video analysis techniques rely on temporal segmentation as a preliminary processing step, because it provides a basic understanding of the movie temporal structure. At its highest level of granularity, temporal segmentation means parsing the video into its basic temporal units or *video shots*. A shot is defined as a continuous sequence of frames recorded between a camera switch on and off. In order to constitute the final video (usually denoted the final cut), in the editing phase video shots are linked together by means of *video transitions*. From this perspective, temporal segmentation roughly means detecting the video transitions that make the connection between consecutive shots.

There are two categories of video transitions: sharp and gradual. The most frequent are the *sharp transitions*, or cuts. A cut is a direct concatenation of two consecutive shots and produces an important visual discontinuity in the visual flow. Depending on video genre, 30 minutes of video may account for up to 300 cuts [28]. On the other hand, there are the *gradual transitions*, such as fades, dissolves, mattes, wipes, etc [29]. Gradual transitions are short effects. Their occurrence frequency in the sequence is more reduced, being at least one order measure less than cuts.

From the gradual transitions, the most commonly used within entertainment videos are the *fades* and *dissolves*. Fades are a gradual emerging of a certain image

from a constant image, typically black (i.e., a fade-in sequence) or vice-versa, the gradual disappearance of an image into a black frame (i.e., a fade-out sequence). Dissolves are closely related to fades (at some level they can be perceived as the superposition of a fade-out with a fade-in transition) and involve a gradual transition at pixel-intensity level of a certain image into another [30]. Compared to cuts, for which most of the actual detection techniques are highly accurate and common detection ratios are above 95% (see early TRECVid campaign [31]), gradual transitions are much difficult to detect. This is mainly due to the highly complex content transformations involved. Dissolve detection is one of the still open issues, current detection ratios being in average situated around 80%.

6.2.1 Cut detection

During my PhD I have developed an algorithm for the detection of cut transitions that exploits second order derivatives of histogram distances between consecutive frames and adaptive thresholding [28].

6.2.2 Fade detection

I have also improved a fade detection that exploits the specific signature of the variance and mean values of the intensity and chromatic image signals. This work was also conducted during my PhD [22].

6.2.3 Dissolves detection

After my PhD I have continued this direction in order to solve the more complex problem of dissolve detection[1]. As for the previous cases of cut and fade detection, research was carried out in the particular case of animated movies.

Artistic animated movies are very different from natural movies and even cartoons in many respects [25]. First, they are created using a large variety of animation techniques: paper drawing, salt animation, 3D synthesis, puppet animation, etc. Therefore, contrary to natural movies, many animated movies are created frame by frame thus affecting the continuity of the visual flow. In the dissolve context, this

[1]this work was developed in cooperation with Prof. Patrick Lambert, from LISTIC, Université de Savoie, Annecy-France. The presented results were published in:

[23] *B. Ionescu*, C. Vertan, P. Lambert, "*Dissolve Detection in Abstract Video Contents*", IEEE ICASSP - International Conference on Acoustic, Speech and Signal Processing, Prague, Czech Republic, 22-27 May, 2011;

[11] *B. Ionescu*, P. Lambert, "*An Intensity-Driven Dissolve Detection Adapted to Synthetic Video Contents*", SPIE - Journal of Electronic Imaging, 22(2), 023011, 2013.

may rend inefficient the general assumptions on the gradual or parabolic evolution of some intensity-based parameters, like the variance [33]. Also, this affects the motion content which is usually discontinuous (e.g., stop motion). Each animated movie has a specific color palette, as colors are selected and mixed by the artists to express particular feelings. Therefore, there is a strong color similarity between shots. We mainly deal with fiction or highly abstract movies, rich in visual effects. Usually, events don't follow any physical rules: characters appear/dissapear, they can take any shape, color, etc. Contour/edge information changes often from one image to another and exiting/entering contour pixels may not necessarily be related to dissolve transitions [34]. Overall, we deal with very complex visual contents and particularly, dissolve transitions usually show atypical patterns.

Contribution to state-of-the-art

To address these constraints, we proposed a straightforward efficient dissolve detection [23, 11][2] that exploits the hypothesis advanced in [35] according to which the pixel intensity in terms of amount of fading-out and fading-in pixels should be high during dissolves. The main novelty of our method is in the way we carry out the localization of the dissolves within the discontinuity function. Instead of just applying a global threshold, as most of the existing approaches do, we use a twin-thresholding approach and the shape analysis of the signal. This approach allows to reduce false detections caused by steep intensity fluctuations (due to noise, movement, visual effects, etc), as well as to retrieve dissolves caught up in other visual effects or scene movements (very frequent in animated movies). Additionally, to overcome the restraint visual continuity of the animated movies, fading-out and fading-in pixels are selected at intensity level from a reduced time window of only several frames.

Algorithm

For the detection, we use only pixel intensity information which is obtained with the Y luminance component after converting initial RGB images to YCbCr color space (Y - intensity, Cb, Cr - chromatic components). The method diagram is presented in Figure 6.2.

For each analyzed frame at time index k, denoted I_k, we first determine the number of fading-out pixels (denoted FOP_k), i.e., pixels whose intensity decreases during next w frames, and fading-in pixels (denoted FIP_k) whose intensity increased during previous w frames. Due to the reduced visual continuity of animated movies,

[2] a demo of the system is available here: `http://imag.pub.ro/~bionescu/index_files/DemoDissolves.wmv`.

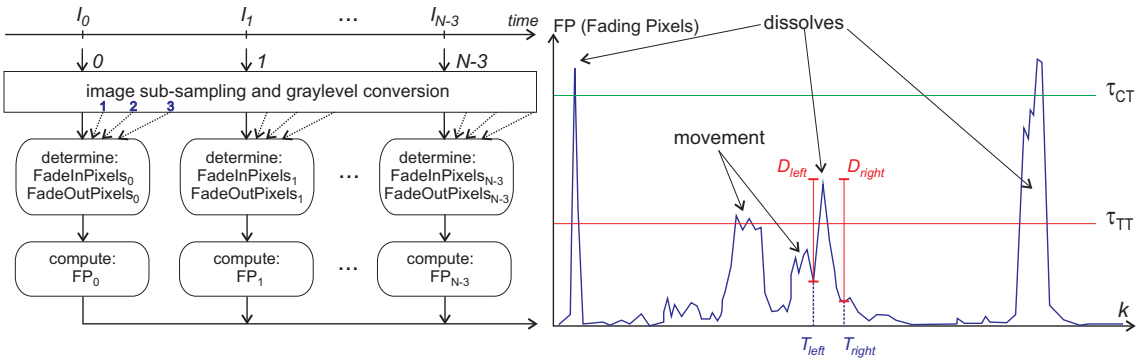


Figure 6.2: Diagram of the proposed dissolve detection [11] (I_k is the frame at time index k, N is the sequence length, FP represent the fading pixels, $1, 2, 3$ denote frames from time window w). Left image exemplifies the twin-thresholding mechanism used for the detection.

we have restricted the search window to only a few frames. At the beginning of the dissolve there are more FOP than FIP. As the first image starts disappearing, the number of FOP increases but also FIP as the final image starts to appear. Both ratios reach their maximum for the middle frame of the dissolve. Finally, as the final image emerges more and more, FIP become more predominant than FOP, which finally disappear in the end.

In animated movies the constant presence of displacements/movements or of color effects make this process to be likely unbalanced, i.e., proportions of FOP and FIP are not equal during dissolve. Therefore, instead of monitoring high values of FOP and FIP, independently, we determine a normalized visual discontinuity function by taking a simple ratio of the two values, thus:

$$FP_k = \frac{(FOP_k + FIP_k)}{H \cdot W} \tag{6.1}$$

where $H \cdot W$ is the image size. Defined in this way, as stated before, ideally, FP should reach its maximum for the dissolve middle frame (when both shot images are as much as visible).

The main novelty of our approach lies however in the way we carry out the localization of a dissolve within the FP function. We propose the following twin-thresholding approach:

- **Case I**: if FP_k for the current frame I_k is greater than a first threshold, denoted τ_{CT} (Certain Threshold), I_k is very likely to be the middle frame of the dissolve, being characterized by, both, high values of FOP_k and FIP_k respectively. If

this value is a local maximum (both, previous and next values are decreasing), then a dissolve is declared in the time interval $[k - t_{max}/2; k + t_{max}/2]$, where t_{max} is an average estimate of a maximum dissolve length;

- **Case II**: on the other hand, if FP_k for image I_k is greater than a second threshold, denoted τ_{TT} (Tolerant Threshold), but still beneath τ_{CT}, then the image is considered to be a potential dissolve middle frame. Further validation is to be performed and consists mainly on the shape analysis of FP values in the neighborhood of the frame I_k.

In the case of a dissolve, values of FP_k should decrease, both, on the positive and negative time axis. Therefore, we seek for the time moments $T_{left} < k$ and $T_{right} > k$, when FP_k starts increasing again, thus:

$$FP_{T_{left}} < FP_{T_{left}-1} \ \wedge \ FP_{T_{right}} < FP_{T_{right}+1} \tag{6.2}$$

To quantify the relevance of FP_k with respect to neighbor values, we compute the height of the peak on both sides, thus:

$$D_{left} = |FP_k - FP_{T_{left}}|, \quad D_{right} = |FP_k - FP_{T_{right}}| \tag{6.3}$$

Similar to the previous case, we decide that a dissolve occurred in the time interval $[k - t_{max}/2; k + t_{max}/2]$ if the distance values are greater then a fraction of FP_k, that is:

$$D_{left} > 0.5 \cdot FP_k \ \wedge \ D_{right} > 0.5 \cdot FP_k \tag{6.4}$$

In this way, we ensure that FP_k is a local maximum, significant enough compared to neighbor values and which has an increase on both sides of at least 50%, compared to local neighbor minimum.

Intensity fluctuations may also result in several representative peaks of the FP_k function during the same dissolve. Therefore, we may by mistake select multiple frames as dissolve middle frames within same transition. To avoid this situation, the final step consists on fusing close overlapping dissolves.

Validation results

To test our approach we have selected movies created with a high diversity of animated techniques that fall in two categories of contents: highly complex (abstract, very complex visual contents, motion discontinuity - denoted ↑) and average complexity (average amount of visual effects, motion content less discontinuous - denoted

Table 6.1: Dissolve detection results [11].

movie	*count*	GD	FD	P	R
"Ex-Enfant"↑	75	65	8	89%	86.7%
"Le Moine et le Poisson"↔	61	47	2	95.9%	77.1%
"M. Pascal"↑	98	76	2	97.4%	77.6%
"Une Bonne Journée"↔	19	19	0	**100%**	**100%**
"Paradise"↑	60	44	7	86.3%	73.3%
"Cœur de Secours"↑	67	47	2	95.9%	70.2%
"The Sand Castle"↔	72	62	2	96.9%	86.1%

↔). The test data set consists of 61 minutes of video footage and contains a total number of 452 dissolve transitions[3].

To assess performance we use classic precision and recall:

$$P = \frac{GD}{GD + FD},\ R = \frac{GD}{GD + ND} \tag{6.5}$$

where GD is the number of good detections (True Positives - TP), FD is the number of false detections (False Positive - FP) and ND is the number of non-detections (False Negative - FN), where $GD + ND = 452$ (i.e., the total number of dissolves). The detection results are summarized with Table 6.1.

Overall, we score 360 good detections and a very reduced number of false detections, i.e., only 23 (for most of the sequences < 2). This leads to an average precision of 94% and a recall of 79.6%. At sequence level, precision and recall ranges from $[86.3; 100]\%$ and $[70.2; 100]\%$, respectively. The highest detection ratio is obtained for movie "Une Bonne Journée" which has a more accessible content ($P = R = 100\%$), while the lowest detection ratio is obtained for the movie "Paradise" due to its very complex content ($P = 86.26\%$, $R = 73.33\%$).

We attempted to compare our approach against other reference methods from the literature. We use two confirmed approaches, namely: the assessment of the variance of pixel intensities, which during dissolves should yield a quadratic behavior [33], and the use of Edge Change Ratio [34] for which edges were obtained with a Canny edge detector with automatic thresholding. The experimental results proved that classic methods tend to be rather inefficient when used on this type of contents.

Figure 6.3 exemplifies the three approaches for several representative movies (for brevity reasons, we limit to exemplify only four movies). Due to the discontinuous nature of the motion content and to the presence of visual effects, variance of pixel

[3]movies are available for free preview or for purchasing at CITIA (`http://www.citia.info/`) and/or on YouTube (`http://www.youtube.com/`).

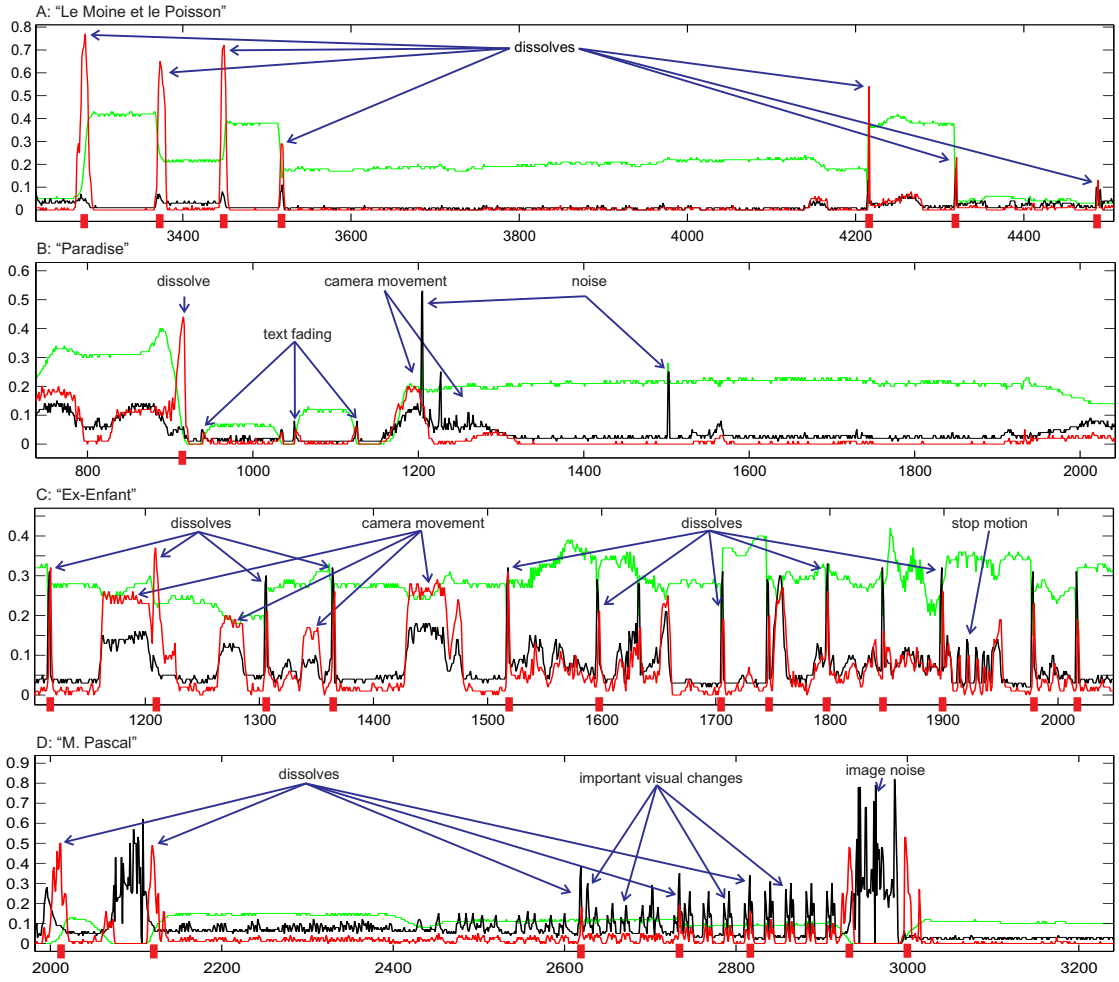


Figure 6.3: The proposed discontinuity function FP_k (in Red) [11] against intensity variance (in Green, values are scaled to fit the other functions) and Edge Change Ratio (in Black). Dissolves which were successfully detected are marked on the temporal axis (oX) with vertical Red segments.

intensities do not follow a parabolic shape. Instead, it has an unpredictable behavior (see the Green line in Figure 6.3, e.g., movie C) or unexpectedly decreasing or increasing during dissolves (e.g., movies A, B or D). On the other hand, contour information (i.e., ECR), whether for some particular cases it provides good discrimination (similar to FP, see in Figure 6.3 the Black line for movie C), in general is either non-discriminant (see movie A where ECR has small values during dissolves) or highly sensitive to noise and visual changes (see movies B and D where important peaks are due not to dissolves but to noise, fading effects and important changes in object structure).

Table 6.2: Dissolve detection processing time [11].

image size (pixels)	frames/s	time processing 10 min.
60×45	134	112s
120×90	128	117s
240×180	94	160s
480×360	72	208s
740×480 (original)	43	349s

On the other hand, the proposed method provides good results in all situations (see the good detections in Figure 6.3). Thanks to the shape analysis, which adapts to local contents, it is discriminant enough to retrieve dissolves even when mixed-up with motion (see the Red line in Figure 6.3, e.g., the first detected dissolve in movie B or the second detected dissolve in movie C which are successfully separated from camera movement) and to avoid false detections (see movie C where camera movement and other visual effects are not taken as dissolves despite their high FP).

The achieved results are very promising considering the difficulty of the test sequences or even compared to existing approaches, which applied to natural movies achieve average detection ratios around 80%.

Finally, the proposed approach provides also good computational performance. Table 6.2 presents the results obtained on a regular laptop computer, CPU Intel(R) Core(TM) i5 M460@2.53GHz, 4GB of RAM running on Microsoft Windows 7 - 64 bit (for calculations we consider a frame rate of 25 frames per second). The presented processing time includes also the delays caused by the visual interface, as images are displayed as being processed (application developed under Borland C++ Builder 6). For instance, at a frame resolution of 120×90 pixels it achieves more than 128 frames per second (5 times faster than real time). Compared to using original frame resolution, it is three times faster, for the later we achieve only 43 frames per second. Nevertheless, even in this case, the detection performs almost twice faster than real time. In terms of detection errors, we obtain results very similar at all scales, therefore, by reducing the image size we increase the performance efficiency.

Conclusions and future work

We proposed a dissolve detection approach that addresses the specificity of the animated videos. The proposed method exploits pixel intensities in terms of amount of fading-out and fading-in pixels. The main novelty of our method is in the way we carry out the localization of the dissolves within the discontinuity function. We use a twin-thresholding mechanism and the shape analysis of the signal.

Experimental tests conducted on more than 452 dissolve transitions show the

potential of this approach in cases where traditional methods (adapted to natural movies) tend to fail. It allows to reduce false detections caused by steep intensity fluctuations (e.g., due to noise, movement, visual effects, etc), as well as to retrieve dissolves caught up in other visual effects or scene movements (very frequent in animated movies) leading to precision and recall ratios up to 100%. In terms of computational complexity, the proposed method performs five times faster than real time on a regular computer.

The method seems to be less efficient when dealing with some very complex scene changes and fade transitions that involve camera movement and special effects. Another limitation is the impossibility of determining the exact dissolve boundaries. Future work will mainly consists on addressing this limitation by investigating the behavior of various features in the neighborhood of the start/end frames of a dissolve.

6.3 Video content description

As introduced at the beginning of this chapter, the content-based retrieval mechanism relies mainly of the data annotation process, i.e., the computation of content descriptors. Content descriptors should be computed to be as representatives as possible for the underlying contents, invariant to various transformations and noise as well as in the same time efficiently represented (i.e., reduced in size). Their representative power influence dramatically the performance of the retrieval [9]. For multimedia data, content descriptors can be extracted from three major information sources: *visual* (e.g., information about colors, shape, texture, motion), *audio* (e.g., information that can be extracted from sound, such as speech and music parameterization) and *text* (e.g., information from inlaid image text, subtitles, associated metadata, text obtained with speech recognition, etc).

6.3.1 Multimodal content description

One of my first contributions to content description was in the area of multi-modal video representation with visual and audio information[4].

[4]this work was developed in cooperation with Dr. Klaus Seyerlehner, from Department of Computational Perception, Johannes Kepler University, Linz-Austria, Dr. Christoph Rasche, Dr. Ionuţ Mironică, Prof. Constantin Vertan, from LAPI, University Politehnica of Bucharest, Romania, and Prof. Patrick Lambert, from LISTIC, Université de Savoie, Annecy-France. The presented results were published in:

[16] *B. Ionescu*, K. Seyerlehner, C. Rasche, C. Vertan, P. Lambert, "*Video Genre Categorization and Representation using Audio-Visual Information*", Journal of Electronic Imaging, 21(2), 2012.

[4] *B. Ionescu*, K. Seyerlehner, I. Mironică, C. Vertan, P. Lambert, "*An Audio-Visual Approach to Web Video Categorization*", Multimedia Tools and Applications, 70(2), 2014.

Contribution to state-of-the-art

We developed and adapted a certain number of techniques for audio, temporal structure, color content and spatial structure representation [16, 4].

The proposed *audio features* are block-level-based and have the advantage of capturing local temporal information by analyzing sequences of consecutive frames in a time-frequency representation [40]. *Visual information* is described using temporal information, color, and structural properties. Temporal descriptors are derived using a classic confirmed approach, that is, analysis of the shot change frequency [36]. However, we use a novel way of measuring action content that assesses action perception. Color information is extracted globally. In contrast to existing approaches, which mainly use local or low-level descriptors such as predominant color, color variance, color entropy, and frame based histograms [37], our approach analyzes color perception. Using a color naming system, we quantify color perception with statistics of color distribution, elementary hues distribution, color properties (e.g., percentage of light colors, cold colors, saturated colors), and relationships between colors [24]. The final type of visual descriptor is related to contour information. Unlike most existing approaches, which describe closed region shapes (e.g., with MPEG-7 visual descriptors [38]), contours are broke down into segments and describe curve contour geometry both individually and relative to neighbor contours [39].

Approach

Audio descriptors. The proposed set of audio descriptors, called *block-level audio features*, has the key advantage of capturing temporal information from the audio track at a local level [40]. In contrast to standard spectral audio features (e.g., Mel Frequency Spectral Coefficient, Spectral Centroid, or Spectral Roll Off), which are typically extracted from each spectral frame (capturing a time span of 20 ms) of the time-frequency representation of an audio signal, the proposed features are computed from a sequence of consecutive spectral frames called a *block*. Depending on the extracted block-level feature, a block consists of 10 up to about 512 consecutive spectral frames. Thus, local features can themselves capture temporal properties (e.g., rhythmic aspects) of an audio track over a time span ranging from half a second up to 12 seconds of audio. Blocks are analyzed at a constant rate and their frames overlap by default by 50%, which results in one local feature vector per block. These local vectors are then summarized by computing simple statistics (e.g., mean, variance, or median) separately for each dimension of the local feature vectors. A schematic diagram of this procedure is depicted in Figure 6.4.

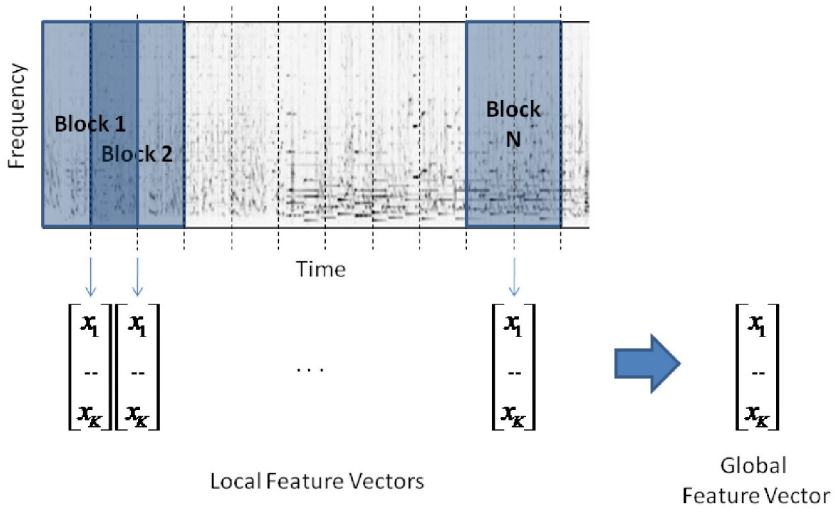


Figure 6.4: Processing a time (OX axis) - frequency (OY axis) representation in terms of spectral blocks (N is the number of blocks) [40].

To obtain a perceptual time-frequency representation of the video soundtrack, the audio track is first converted into a $22kHz$ mono signal. Then we compute the short-time Fourier transform and perform a mapping of the frequency axis according to the logarithmic cent-scale because human frequency perception is logarithmic. The resulting time-frequency representation consists of 97 logarithmically spaced frequency bands. The following complex block-level audio features are derived:

- **spectral pattern** (number of frames constituting a block: 10 frames, 0.9 percentile statistics): characterizes the timbre of the soundtrack by modeling the frequency components that are simultaneously active. The dynamic aspects of the signal are retained by sorting each frequency band of the block along the time axis. The block width varies depending on the extracted patterns, which allows capturing temporal information over different time spans;

- **delta spectral pattern** (14 frames, 0.9 percentile statistics): captures the strength of onsets. To emphasize onsets, we first compute the difference between the original spectrum and a copy of the original spectrum delayed by three frames. As with the spectral pattern, each frequency band is then sorted along the time axis;

- **variance delta spectral pattern** (14 frames, variance statistics): is basically an extension of the delta spectral pattern and captures the variation of the onset strength over time;

- **logarithmic fluctuation pattern** (512 frames, 0.6 percentile statistics): captures the rhythmic aspects of the audio signal. In order to extract the amplitude modulations from the temporal envelope in each band, periodicities are detected by computing the FFT (Fast Fourier Transform) along each frequency band of a block. The periodicity dimension is then reduced from 256 to 37 logarithmically spaced periodicity bins;

- **spectral contrast pattern** (40 frames, 0.1 percentile statistics): roughly estimates the *"tone-ness"* of an audio track. For each frame within a block, the difference between spectral peaks and valleys in 20 sub-bands is computed, and the resulting spectral contrast values are sorted along the time axis in each frequency band;

- **correlation pattern** (256 frames, 0.5 percentile statistics). To capture the temporal relation of loudness changes over different frequency bands, the correlation coefficients for all possible pairs of frequency bands within a block are used. The resulting correlation matrix forms the correlation pattern. The correlation coefficients are computed for a frequency resolution of 52 bands.

Temporal structure descriptors. Temporal descriptors are derived by means of a classic confirmed approach, that is, analysis of the shot change frequency [36]. Unlike existing approaches, we determine the action content based on human perception. First, we detect video transitions [41] using appropriate tools (e.g., cuts, fades, dissolves). The temporal descriptors are then computed as follows:

- **rhythm**. To capture the movie's tempo of visual change, we define a basic indicator, denoted $\zeta_T(i)$, which represents the relative number of shot changes occurring within the time interval T, starting at the frame at time index i. Based on ζ_T, we define the movie rhythm as the movie's average shot change speed, denoted $\bar{v}_T$, which is the average number of shot changes in the time interval T for the entire movie, thus $E\{\zeta_T\}$;

- **action**. We aim to define two opposite situations: video segments with a high action content (denoted "hot action", e.g., fast changes, fast motion, visual effects) and video segments with low action content (denoted "low action", containing mainly static scenes). The process is illustrated in Figure 6.5.

  First, at a coarse level, we identify segments containing a high number of shot changes ($\zeta_T > 3.1$), which are "hot action" candidates, and a reduced number of shot changes ($\zeta_T < 0.6$), which are low action candidates (see step a in Figure 6.5). Thresholds were determined experimentally based on human

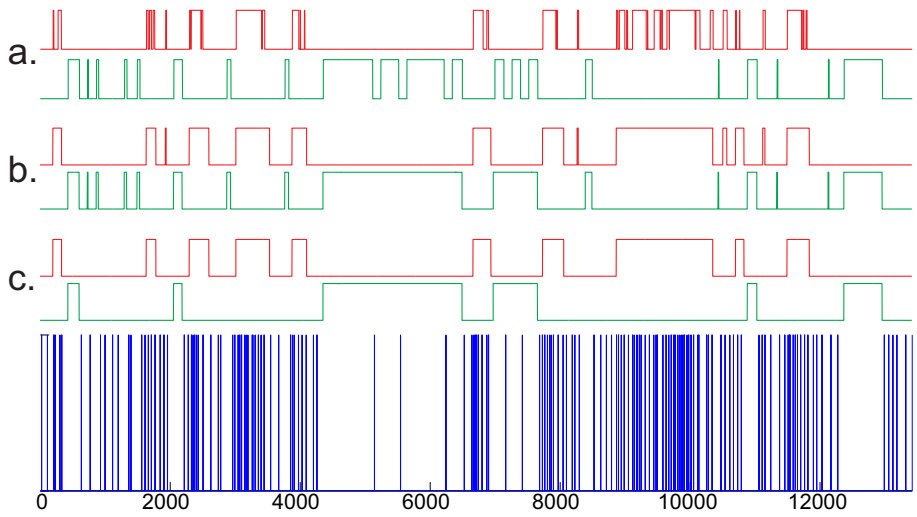


Figure 6.5: Action-based temporal segmentation [16] (the OX axis is the temporal axis, vertical blue lines correspond to shot changes). "Hot action" and "low action" segments are indicated in red and green respectively. Letters denote the processing steps as described in the text.

perception: several persons were asked to manually classify video segments into the two categories mentioned. On the basis of this ground truth, we determined average ζ_T intervals for each type of action content.

To avoid over-segmentation, we merge neighboring action segments at a time distance smaller than T seconds (the size of the time window, see step b in Figure 6.5). Further, we remove unnoticeable and irrelevant action segments by erasing small action clips of length less than the analysis time window T. Finally, all hot action clips containing fewer than $N_s = 4$ video shots are removed because they are very likely the result of false detection and contain one or several gradual transitions (e.g., a "fade-out" - "fade-in" sequence, see step c in Figure 6.5). Using this information, we quantify the action content by two parameters, hot-action ratio (HA) and low-action ratio (LA):

$$HA = \frac{T_{HA}}{T_{total}}, \quad LA = \frac{T_{LA}}{T_{total}} \tag{6.6}$$

where T_{HA} and T_{LA} represent the total lengths of hot and low action segments, respectively, and T_{total} is the total length of the movie.

- **gradual transition ratio**. The gradual transition ratio (GT) is computed

as:

$$GT = \frac{T_{dissolves} + T_{fade-in} + T_{fade-out}}{T_{total}} \tag{6.7}$$

where T_x represents the total duration of all the gradual transitions of type x. This provides information about editing techniques which are specific to certain genres.

Color descriptors. Color information is a powerful mean for describing visual perception. We propose an elaborated strategy which addresses the perception of color content. A simple and efficient way to accomplish this is using color names; associating names with colors allows creating a mental image of a given color or color mixture. We project colors onto a color naming system [43], and color properties are described using statistics of color distribution, elementary hue distribution, color visual properties (e.g., percentage of light colors, warm colors, saturated colors, etc.), and relationships between colors (adjacency and complementarity). Color descriptors are extracted globally taking the temporal dimension into account.

Prior to parameter extraction, we project colors onto a more manageable color palette (initial images are true color). We selected the non-dithering 216 color Webmaster palette because of the high color diversity and its efficient color naming system: each color is named according to the degree of hue, saturation, and intensity [27]. Color mapping is performed with a classic dithering scheme, and colors are selected in the L*a*b* color space. Further, the proposed color descriptors are computed as follows:

- **global weighted color histogram** captures the global color distribution of the movie. It is computed as the weighted sum of each individual shot average color histogram:

$$h_{GW}(c) = \sum_{i=0}^{M} \left[\frac{1}{N_i} \sum_{j=0}^{N_i} h^j_{shot_i}(c) \right] \cdot \frac{T_{shot_i}}{T_{total}} \tag{6.8}$$

where M is the total number of video shots, N_i is the total number of frames retained from shot i (to reduce computational load, each shot is summarized by retaining $p = 10\%$ of its frames), $h^j_{shot_i}()$ is the color histogram of frame j from shot i, c is a color index from the Webmaster palette, and T_{shot_i} is the total length of shot i. The longer the shot, the more important the contribution of its histogram to the global histogram of the movie. Thus, values of $h_{GW}()$ describe global percentages of colors appearing in the movie;

- **elementary color histogram**. This feature is computed by:

$$h_E(c_e) = \sum_{c=0}^{215} h_{GW}(c)|_{Name(ce) \subset Name(c)} \tag{6.9}$$

where c_e is an elementary color from the Webmaster color dictionary, $c_e \in \Gamma_e$ with Γ_e = {"Orange", "Red", "Pink", "Magenta", "Violet", "Blue", "Azure", "Cyan", "Teal", "Green", "Spring", "Yellow", "Gray", "White", "Black"}, and $Name()$ is an operator that returns a color name from the palette dictionary.

Thus, each available color is projected in $h_E()$ onto its elementary hue, while saturation and intensity information are disregarded. This mechanism removes susceptibility to color fluctuations (e.g., illumination changes) and provides information about predominant hues;

- **color properties**. These parameters aim to describe color perception by means of light/dark, saturated/non-saturated, warm/cold colors and color richness by quantifying color variation/diversity. Using the previously determined histogram information in conjunction with the color naming dictionary, we define several color ratios. For instance, the light color ratio, P_{light}, which reflects the percentage of bright colors in the movie, is computed by:

$$P_{light} = \sum_{c=0}^{215} h_{GW}(c)|_{W_{light} \subset Name(c)} \tag{6.10}$$

where c is the index of a color whose name (provided by $Name(c)$) contains one of the words defining brightness, and $W_{light} \in$ {"light", "pale", "white"}. Using the same reasoning and keywords specific to each color property, we define:

  - *dark color ratio*, denoted P_{dark}, where $W_{dark} \in$ {"dark", "obscure", "black"};
  - *hard color ratio*, denoted P_{hard}, which reflects the number of saturated colors. $W_{hard} \in$ {"hard", "faded"} $\cup \Gamma_e$, where Γ_e is the elementary color set;
  - *weak color ratio*, denoted P_{weak}, which is opposite to P_{hard}, $W_{weak} \in$ {"weak", "dull"};
  - *warm color ratio*, denoted P_{warm}, which reflects the number of warm colors; in art, some hues are commonly perceived to exhibit levels of warmth, namely: "Yellow", "Orange", "Red", "Yellow-Orange", "Red-Orange", "Red-Violet", "Magenta", "Pink" and "Spring";

– *cold color ratio*, denoted P_{cold}, where "Green", "Blue", "Violet", "Yellow-Green", "Blue-Green", "Blue-Violet", "Teal", "Cyan" and "Azure" describe coldness.

Further, we capture movie color richness with two parameters. Color variation, P_{var}, which represents the number of significant colors, is defined as:

$$P_{var} = \frac{Card\{c|h_{GW}(c) > \tau_{var}\}}{216} \tag{6.11}$$

where c is a color index, h_{GW} is the global weighted histogram defined in equation 6.8, and $Card()$ is the cardinal function, which returns the size of a data set. We consider a color significant if it has a frequency of occurrence in a movie of more than 1% (i.e., $\tau_{var} = 0.01$). Color diversity, P_{div}, which reflects the number of significant color hues in the movie, is defined using the same principle, but based on the elementary color histogram h_E;

- **color relationship**. The final two parameters are related to the concept of perceptual relationships between colors in terms of adjacency and complementarity. The parameter, P_{adj} reflects the number of perceptually similar colors in the movie (neighborhood pairs of colors on a perceptual color wheel, e.g., Itten's color wheel [44, 27]), and P_{compl} reflects the number of perceptually opposite color pairs (antipodal).

Spatial structural descriptors. The final set of parameters provides information based on structure, that is, on contours and their relations. So far, contour information has been exploited to a very limited extent in video description. For instance, some approaches use MPEG-7-inspired contour descriptors [38], such as texture orientation histograms, edge direction histograms, edge direction coherence, [45], which do not exploit real contour geometry and properties.

The adopted appraoch, in contrast, proposes a novel method which uses curve partitioning and curve description [39]. The contour description is based on a characterization of geometric attributes of each individual contour, for instance, degree of curvature, angularity, and "wiggliness":

- **contour characterization**. Contour processing starts with edge detection, which is performed with the Canny edge detection algorithm [46]. For each contour, a type of curvature space is created. This space is then abstracted into spectra-like functions, from which a number of geometric attributes, such as the degree of curvature, angularity, circularity, symmetry and "wiggliness", are derived. In addition to these geometric parameters, a number of "appearance"

parameters are extracted. They are based on simple statistics obtained from the luminance values extracted along the contour, such as contrast (mean and standard deviation) and "fuzziness", obtained by convolution of the image with a blob filter. In summary, for a given image with n extracted and partitioned contours, we obtain a list of 7 geometric and 4 appearance attributes for each contour. For each attribute, a 10-bin histogram with n values, is generated;

- **pair relations**. In addition to individual contour attributes, we also obtain attributes for pairs of contours. Contour segments are first grouped as a list of all $n!$ pairs. From this long list of pairs, only a subset (approximately $3 \times n$) is selected based on spatial proximity, meaning that their contour endpoints are either proximal or in the proximity of other segments. For each selected pair, a number of geometric attributes is determined: the angular direction of the pair, the distance between the proximal contour endpoints, the distance between the distal contour end points, the distance between segment center (middle) points, the average segment length, the symmetry of the two segments, the degree of bendiness of each segment, the structural biases which express to what degree the pair alignment is an L feature, a T feature, or a "closed" feature (two curved segments facing each other as '( )'). In total, 12 geometric relational attributes are extracted for the selected pairs. Again, for each attribute a 10-bin histogram is generated;

The structural information is extracted only from a summary of the movie. In this case, we retained around 100 images that are evenly distributed with respect to video transitions. As previously mentioned, at image level, contour properties are captured with histograms. To address the temporal dimension - at sequence level - the resulting concatenated feature vectors are averaged to form so the structure signature of the movie.

Validation for animated genre detection

In the context of automatic content-based retrieval, a common task is the automatic selection of the "animated" content from other video genres. The first validation of the proposed temporal structure and color descriptors was for the construction of such a detector. Applications are with children program selection tasks.

To approach this task, descriptor parameters were tuned to optimal values with respect to the specificity of animated movies, e.g., action thresholds were determined after a user study on some representative movies, color perceptual schemes were consider, etc [24].

Validation tests were conducted on a diverse video database, i.e., 749 clips, with a high diversity of genres and sub-genres (more than 159 hours of video footage retrieved mainly from several TV chains). The animated genre is represented with 209 sequences (54 hours) containing: artistic animated movies (source CITIA [32]), films and cartoon series (source Disney, Pixar, DreamWorks animation companies). The non animated genre is represented with 541 sequences (105 hours), namely: 320 commercials (4 hours, source 1980th TV commercials and David Lynch clips; some clips are containing both animated graphics and natural scenes); 74 documentaries (32 hours, both outdoor and indoor series, source BBC, IMAX, Discovery Channel); 57 movies (43 hours, both long movies and soap series, e.g., Friends, X-Files); 43 news broadcasting (19 hours, source TVR Romanian National Television Channel); 16 sports (4 hours, mainly soccer and outdoor extreme sports); 30 music clips (3 hours, source MTV Channel: dance, pop, techno music).

Animated genre detection is carried out with a binary classification approach, i.e., considering two classes: animated and non-animated. Each movie is represented with a feature vector, according to the previously presented content descriptors (several combinations are tested). For the classification, we use three approaches: the k-Nearest Neighbors algorithm (KNN, with k=5, cosine distance and majority rule), Support Vector Machines (SVM, with a linear kernel) and Linear Discriminant Analysis (LDA, applied on a PCA-reduced feature space) [47]. The method parameters were set to optimal values for this scenario after several preliminary tests.

As the choice of the training set may distort the accuracy of the results, we have adopted an exhaustive testing. Tests were performed for different amounts of training data (see the beginning of Table 6.3). For each set, tests are repeated using a cross validation approach, thus generating all possible combinations between training and test data, in order to shuffle all sequences.

To assess performance, we use the classic precision and recall ratios - see equation 6.5 - which are averaged over all the experiments.

Figure 6.6 depicts the obtained precision vs. recall curves for different amounts of training data and different runs. With KNN we obtain a precision and recall up to 90.1% and 78.6%, respectively (using all descriptors), with LDA the precision and recall are up to 74.7% (using all descriptors) and 92.4%, respectively (using only action descriptors) while with SVM precision and recall are up to 74.7% (using h_E histogram) and 74.9%, respectively (using all descriptors). Overall, the highest precision is achieved with KNN on all the descriptors, thus 90.1%, while the highest recall is obtained with LDA on action descriptors, namely 92.4%.

However, the best method in terms of both precision and recall proves to be KNN run on all action-color descriptors together. The resulted average precision,

Table 6.3: Animated genre detection [24]: KNN on all temporal structure and color descriptors.

train (%)	#train seq.	#train anim.	#test seq.	#test anim.	P (%)	R (%)	$\overline{GD}$	$\overline{FD}$ #	pub.	doc.	mov.	news	sp.	mus.	$\overline{ND}$
10	75	21	674	188	68	66	124	58	24	5	15	5	9	0	64
20	150	42	599	167	76	70	117	37	12	4	10	4	7	0	50
30	225	63	524	146	80	73	106	26	7	4	6	3	6	0	40
40	300	84	449	125	84	74	92	17	5	3	3	2	4	0	33
50	375	105	374	104	87	75	78	12	4	2	2	1	3	0	26
60	450	126	299	83	88	76	63	9	3	2	1	1	2	0	20
70	524	147	225	62	89	77	48	6	2	1	1	1	1	0	14
80	599	168	150	41	89	78	32	4	1	1	1	0	1	0	9
90	674	189	75	20	90	79	16	2	1	1	0	0	0	0	4

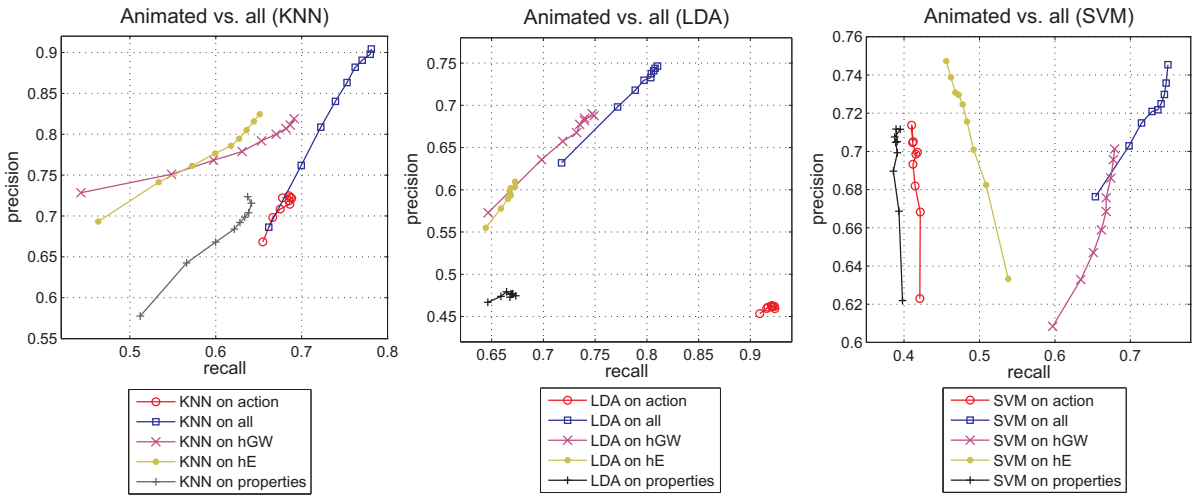


Figure 6.6: Animated genre detection [24]: precision vs. recall curves for different runs (action descriptors, h_{GW} - equation 6.8, h_E - equation 6.9, color properties and all parameters together) and amounts of training data (% of training is increasing along the curves).

recall, $\overline{GD}$ (good detections), $\overline{FD}$ (false detections) and $\overline{ND}$ (non detections) are presented in detail in Table 6.3 (for visualization purpose, actual real data values are to be rounded to nearest integer value).

The results are very promising considering the diversity of the video material (including a high variety of animated genres). For only 10% of training, average precision and recall are around 70% when testing on 674 sequences from which 188 are animated, while for 50% training precision approaches 90% and recall 80%. Also, one may observe the reduced number of false detections while maintaining a good detection ratio. For instance, using 70% training we obtain in average 48 good detections, only 6 false detections and 14 non detections.

To analyze which genres were wrongly classified in the animated category, we present in Table 6.3 the distribution of false positives to the other genres (we use the notations: pub. = commercials, doc. = documentaries, mov. = movies, sp. = sports, mus. = music). From the six genres, the most distinctive proves to be the music genre. None of the clips were classified as animated (regardless the amount of training). This is due to their very distinctive color signature (typically darker colors due to the intensive use of visual effects) and a high visual rhythm (lot of changes over a short period of time). On the other hand, the most wrongly classified genre are the commercials. This is mainly because many of them involve a lot of computer graphics and animation (also there is a practical reason, the test database includes a lot of commercials, compared to the other genres). On the third place are the movies, which for a small amount of training tend to be confounded with animation (several movies are science fiction series, thus involving an abstract contents). Other genres, are misclassified occasionally.

Validation for web genre classification

Automatic labeling of video footage according to genre is a common requirement in indexing large and heterogeneous collections of video material, especially to web collections. The validation of the proposed audio-visual descriptors was carried out in the context of automatic classification of video according to genres, e.g., "cartoons", "music", "news", "sports", "documentaries". Applications of such systems are in the automatic categorization of videos for TV programs, video selling platforms or genre based visualization of web media in dedicated platforms such as YouTube.

Validation tests were carried out in the context of the 2011 MediaEval Video Genre Tagging Task [48] and addressed a real-world scenario, namely the automatic categorization of web video genres from the blip.tv media platform[5].

The test dataset consisted of 2,375 sequences (around 421 hours of video footage) labeled according to 26 video genre categories (the numbers in brackets are the numbers of available sequences): "art" (66), "autos and vehicles" (36), "business" (41), "citizen journalism" (92), "comedy" (35), "conferences and other events" (42), "documentary" (25), "educational" (111), "food and drink" (63), "gaming" (41), "health" (60), "literature" (83), "movies and television" (77), "music and entertainment" (54), "personal or auto-biographical" (13), "politics" (597), "religion" (117), "school and education" (11), "sports" (117), "technology" (194), "environment" (33), "mainstream media" (47), "travel" (62), "video blogging" (70), "web development and sites" (40) and "default category" (248, comprises movies that cannot be assigned to any of the previous categories).

[5] `http://blip.tv/`.

The main challenge in classifying these videos comes from the high number of different genres. Also, each genre category has a high variety of video material, which makes training difficult. Finally, video content available on web video platforms is typically video reports, and differs from classic TV footage. Video material is usually assembled in a news broadcasting style, which means genre-specific content is inserted periodically into a dialogue or interview scene.

Classification perspective. For classification we used the Weka [49] environment, which provides a great selection of existing machine learning techniques. We tested methods ranging from simple Bayes to function-based, rule-based, lazy classifiers and tree approaches (from each category of methods, we selected the most representatives). Method parameters were tuned on the basis of preliminary experiments.

As the choice of training data may distort the accuracy of the results, we used a cross-validation approach. We split the data set into training and test sets, using values ranging from 10% to 90% for the percentage split. For part of the training data classification was repeated for all possible combinations between training and test sets in order to shuffle all sequences. Additionally, we tested different combinations of descriptors.

To assess performance, we used several measures. At the genre level, we computed the classic precision and recall ratios - see equation 6.5 (averaged over all experiments for a given percentage split). As a global measure, we computed F_{score} and average correct classification ($\overline{CD}$):

$$F_{score} = 2 \cdot \frac{P \cdot R}{P + R}, \quad \overline{CD} = \frac{\overline{N_{GD}}}{N_{total}} \tag{6.12}$$

where $\overline{N_{GD}}$ is the average number of correct classifications, and N_{total} is the number of test sequences.

The most accurate classification was obtained by using all audio-visual descriptors in combination. For reasons of brevity, we present only these results. Figure 6.7 shows the overall average F_{score} and average correct classification $\overline{CD}$ for a selection of seven machine learning techniques (those providing the most significant results).

The global results are very promising considering the high difficulty of this classification task. The highest average F_{score} is 46.3%, while the best average correct classification is 55% (out of 475 test sequences, 261 were correctly labeled, obtained for 80% training data). The most accurate classification technique proved to be an SVM with linear kernel, followed very closely by Functional Trees (FT), and then k-NN (with k=3), Random Forest trees, Radial Basis Function (RBF) Network, J48 decision tree, and finally Bayes Network.

The most interesting results, however, were obtained at genre level. Due to the

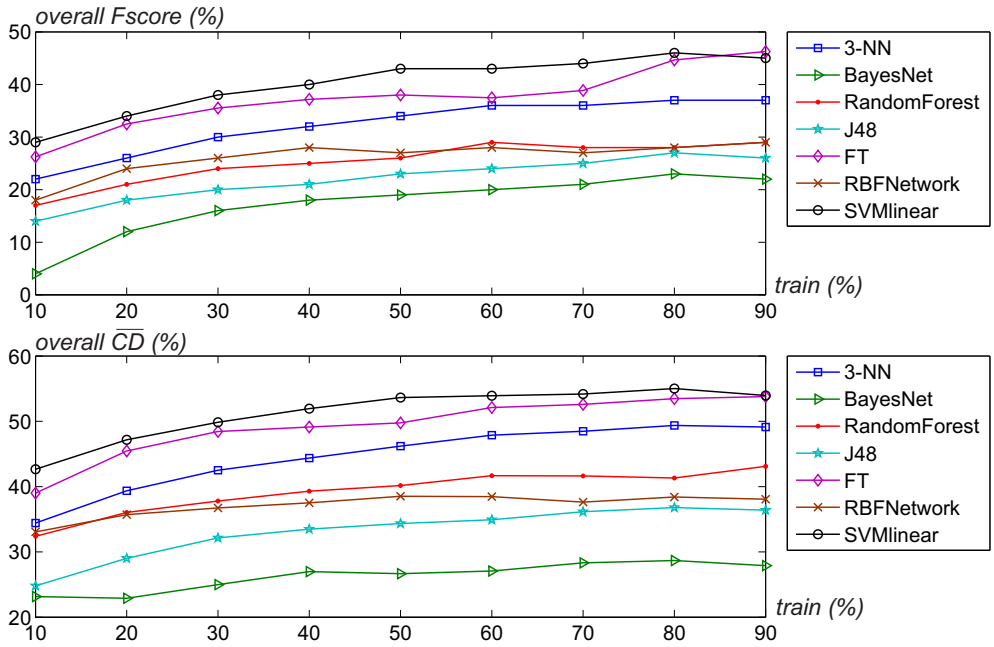


Figure 6.7: Overall average F_{score} and average correct classification $\overline{CD}$ achieved by various machine learning techniques using all audio-visual descriptors [4].

high semantic content, not all genres can be classified correctly with audio-visual information. We sought to determine which categories are better suited for this approach. Figure 6.8 shows the genre average F_{score} achieved by the linear SVM and FT trees.

The best performance was obtained for the following genres (we present results for a 50% percent split and give the highest value): "literature" ($F_{score} = 83\%$, highest 87%) and "politics" ($F_{score} = 81\%$, highest 84%), followed by "health" ($F_{score} = 78\%$, highest 85%), "citizen journalism" ($F_{score} = 65\%$, highest 68%), "food and drink" ($F_{score} = 62\%$, highest 77%), "web development and sites" ($F_{score} = 63\%$, highest 84%), "mainstream media" ($F_{score} = 63\%$, highest 74%), "travel" ($F_{score} = 57\%$, highest 60%), "technology" ($F_{score} = 53\%$, highest 56%). Less successful performance was achieved for genres such as "documentary" ($F_{score} = 7\%$ which is also the highest), "school" ($F_{score} = 10\%$, highest 22%) or "business" ($F_{score} = 9\%$, highest 14%).

Globally, classification performance increases with the amount of training data. However, for some genres, due to the large variety of video materials, increasing the number of examples may result in overtraining and thus in reduced classification performance. It can be seen in Figure 6.7 that classification performance decreases as the proportion of training data increases (e.g., SVM linear for 90% training data).

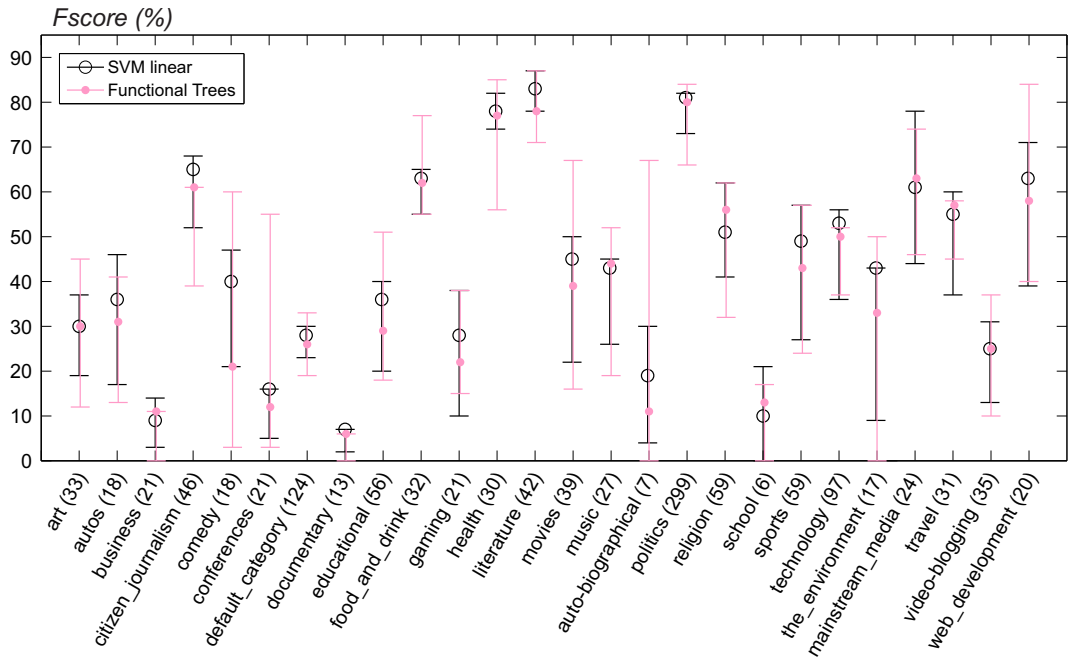


Figure 6.8: Average F_{score} for linear SVM and Functional Trees (FT) using all audio-visual descriptors and for a training-test data percentage split of 50% (numbers in brackets are the numbers of test sequences in each genre). Vertical lines indicate the min-max F_{score} intervals of each genre (percentage split ranging from 10% to 90%).

A clear difference between FT and SVM is visible at genre level. Globally, the SVM tends to perform better on a reduced training set, while the FT tends to be superior for higher amounts of training data (e.g., training data $> 70\%$, see min-max intervals in Figure 6.8).

Retrieval perspective. In this experiment, we assessed the classification performance of the proposed descriptors from the perspective of an information retrieval system. We present the results obtained for the 2011 MediaEval Video Genre Tagging Task [48]. The challenge was to develop a retrieval mechanism that works with all 26 genre categories. Each participant was provided with a development set consisting of 247 sequences, unequally distributed with respect to genre. Some genre categories were represented with very few (even just one or two) examples. This initial set was to serve as a reference point for developing the proposed solution. The participants were encouraged to build their own training sets if required by their approach. Consequently, to provide a consistent training data set for the classification task, we extended the data set to up to 648 sequences. Additional videos were retrieved from the same source (blip.tv), using genre-related keywords (we checked

for duplicates in the official development and test sets).

The final retrieval task was performed on a test set consisting of 1,727 sequences. In this case, the training-classification steps are to be performed only once. Up to 10 teams competed at this task, each one submitting up to 5 different runs (3 were restricted to using only textual descriptors extracted from speech transcripts, user tags, and metadata).

In our case, the retrieval results were obtained using a binary ranking in which the maximum relevance of 1 is associated with the genre category into which the document was classified, while other genres have 0 relevance.

To assess performance, we use the overall Mean Average Precision (MAP) as defined by TRECVid[6]:

$$MAP = \frac{1}{|Q|} \cdot \sum_{j=1}^{|Q|} \frac{1}{m_j} \cdot \sum_{k=1}^{m_j} P(R_{j,k}) \tag{6.13}$$

where $Q = \{q_1, ..., q_{|Q|}\}$ denotes a set of queries q_j which are represented in the data set by $\{d_1, ..., d_{m_j}\}$ relevant documents, R_{j_k} is the set of ranked retrieval results from the top result to document d_k, and $P()$ is the precision (see equation 6.5). When a relevant document is not retrieved at all, the precision value in the above equation is taken to be 0.

For classification we used the approach providing the most accurate results, namely the SVM with a linear kernel. In Table 6.4 we compare our results with several other approaches using various modalities of the video, from textual information (e.g., speech transcripts, user tags, metadata) to audio-visual.

The proposed descriptors achieved an overall MAP of up to 12% (see team RAF [57]). These were the best results obtained using audio-visual information alone. Use of descriptors such as cognitive information (face statistics), temporal information (average shot duration, distribution of shot lengths) [50], audio (MFCC, zero-crossing rate, signal energy), color (histograms, color moments, autocorrelogram - denoted autocorr.), and texture (co-occurrence - denoted co-occ., wavelet texture grid, edge histograms) with SVM resulted in a MAP of less than 1% (see team KIT [56]), while clustered SURF features and SVM achieved a MAP of up to 9.4% (see team TUB [58]). We achieved better performance even compared to some classic text-based approaches, such as the Term Frequency-Inverse Document Frequency (TF-IDF - MAP 9.8%, see team UAB [53]) and the Bag-of-Words (MAP 5.5%, see team SINAI [52]) approaches. Compared to visual information, audio descriptors seem to provide better discriminative power for this task.

[6]see trec_eval scoring tool at `http://trec.nist.gov/trec_eval/`.

Table 6.4: Comparative results: MediaEval benchmarking [48] (selective results).

descriptors	*modality*	*method*	*decision*	MAP	*team*
metadata	text	Negative multinomial diverg.	ranked list	39.37%	TUD [55]
clustered SURF, metadata	visual, text	Naive Bayes, SVM + serial fusion	binary	30.33%	TUB [58]
proposed	audio, visual	SVM with linear kernel	binary	12.08%	RAF [57]
speech transcripts	text	Support Vector Machines	ranked list	11.79%	LIA [51]
speech transcripts, metadata, user tags	text	Bag-of-Words + Terrier IR	ranked list	11.15%	SINAI [52]
speech transcripts, Delicious tags, metadata	text	BM25F + Kullback - Leibler diverg.	ranked list	11.11%	UNED [54]
proposed	audio	SVM with linear kernel	binary	10.29%	RAF [57]
clustered SURF	visual	Visual-Words + SVM with RBF kernel	binary	9.43%	TUB [58]
speech transcripts, metadata, user tags	text	TF-IDF + cosine dist.	binary	9.4%	UAB [53]
speech transcripts	text	Bag-of-Words	ranked list	5.47%	SINAI [52]
proposed	visual	SVM with linear kernel	binary	3.84%	RAF [57]
hist., moments, autocorr., co-occ., wavelet, edge hist.	visual	multiple SVMs	binary	0.35%	KIT [56]
structural (shot statistics)	visual	multiple SVMs	binary	0.3%	KIT [56]
color, texture, aural, cognitive, structural	audio, visual	multiple SVMs	binary	0.23%	KIT [56]
MFCC, zero cross. rate, signal energy	audio	multiple SVMs	binary	0.1%	KIT [56]
cognitive (face statistics)	visual	multiple SVMs	binary	0.1%	KIT [56]

The most efficient retrieval approach remains the inclusion of textual information, as it provides a higher semantic level of description than audio-visual information. The average MAP achieved by including textual descriptors is around 30% (e.g., see team TUB [58] in Table 6.4). Retrieval performance is boosted by including information such as movie names, movie ID from *blip.tv*, or the username of the video uploader; in this particular case, the reported MAP was up to 56% (which is also the highest obtained).

Conclusions and future work

We proposed four categories of content descriptors: block-level audio features, temporal based descriptors, color perceptual descriptors and statistics of contour geometry. These descriptors were used with several binary classification techniques to classify video footage into animated and non-animated content. We achieved very

promising results when using temporal structure and color descriptors together, namely an average precision and recall ratios up to 90% and 92%, respectively, and a global correct detection ratio up to 92%.

Another validation consisted in approaching a real-world video genre classification scenario, i.e., the categorization of 26 video genres from the blip.tv media platform. With a classification approach, the use of audio-visual information may be highly efficient in detecting particular genres, for instance, in our case "literature" (we obtain $F_{score} = 87\%$), "politics" ($F_{score} = 84\%$), and "health" ($F_{score} = 85\%$), and less successful for others, such as "school" ($F_{score} = 22\%$), and "business" ($F_{score} = 14\%$). One can envisage a classification system which adapts the choice of parameters to the target categories, for instance, using audio-visual descriptors for genres which are best detected with this information, using text for text-related categories, and so on. With a retrieval approach, the proposed descriptors achieved the best results of all audio-visual descriptors in the context of the 2011 MediaEval Video Genre Tagging Task [48]. They provided better retrieval performance than other descriptors such as cognitive information (face statistics), audio (MFCC, zero-crossing rate, signal energy), and excelled even compared to some classic text-based approaches, such as the Term Frequency-Inverse Document Frequency approach.

Future work on this subject should push forward descriptors to a higher semantic level, like exploiting human concept detection as well as more sophisticated fusion techniques.

6.3.2 Fisher kernel representation

In video, global features are often used for reasons of computational efficiency, where each global feature captures information of a single video frame. But frames in video change over time (e.g., due to motion, changes, etc) which is one of the video representative information. Therefore, one should search for meaningful approaches to capture that variation in time.

Another contribution related to content description is using Fisher Kernel representation to capture temporal changes in order to derive highly representative and efficient content descriptors[7].

[7]this work was developed in collaboration with Dr. Ionuţ Mironică, from LAPI, University Politehnica of Bucharest, Romania, Dr. Jasper Uijlings, Negar Rostamzadeh and Prof. Nicu Sebe, from MHUG, University of Trento, Italy. The presented results were published in:

[12] I. Mironică, J. Uijlings, N. Rostamzadeh, *B. Ionescu*, N. Sebe, "*Time Matters! Capturing Variation in Time in Video using Fisher Kernels*", ACM Multimedia, 21-25 October, Barcelona, Spain, 2013.

Contribution to state-of-the-art

Bag of Local Visual Features [61], which are currently some of the most popular and efficient description approaches, capture the visual variation in space for images and in both space and time for video. The proposed Fisher Kernel approach [12] improves over the common k-means vocabulary by modelling the distribution of features within each visual word. In contrast to Local Visual Features, Fisher representation is applied at frame-based features, effectively capturing variation in time only (as there is no variation in space). Like Bag of Local Visual Features, all ordering is lost but all variation is captured. Using the Fisher representation for modelling variation in time, dissimilar frames will be represented by different mixture components (i.e., clusters), preventing blending of unrelated features while enabling them to co-exist in a single representation. This enables representing videos which consist of dissimilar parts (which may not even have a fixed temporal order) such as news broadcasts that switch between the news-anchor and on-site footage. Furthermore, similar frames that fall in the same mixture component will be modelled with respect to the general distribution of that component, capturing subtle variations in time such as the different appearances of a person walking by.

The proposed description framework is general enough in terms of descriptor use and applicability so that is able to address a broad range of applications, e.g., genre-recognition, sports-recognition, daily activity recognition; all of these while keeping a significantly reduced size of the descriptor compared to similar approaches and a performance close to real-time.

Algorithm

The Fisher Kernel [59] represents a signal as the gradient with respect to the probability density function that is a learned generative model of that signal. Recently, [60] introduced the Fisher Kernel as an improved visual vocabulary for Bag-of-Words. Its success shows that it meaningfully captures the visual variation of local descriptors.

We follow [60] and use a Gaussian Mixture Model with diagonal covariance matrices as generative distribution. Specifically, let μ_i and σ_i be the mean and standard deviation of the i-th Gaussian centroid, let $\gamma(i)$ be the soft assignment to the i-th Gaussian of the d-dimensional feature x_t captured at frame t. The gradient of the GMM with respect to μ_i and σ_i are calculated as [60]:

$$\mathcal{G}^x_{\mu,i} = \frac{1}{T\sqrt{\omega_i}} \sum_{t=1}^{T} \gamma(i) \frac{x_t - \mu_i}{\sigma_i} \tag{6.14}$$

$$\mathcal{G}^x_{\sigma,i} = \frac{1}{T\sqrt{2\omega_i}} \sum_{t=1}^{T} \gamma(i) \left[\frac{(x_t - \mu_i)^2}{{\sigma_i}^2} - 1 \right]. \quad (6.15)$$

The final Fisher descriptor is obtained by the concatenation of the $\mathcal{G}^x_{\mu,i}$ and $\mathcal{G}^x_{\sigma,i}$ for $i = 1, ..., k$ and has a dimensionality of $2kd$. Interpreting the formulas in terms of variation in time, equation 6.14 averages related features over time, which are related as they fall in the same mixture component. Equation 6.15 models the variation of related features over the video sequence, capturing subtle visual changes (e.g., a car driving by). The different mixture components capture drastic variations in time such as a shot changes.

Validation results

Validation of this content representation scheme was carried out for three different scenarios, namely: video genre classification, human action recognition and recognition of daily activities. We normalise the Fisher vector by taking the square root followed by the $L2$-norm [60]. In contrast to [60], for classification we use Support Vector Machines (SVM) with RBF-kernels as these performed better than linear SVMs, even at an increased number of clusters for the latter. When combining different types of features we use weighted late fusion, learning weights on our optimization sets (see equation 6.16).

Video genre classification. For video genre classification, experiments were conducted on the 2012 MediaEval Genre Tagging Task [62], consisting of 2,000 hours in 14,838 videos, labelled according to 26 genres such as "art", "autos", and "comedy" (see the complete list in Section 6.3.3 with the Experimental results). Performance is measured in terms of Mean Average Precision (MAP) as defined in equation 6.13. We perform all parameter optimization on the training set which we split in two fixed, equally sized parts. We compare with the state-of-the-art using the official training set (5,288 videos) and test set (9,550 videos), as used with the task.

The Fisher representation is computed using the following type of descriptors [12]: global Histogram of Oriented Gradients (HoG, 81 dimensions) which calculates HoG over the whole frame using a 3x3 spatial division, Colour Naming histogram (CN, 11 dimensions) of the whole frame, audio features (98 dimensions) which are general purpose audio descriptors extracted over a standard period of 1.28 seconds around the frame. Results of averaging features over the whole video are presented as the horizontal lines in Figure 6.9.

We determine the optimal number of clusters for each feature as shown in Figure 6.9. First of all, one can observe the important improvement of the Fisher representation over the baseline which simply averages the features: even when us-

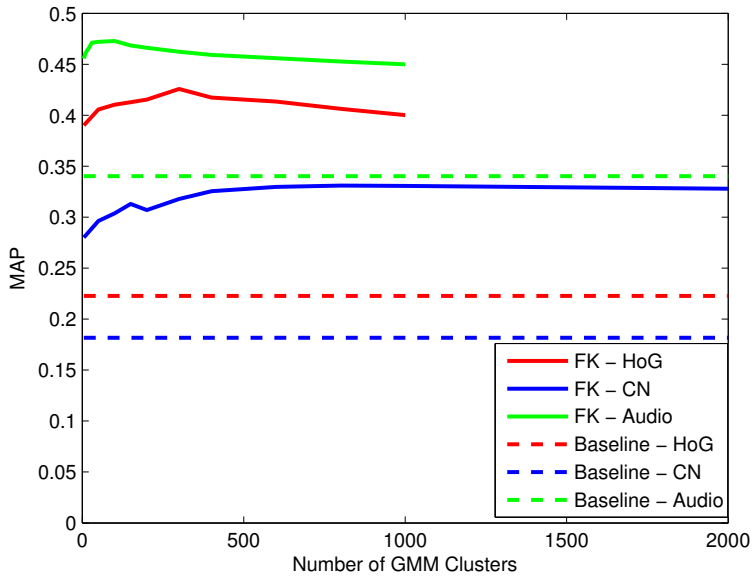


Figure 6.9: Mean Average Precision (MAP) while varying the number of cluster centres on the MediaEval 2012 training set [62][12].

ing only a single centroid, CN achieves an improvement from 18% MAP to 28% MAP, HoG from 22% MAP to 38% MAP, and Audio from 34% MAP to 45% MAP. The modelling of variation in time therefore significantly improves results. Increasing the number of clusters increases performance even further: both CN and HoG increase an extra 5% MAP, reaching 33% MAP and 43% MAP at 800 clusters and 200 clusters respectively. Audio features increase to 47% MAP at 50 clusters. We will use this number of clusters in the next experiment. The final sizes of the Fisher vectors are reasonable at 17,600 for CN, 42,000 for HoG, and 9,000 for audio features.

Results on 2012 MediaEval Genre Tagging Task [62] are shown in Table 6.5. For audio features our results are at 47% MAP much better than the best result of 19% reported by team ARF [63]. For visual features only, at 46% MAP we perform significantly better than the best result of 35% MAP reported by team KIT [64]. Remarkably, our combination of audio and visual features yields with 55% MAP a better performance than the use of text from automatic speech recognition and meta-data, which had the highest performance with 53% MAP (TUB [65]).

Human action recognition. For the second validation, we use the UCF50 Human Action Recognition dataset [69], which contains 6,600 realistic videos from YouTube with large variations in camera motion, object appearance and pose, illumination conditions, scale, etc. It has 50 mutual exclusive categories such as

Table 6.5: Comparison with State-of-the-Art in terms of Mean Average Precision (MAP) on MediaEval Genre Tagging [62][12].

Feature type	*method*	MAP (reported)	MAP (**our method**)
Audio	Block Based Audio Features and 5-NN ARF [63]	19.2%	47.5%
Visual	Visual descriptors (Color, Texture, rgbSIFT) KIT [64]	35%	46%
Audio and Visual	-	-	55%
Metadata and Text ASR	BoW Text ASR and metadata TUB [65]	52.3%	-
Notations: SIFT - Scale Invariant Features Transform, BoW - Bag-of-Words, ASR - text extracted with automatic speech recognition, KNN - K Nearest Neighbors			

"biking", "diving", "drumming" and "fencing". Performance is evaluated in terms of classification accuracy (i.e., the percentage of the items correctly classified from the total number). We perform all optimization on half of the dataset, using 8-fold cross-validation. We compare with the state-of-the-art using the standard leave-one-group-out cross-validation on the full dataset [69].

The Fisher representation is computed using the following type of descriptors [12]: global Histogram of Oriented Gradients (HoG, with 9, 36, 81, and 144 dimensions) which calculates HoG over the whole frame using a 1x1, 2x2, 3x3, and 4x4 spatial division, global Histogram of Optical Flow (HoF, with 9, 36, 81, and 144 dimensions) which measures the average velocity of non-stationary pixels over a region in 9 orientations. We use a 1x1, 2x2, 3x3, and 4x4 spatial division; Colour Naming histogram (CN with 11, 44, 99, and 176 dimensions) using a 1x1, 2x2, 3x3, and 4x4 spatial division. In all experiments, we combine different spatial divisions for a single feature type using late fusion with equal weights. Results of averaging each feature over the whole video are shown as horizontal lines in Figure 6.10.

In Figure 6.10 we evaluate the performance with respect to the number of GMM clusters, where we use the same number of clusters for all spatial divisions of a single feature type. For CN and HoG the use of a single cluster improves the baseline with 6% and 5% respectively. More, clusters degrade performance as for this dataset the visual changes are subtle and do not require different mixture components. For HoF, using 50 clusters improves the baseline of 54% to 67%, a 13% improvement. Hence the optical flow changes drastically in time which is best captured in multiple clusters. Indeed, for example a baseball pitch has at least three distinct movement patterns: static (before the action), the pitch, and the batting. In the next experi-

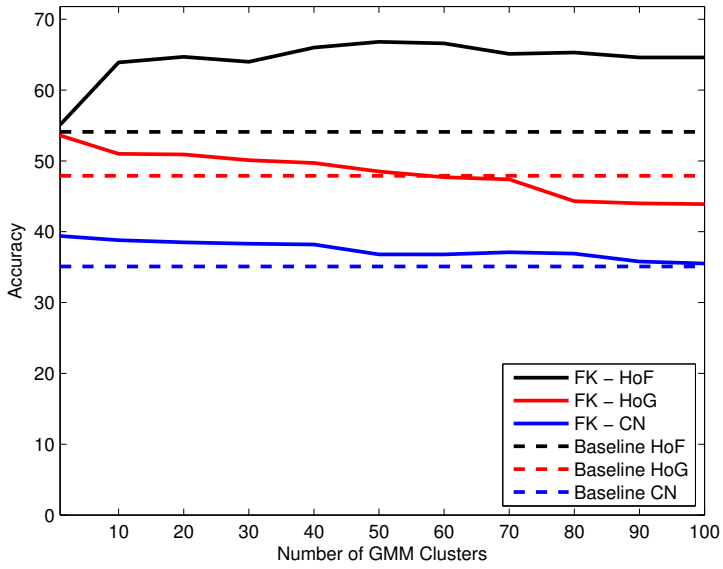


Figure 6.10: Classification accuracy on half of UCF50 sports [69] while varying the number of cluster centres (8-fold cross-validation) [12].

Table 6.6: Comparison with State-of-the-art on UCF50 dataset [69].

Method	*Accuracy*
Reddy et al. [69]	76.9%
proposed	74.7%
Solmaz et al. [71]	73.7%
Everts et al. [72]	72.9%
Kliper-Gross et al. [70]	72.6%
Solmaz et al. [71]: GIST3D	65.3%

ment we use 1 cluster for CN and HoG, and 50 clusters for HoF.

We present the state-of-the-art in Table 6.6. As can be seen, the proposed approach rank second with 74.7% accuracy after the 76.9% accuracy of [69]. However, we use only global features whereas all other good performing methods use computationally more expensive Space-Time Interest Points (STIPs). Only the GIST3D entry of [70] does not use STIPs. They use global, frame-based features plus linear quantization. Our performance using the Fisher vector is a significant 9.4% higher.

Daily activities recognition. The final validation was carried out on the Daily Activity Recognition - ADL dataset [73], consisting of ten human activities such as dialling a phone, peeling banana, and chopping banana. Each activity is performed

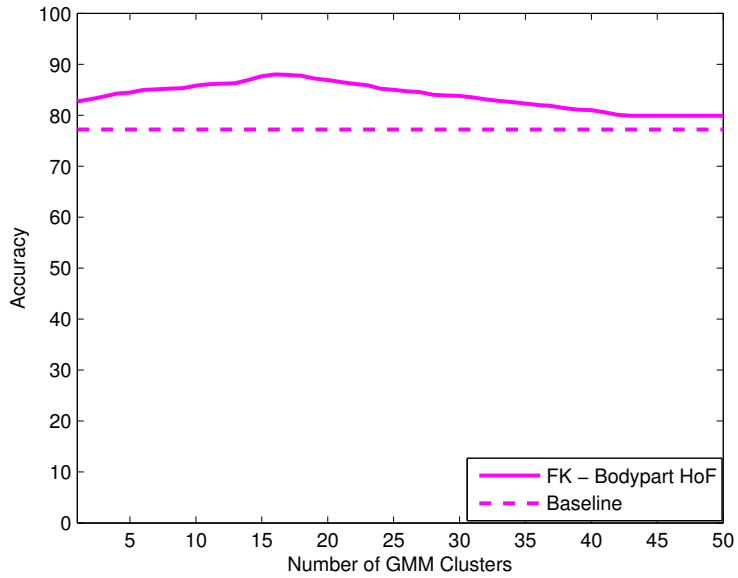


Figure 6.11: Classification accuracy on ADL daily activity [73] recognition on half the dataset while varying the number of cluster centres [12].

Table 6.7: Comparison with state-of-the-art on ADL dataset [73].

Method	*Accuracy*
proposed	97.3%
Wang et al. [76]	96%
Lin et al. [77]	95%
Messing et al. [73]	89%

three times by five people, totalling 150 videos. Performance is measured in accuracy. We do all optimization on half of the dataset and report final results on the full dataset. In both cases we use leave-one-person-out cross-validation [73].

As human pose and body-part motion are important for distinguishing the different categories, for content description we extract body-part features [74]. We use the state-of-the-art body-part detector of [75] and extract at every frame for all 18 body-parts a Histogram of Optical Flow in 8 orientations (144 dimensions). The result of averaging this feature over the video is shown as the horizontal line in Figure 6.11.

Figure 6.11 shows accuracy with respect to the number of GMM clusters. Using only a single cluster yields a performance improvement from 77% to 82% accuracy. The best accuracy of 88% is obtained using 17 clusters. Note that the number of

clusters is relatively low, likely due to the smaller dataset. At 17 clusters, the final feature has 4,896 dimensions. We use 17 clusters when testing on the full dataset.

We compare our work with other approaches in Table 6.7. As can be seen, the proposed approach yields the highest accuracy of 97.3%. This shows that the Fisher representation is also effective for modelling variation in time using local body-part features.

Conclusions and future work

We proposed a scheme for using Fisher Kernel theory to model variation in time for frame-based video features. While the temporal order is lost, the temporal variation is captured at two levels: similar features are grouped together while retaining variation, which enables capturing subtle variations over time such as a exhibited by a moving car. Dissimilar features are kept separate, preventing mixing features from unrelated parts of the video while keeping them in a single representation, which enables capturing different shots in a video.

We demonstrated that our framework is highly general: it allowed significant improvements on a wide variety of features, ranging from global visual features, to body-part features, and to audio features. We also demonstrated that our method works on a wide variety of datasets: we obtained state-of-the-art performance on UCF50 using global features instead of the more complex STIPs used in other methods. We improved the state-of-the-art on ADL daily activity recognition. We significantly improved the state-of-the-art on the 2012 MediaEval Genre Tagging task.

In future work we plan to model variation in time using Fisher kernels on more advanced features such as STIPs.

6.3.3 Multimodal fusion

The essence of multimedia is in taking advantage of the information provided with different modalities. As presented in the previous sections, dealing with content-based retrieval requires handling multi-modal information such as visual, audio and text. A dedicated domain is with the *fusion* techniques for aggregating different information sources thus to benefit from the advantages of each of them while reducing the redundant information.

Currently, there are two fundamentally different directions: *early fusion* techniques that aggregate content descriptors in the early stage of the processing (the fused descriptors are the concatenation of all the descriptors) and *late fusion* techniques that perform the fusion later in the processing chain, e.g., fusing the outputs of some single-modal classification systems. At the moment, there is no clear

supremacy of one approach over the other, both approaches providing state-of-the-art results in rather different situations. Therefore, fusion should be adapted to the specificity of the task.

The following contribution is in the investigation of the use of various state-of-the-art content descriptors extracted from different modalities and the efficiency of early and late fusion mechanisms in the content-based retrieval paradigm[8].

Contribution to state-of-the-art

We have conducted an in-depth evaluation and analysis [14] of the performance of multimodal video descriptors and early/late fusion schemes with the objective of evaluating an adequate combination of various modalities for achieving highly accurate classification. As case study, we experimented with the video genre classification task. We identify the following contributions of this work: providing an in-depth evaluation of truly multimodal video description (automated and user-generated text, audio and visual) in the context of a real-world genre-categorization scenario which can serve as guideline for other practitioners in the field; demonstrating the potential of appropriate late fusion techniques and achieving very high categorization performance; demonstrating that notwithstanding the superiority of user text based descriptors, late fusion can boost performance of automated content descriptors to achieve close performance.

Approach

To perform early fusion, descriptors are combined before the classification. The combination takes place in the feature space, namely the features are concatenated into one vector. The major drawback in this case is the dimensionality of the resulting feature space that is basically the sum of all the concatenated dimensions. High-dimensional spaces tend to scatter the homogeneous clusters of instances belonging to the same concepts reducing the performance.

In contrast, late fusion combines the confidence values from classifiers run on different descriptors. In our scenario, a classifier is supposed to provide some relevance scores indicating the probability of each classified item of belonging to some

[8]this work was developed in collaboration with Dr. Ionuţ Mironică, from LAPI, University Politehnica of Bucharest, Romania, Assist. Prof. Peter Knees, from Department of Computational Perception, Johannes Kepler University, Linz-Austria, and Prof. Patrick Lambert, from LISTIC, Université de Savoie, Annecy-France. The presented results were published in:

[14] I. Mironică, *B. Ionescu*, P. Knees, P. Lambert, "*An In-Depth Evaluation of Multimodal Video Genre Categorization*", IEEE International Workshop on Content-Based Multimedia Indexing - CBMI 2013, 17-19 June, Veszprém, Hungary, 2013.

class. Naturally, each of the classifiers will tend to provide different scores for each class. Late fusion involves the design of an aggregated classifier combination function, $f(x_1, ..., x_N)$, with x_i the relevance output of the classifier i, whose result is better than any of its individual classifiers and as good as possible. The aggregation is carried out for each individual class. To achieve the final categorization, videos are then sorted according to the new aggregated scoring formula results. Late fusion focuses on the individual strength of modalities, whereas early fusion uses the correlation of features in the mixed feature space.

For our study, we have selected several popular late fusion approaches. For each item and feature pair, classification yields C confidence scores, one for each of the target classes. Simple linear combination represents a weighted average of the multimodal confidence values of each of the considered classifiers:

$$CombMean(d, q) = \sum_{i=1}^{N} \alpha_i \cdot cv_i \tag{6.16}$$

where cv_i is the confidence value of classifier i for class q ($q \in \{1, ..., C\}$), d is the current item, α_i are some weights and N is the number of classifiers to be aggregated. In case of considering equal weights, $\alpha_1 =, ..., = \alpha_N$, this is referred to as CombSum.

An extension of CombMean can be obtained by giving more importance to the items that are more likely to be relevant for current concepts, which leads to:

$$CombMNZ(d, q) = F(d)^{\gamma} \cdot \sum_{i=1}^{N} \alpha_i \cdot cv_i \tag{6.17}$$

where $F(d)$ is the number of classifiers for which item d appears in the top K of the retrieved items and $\gamma \in [0, 1]$ is a parameter.

Finally, another useful perspective is to consider the rank of each confidence level. The score-based late fusion strategies require a normalization among all confidence values in order to balance the importance of each of them, which is not the case of the rank-based strategies. In our scenario, we use a common method for rank-based fusion, that is Borda Count. The item with the highest rank on each rank-list gets n votes, where n is the size of the dataset:

$$CombRank(d, q) = \sum_{i=1}^{N} \alpha_i \cdot rank(cv_i) \tag{6.18}$$

where $rank()$ represents the rank of classifier i, α_i are some weights and N is the number of classifiers to be aggregated.

Experimental results

Experimentation was conducted in the context of the 2012 MediaEval Genre Tagging Task [62]. The data set consists of up to 14,838 blip.tv videos that are divided into a training set of 5,288 videos (36%) and a test set of 9,550 movies (64%; we use the same scenario as for the official benchmark). Videos are labeled according to 26 video genre categories specific to the blip.tv media platform, namely (the numbers in brackets are the total number of videos): "art" (530), "autos and vehicles" (21), "business" (281), "citizen journalism" (401), "comedy" (515), "conferences and other events" (247), "documentary" (353), "educational" (957), "food and drink" (261), "gaming" (401), "health" (268), "literature" (222), "movies and television" (868), "music and entertainment" (1148), "personal or auto-biographical" (165), "politics" (1107), "religion" (868), "school and education" (171), "sports" (672), "technology" (1343), "environment" (188), "mainstream media" (324), "travel" (175), "video blogging" (887), "web development" (116) and "default category" (2349, comprises movies that cannot be assigned to any of the previous categories). The main challenge of this scenario is in the high diversity of genres, as well as in the high variety of visual contents within each genre category.

For content description we use a broad range of state-of-the-art descriptors [14]:

- *aural information*: standard audio features (196 values) - a common set of general-purpose audio descriptors, namely: Linear Predictive Coefficients, Line Spectral Pairs, MFCCs, Zero-Crossing Rate, spectral centroid, flux, rolloff and kurtosis, augmented with the variance of each feature over a certain window (a common setup for capturing enough local context is taking 1.28 s). Video temporal integration is achieved by taking the mean and standard deviation of these descriptors over all frames;
- *visual information*:
  - MPEG-7 related descriptors (1,009 values) - describe the global color and texture information over all the frames. We selected the following representative descriptors: Local Binary Pattern, autocorrelogram, Color Coherence Vector, Color Layout Pattern, Edge Histogram, Scalable Color Descriptor, classic color histogram and color moments [38]. For each sequence, we aggregate the features by taking the mean, dispersion, skewness, kurtosis, median and root mean square statistics over all frames;
  - structural descriptors (1,430 values) - account for contour information and relation between contours. We use the approach in [39] to parameterize the geometry and appearance of contours and regions;

- global Histograms of oriented Gradients (HoG - 81 values) - represent the average of the well known HoG features. For the entire sequence we compute the average histogram over all frames;
- Bag-of-VisualWords of SIFT descriptors (20,480 values) - we compute a Bag-of-VisualWords (B-o-VW) model over a selection of key frames (uniformly sampled). For this task, we extract a visual vocabulary of 4,096 words. The keypoints are extracted with a dense sampling strategy and described using rgbSIFT features [78]. Descriptors are extracted at two different spatial scales of a spatial pyramidal image representation (entire image and quadrants).

- *textual information*: we adapted a classic Term Frequency-Inverse Document Frequency (TF-IDF) approach. First, we filter the input text by removing the terms with a document frequency less than 5%-percentile of the frequency distribution. We reduce further the term space by keeping only those terms that discriminate best between genres according to the χ^2-test. We generate a global list by retaining for each genre class, the m terms with the highest χ^2 values that occur more frequently than in complement classes. This results in a vector representation for each video sequence that is subsequently cosine normalized to remove the influence of the length of text data. We consider following TF-IDF descriptors:
  - TF-IDF of ASR data (3,466 values, $m = 150$) - describes textual data obtained from Automatic Speech Recognition of the audio signal. For ASR we use the transcripts provided with the dataset;
  - TF-IDF of metadata (504 values, $m = 20$) - describes textual data obtained from user metadata such as synopsis, user tags, video title, information that typically accompanies videos posted on the blip.tv platform.

For classification, we have selected five of the most popular approaches that proved to provide high performance in various information retrieval tasks, namely Support Vector Machines (SVM, with various kernel functions: linear, Chi-square - CHI, Radial Basis Functions - RBF), k-Nearest Neighbor (k-NN), Random Trees (RT) and Extremely Random Forest (ERF).

To assess performance, we report the standard Mean Average Precision (MAP), as defined in equation 6.13.

Performance assessment of individual modalities. The first experiment consisted on assessing the discriminative power of each individual modality and group

Table 6.8: Classification performance of individual modalities (MAP) [14].

Descriptors	*SVM Linear*	*SVM RBF*	*SVM CHI*	*5-NN*	*RF*	*ERF*
HoG	9.08 %	**25.63%**	22.44%	17.92%	16.62%	23.44%
Bag-of-Visual-Words rgbSIFT	14.63 %	17.61%	**19.96%**	8.55%	14.89%	16.32%
MPEG-7	6.12 %	4.26%	17.49%	9.61%	20.90%	**26.17%**
Structural descriptors	7.55 %	17.17%	**22.76%**	8.65%	13.85%	14.85%
Standard audio descriptors	20.68 %	24.52%	35.56%	18.31%	34.41%	**42.33%**
TF-IDF of ASR	32.96 %	**35.05%**	28.85%	12.96%	30.56%	27.93%
TF-IDF of metadata	56.33 %	58.14%	47.95%	57.19%	**58.66%**	57.52%

of descriptors. Table 6.8 summarizes some of the results (the best performance per modality is highlighted in bold).

The highest performance for visual information is achieved using MPEG-7 related descriptors and Extremely Random Forest (ERF) classifiers, MAP 26.17%, followed closely by HoG histograms on SVM and RBF kernel, MAP 25.63%. Surprisingly, Bag-of-Visual-Words representation of feature information (rgbSIFT) is not performing efficiently to this task, MAP is below 20%. The audio descriptors are able to provide a significantly higher discriminative power, the highest MAP of 42.33% being achieved with the ERF classifier.

In what concerns the text modality, the use of metadata and Random Forest classifiers led to the highest MAP of 58.66% which is an improvement of more than 16% over the audio. The use of ASR data alone is able only to provide a MAP up to 35.05% (with SVM and RBF kernel), which is less discriminative than using audio descriptors. Therefore, video descriptors can outperform at this point the automated text descriptors. This is mostly due to the fact that ASR data is extracted automatically, being inherently subject to errors (e.g., due to noise).

Performance of multimodal integration. Fusion techniques tend to exploit complementarity among different information sources. In this experiment we assess the performance of various combination of modalities as well as of different fusion strategies, from late fusion schemes to the simple concatenation of different descriptors (i.e., early fusion).

For late fusion, weights (i.e., α_k and $F(d)$ values) are first estimated on the training set and tuned for best performance. To avoid overfitting, half of the training set is used for training and the other half for parameter evaluation. The actual classification is then carried out on the test set. MAP is reported in Table 6.9 (highest values per feature type are presented in bold).

In all of the cases, late fusion tends to provide better performance than early

Table 6.9: Performance of multimodal integration (MAP) [14].

Descriptors	*Comb SUM*	*Comb Mean*	*Comb MNZ*	*Comb Rank*	*Early Fusion*
all visual	35.82%	36.76%	**38.21%**	30.90%	30.11%
all audio	43.86%	44.19%	**44.50%**	41.81%	42.33%
all text	62.62%	**62.81%**	62.69%	50.60%	55.68%
all	64.24%	65.61%	**65.82%**	53.84%	60.12%

fusion. Using only the visual descriptors the improvement is of more then 8% over simple descriptor concatenation (highest MAP is 38.21% using CombSUM). For audio descriptors, highest MAP of 44.5% is achieved with CombMNZ, that is an improvement of more than 2% over the simple use of all descriptors together. Audio still provides significant superior discriminative power than using only visual.

A significant improvement of performance is also achieved for textual descriptors. We obtain the highest MAP score with CombMean, namely 62.81%, which is an improvement of over 7% compared to early fusion. Although the simple concatenation of modalities manages to boost classification performance up to a MAP of 60.12%, late fusion is able to exploit better the complementarity between descriptors, achieving more than 5% of improvement. In what concerns the late fusion techniques, CombRank tends to provide the least accurate results in most of the cases, while the other approaches tend to provide more or less similar results.

Therefore, in most of the cases late fusion proves to be a better choice for multimodal genre classification. Firstly, it provides significantly higher performance than early fusion. Secondly, late fusion is also less computational expensive than early fusion, because the descriptors used for each of the classifiers are shorter than using the concatenation of all features. Finally, late fusion systems scale up easier because no re-training is necessary if further streams or modalities are to be integrated.

Comparison to state-of-the-art. The final experiment consisted on comparing the late fusion strategies against other methods from the literature. As reference, we use the best team runs reported to 2012 MediaEval Video Genre Tagging Task [62]. Results are presented in Table 6.10 by decreasing MAP values.

The most efficient modality remains the exploitation of textual information as it provides a higher semantic level of description than audio-visual information. In particular, the use of metadata proves to be the most efficient approach leading to the highest MAP at MediaEval 2012, 52.25% (see team TUB [65]). In spite of this high classification rate, late fusion still allows for significant improvement, for instance CombMean on ASR and metadata achieves a MAP up to 62.81% - that is an improvement of more than 10% over the best run at MediaEval 2012 and of around 25% over using the same combination of textual descriptors (ARF [63]).

Table 6.10: Comparison with 2012 MediaEval Video Genre Tagging best runs [62]

Team	*Modality*	*Method*	MAP
proposed	all	Late Fusion CombMNZ with all descriptors	**65.82**%
proposed	text	Late Fusion CombMean with TF-IDF of ASR and metadata	**62.81**%
TUB [65]	text	Naive Bayes with Bag of Words on text (metadata)	52.25%
proposed	all	Late Fusion CombMNZ with all descriptors except for metadata	**51.9**%
proposed	audio	Late Fusion CombMean with standard audio descriptors	**44.50**%
proposed	visual	Late Fusion CombMean with MPEG-7 related, structural, HoG and B-o-VW with rgbSIFT	**38.21**%
ARF [63]	text	SVM linear on early fusion of TF-IDF of ASR and metadata	37.93%
TUD [68]	visual & text	Late Fusion of SVM with B-o-W (visual word, ASR & metadata)	36.75%
KIT [64]	visual	SVM with Visual descriptors (color, texture, B-o-VW with rgbSIFT)	35.81%
TUD-MM [66]	text	Dynamic Bayesian networks on text (ASR & metadata)	25.00%
UNICAMP - UFMG [67]	visual	Late fusion (KNN, Naive Bayes, SVM, Random Forests) with BOW (text ASR)	21.12%
ARF [63]	audio	SVM linear with block-based audio features	18.92%
Notations: SIFT - Scale Invariant Features Transform, TF-IDF - Term Frequency-Inverse Document Frequency, BoW - Bag-of-Words, ASR - text extracted with automatic speech recognition.			

In what concerns the visual modality, best MAP at MediaEval 2012 is up to 35% (see team KIT [64]) and is obtained using a combination of classical color/texture descriptors (e.g., HSV color histogram, L*a*b* color moments, autocorrelogram, concurrence texture, wavelet texture grid and edge histograms) and B-o-VW of rgbSIFT descriptors. Results show that using only B-o-VW of feature descriptors (e.g., SIFT, SURF - Speeded-up Robust Features), in spite of their reported high performance in many retrieval tasks, is not that accurate, e.g., MAP 23.29% using SIFT, 23.01% with SURF-PCA. The CombMean late fusion of visual descriptors provides an improvement over the best run of more than 3% (MAP 38.21%).

Using only audio information, best reported run at MediaEval 2012 achieves a MAP of 18.92% (see team ARF [63]). In this case CombMean late fusion of audio descriptors provides an improvement of more than 25% (MAP 44.5%).

Combining all the descriptors with CombMNZ we achieve a very high classification accuracy as MAP is up to 65.82%, that is an improvement of more than 13% over the MediaEval 2012 best run. In spite of the high discriminative power of textual descriptors, the combination of all the modalities with late fusion is able to exploit data complementarity at some level as the improvement over using only textual information is of 3%. This is a significant achievement considering the scale of the data set.

From the modality point of view, metadata provides the highest discriminative power for genre categorization. However, one should note that this information is user generated (e.g., includes document title, tags and user comments and descriptions) and cannot be determined automatically from the video information, that limits its applicability in real-time categorization scenarios. Approaching the classification using only content information that can be computed automatically from video data (ASR and audio-visual descriptors), late fusion is still able to provide high classification performance leading to a MAP of 51.9%, surpassing even some metadata-based approaches, e.g., see team ARF [63] and TUD-MM [66].

Conclusions and future work

We studied the contribution of various modalities and the role of the fusion mechanisms in increasing the accuracy of the classification results. The study was carried out in a real-world scenario using 26 blip.tv web video categories and more than 3,200 hours of video footage. The design of appropriate descriptors and late fusion integration allows to achieve a MAP up to 65.8%, that is a significant improvement of more than 13% over the best approach reported at the 2012 MediaEval Genre Tagging Task [62]. We prove that notwithstanding the superiority of employing user-generated textual information (e.g., user tags, metadata), the proposed multimodal integration allows to boost performance of automated content descriptors to achieve close performance. Future work will mainly consist in exploring spatio-temporal data representation in this context.

6.4 Video summarization

Video information is massive spatio-temporal data, as just one minute contains up to 1,800 static images, whereas video databases contain as much as millions of recordings. Browsing the database in the search of a specific movie or a particular scene, can be a tedious task, as it is necessary to visualize the whole content of the movie. Visualizing each movie is, firstly inefficient, due to data redundancy, and secondly, can be very time consuming, as it may take months to process all the video footage. An efficient solution for addressing this issue are *movie abstracts*. A movie abstract is a compact representation of the original video, significantly shorter, which preserves most of the essential parts of the original video [79].

There are two fundamentally different types of video abstracts. The still-image abstracts, known as *video summaries* (sometimes called static storyboards), are a small collection of salient images (i.e., key frames) that best represent all the underlying content at different levels of details. The moving-image abstracts, or *video*

skims, consist of a collection of image sequences. The existing video skimming techniques address two different approaches: *summary sequences*, which are classic abstracts covering the entire movie's content, and *movie highlights*, which only summarize some of the most interesting parts of the movie. Video highlighting techniques are related to the characteristics of the events, which should be considered as representative for the underlying movie content [80].

Both video abstraction techniques are useful and have been used extensively with content-based video indexing systems for reducing browsing time, for improving the quality of the search as well as reducing the computational complexity by replacing the original movie in the processing steps. They are efficient in almost complementary situations: static abstracts, or video summaries, are easy to compute (they contain only visual information), the computational complexity can be greatly reduced (e.g., a quick summary can be produced by retaining one image per shot) and are easy to visualize (there is no need to synchronize data); on the other hand, the possibly higher computational effort during the skimming process pays off during the playback time as video skims make more sense providing the dynamic/motion content.

I have contributed to the development of both types of abstracts, namely techniques for trailer-like video highlights and techniques for adaptive summarization which tends to follow the storyboard of the movie[9]. These techniques were adapted to the specificity of animated movies, which pose different processing challenges compared to natural movies (a discussion is presented in Section 6.2.3).

6.4.1 Video storyboard

The first contribution in this area was to the development of techniques for video summarization thus to create an image summary that follows the exact narration of the video in a storyboard manner [25].

Contribution to state-of-the-art

The proposed storyboard-like summary is a collection of key frames which are retrieved at shot level. To capture the shot visual activity, we take advantage of the

[9]this work was developed in collaboration with Dr. Laurent Ott, Prof. Patrick Lambert, Prof. Didier Coquin, from LISTIC, Université de Savoie, Annecy-France, Dr. Alexandra Păcureanu, Prof. Vasile Buzuloiu†, from LAPI, University Politehnica of Bucharest, Romania. The presented results were published in:

[25] *B. Ionescu*, L. Ott, P. Lambert, D. Coquin, A. Pacureanu, V. Buzuloiu, "*Tackling Action - Based Video Abstraction of Animated Movies for Video Browsing*", SPIE - Journal of Electronic Imaging, 19(3), 2010.

particularity of the animated movies of sharing specific color palettes and we use histograms of cumulative inter-frame distances. A variable number of key frames are extracted according to a histogram pattern and thus to visual contents. In this way, we obtain a key frame for each different scene, not only at shot level, but at intra-shot level.

Approach

The proposed video summary aims to present one image for each individual movie scene, in a storyboard manner. The key frames are extracted at shot level and the number of key frames is adapted to the variability of the shot visual content. Basically, we attempt to extract one key frame for each group of similar content images. To do so, we inspired from color median filtering techniques, in which the output of the filter is the most representative value, thus the one which minimizes the cumulative sum of distances to all the other values. To capture the pattern of visual changes, we estimate a histogram of cumulative inter-frame distances.

For each retained frame (typically we use a temporal sub-sampling) of index i, from the current shot k, we compute its color histogram, denoted $H^i_{shot_k}(c)$, in which c is the color index, $c \in \{1, ..., 125\}$ in our experiments. To evaluate the distance between frames, we use the classical Manhattan distance, denoted $d_M()$, which provides a good compromise between the computational complexity and the quality of the results. We use a version of this distance which is normalized to 1:

$$d_M(H^i_{shot_k}, H^j_{shot_k}) = \frac{\sum_{c=1}^{125} |H^i_{shot_k}(c) - H^j_{shot_k}(c)|}{2 \cdot N_p} \tag{6.19}$$

where N_p represents the number of pixels and i and j are two frame indexes.

The normalized inter-frame cumulative distance for the current frame i of the shot k, denoted $D_{shot_k}(i)$, is given by the following equation:

$$D_{shot_k}(i) = \frac{1}{Card(S) - 1} \sum_{j \in S, i \neq j} d_M(H^i_{shot_k}, H^j_{shot_k}) \tag{6.20}$$

where S is the set of the retained frames for shot k and $Card()$ returns the size of a set. This measure gives us information on the correlation between frame i and other frames. If the cumulative distance is low, we may conclude that frame i is similar to most of the shot frames, while if the distance is high, then the image must be different from most of the frames. The normalization has been adopted to be able to compare an histogram of cumulative distances of different shots.

The histogram of cumulative inter-frame distances, $\aleph^D_{shot_k}$ is computed after

quantifying $D_{shot_k}(i)$ values into N_b bins, denoted $D^q_{shot_k}(i)$, in which $i \in S$ and $q = 1, ..., N_b$. $\aleph^D_{shot_k}$ is further determined as:

$$\aleph^D_{shot_k}(d_q) = \sum_{i \in S} \delta(D^q_{shot_k}(i) - d_q) \tag{6.21}$$

in which S is the frame set for shot k, d_q is a quantified value of the normalized cumulative inter-frame distance, q represents the bin index and $\delta(x) = 1$ if $x = 0$ and 0 otherwise. A good tradeoff between computational complexity and the precision of the histogram is $N_b = 100$.

Once the shot histogram is determined, we use the analysis of the histogram shape to measure how the visual activity is related to the distribution of cumulative distances within the shot. After manually observing and analyzing the histograms of cumulative inter-frame distances for more than 50 shots from a large variety of animated movies, we conclude that, despite the diversity of the histograms, they can be projected without significant information lost into only a limited number of patterns, which are related to the type of shot content.

We use only four classes, as follows:

- **histograms with small distance** (pattern 1): all the values of the cumulative distance are small and therefore the variability of the visual content is reduced (shot content is almost constant). This occurs if the maximum cumulative distance over the shot frames is below a certain threshold (e.g., 0.12, empirically determined);

- **histograms with both small and high distances** (pattern 2): most of the cumulative distances are small, but there are a few frames which are very different from the others. This scenario corresponds to shots in which the visual content is mostly constant, but presenting some significant short visual changes (see shot $[8612 - 8657]$ from movie "Ferrailles" in Figure 6.13). To detect these situations, we aim at positioning a mean value of D_{shot_k} with respect to the minimum and maximum values. The pattern 2 histograms are detected if the gap between minimum and mean D_{shot_k} values is less than a fraction (e.g., 0.2) of the gap between maximum and minimum D_{shot_k} values;

- **multi-modal histograms** (pattern 3): in this case the shot contains different groups of similar frames. This scenario corresponds, in general, to several static scenes which are linked through camera motion, e.g., a 3D camera movement with repetitive focuses on several points of the scene (see shot $[78 - 735]$ from movie "The Buddy System" in Figure 6.12);

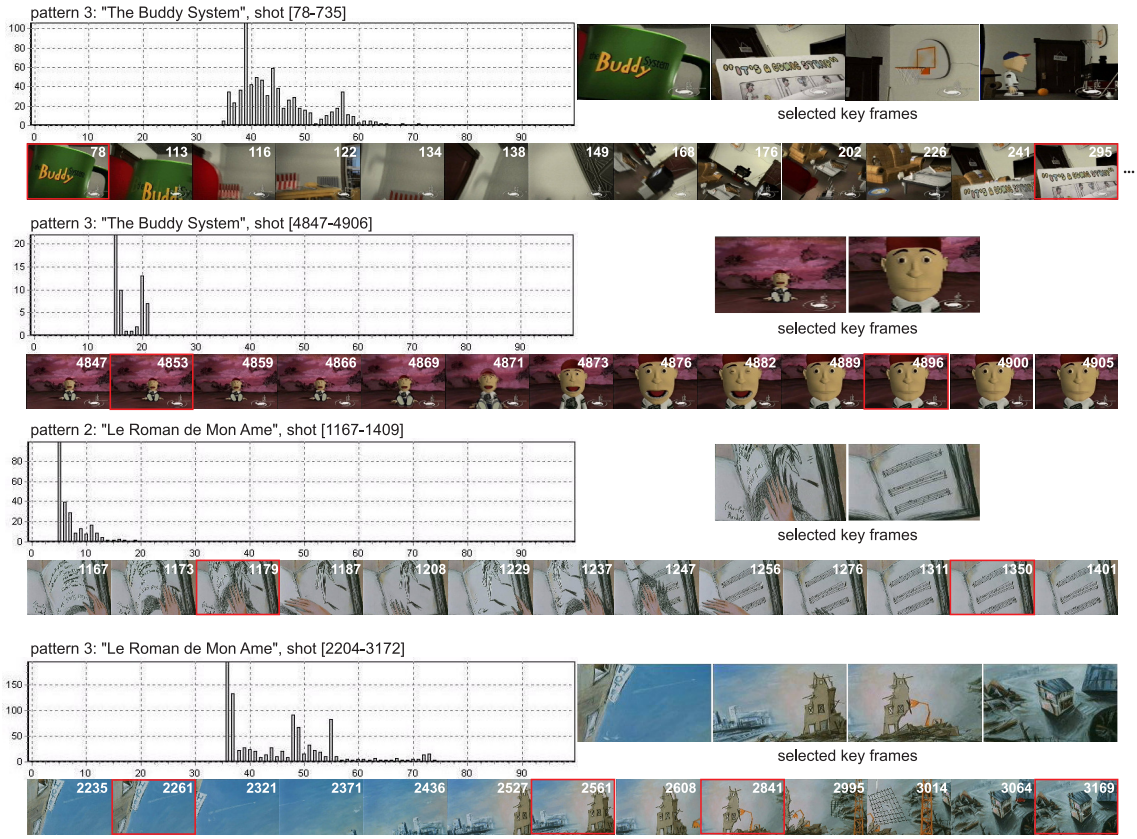


Figure 6.12: Shot summarization using histograms of cumulative inter-frame distances [25]: oX axis corresponds to histogram bins while oY axis corresponds to histogram value. Each shot is summarized with several representative images for visualization purpose (bottom of the histogram). Selected key frames are marked with red rectangles (detailed on the right side of the histogram).

- **single-mode histograms** (pattern 4): the histogram has only one mode, but the cumulative distances are high. Such shots are composed of many different frames suggesting a constantly changing content which may result from continuous motion or due to the use of special color effects.

Some examples of histogram shapes and their corresponding shot contents are presented in Figure 6.12 and 6.13.

To constitute the abstract, we use the following reasoning:

- *rule 1* - For a shot with a cumulative histogram of pattern 1, i.e., small distances, and thus with low visual variability, we extract only one key frame. If $\{frame_i\}$, with $i = 1, ..., N$ (N is the number of frames) denotes the frame set of the current shot, then, the key frame, $frame_k$, is extracted according

Figure 6.13: Shot summarization using histograms of cumulative inter-frame distances (next) [25]: oX axis corresponds to histogram bins while oY axis corresponds to histogram value. Each shot is summarized with several representative images for visualization purpose (bottom of the histogram). Selected key frames are marked with red rectangles (detailed on the right side of the histogram).

to the following equation:

$$k = argmin|_{i \in \{1,...,N\}} \{D_{shot}(i)\} \tag{6.22}$$

in which $D_{shot}(i)$ is the cumulative inter-frame distance given by equation 6.20. Therefore, the key frame is the median image in terms of cumulative distances;

- *rule 2* - A shot with both small and high distances (pattern 2) is represented with two key frames. This type of video shot contains mainly a group of similar frames, as well as some different content images. To capture the content with a reduced visual variability, the first key frame is the median image, selected according to equation 6.22. The second key frame aims at representing the changing content and is selected as $frame_l$, with l given by:

$$l = argmax|_{i \in \{1,...,N\}} \{D_{shot}(i)\} \tag{6.23}$$

This frame is theoretically the most different one, providing the maximal cu-

mulative distance over all the other frames.

- *rule 3* - For shots with cumulative histograms of pattern 3, i.e., multi-modal, which contain different groups of similar content frames, we analyze the histogram peak repartition. The idea is to extract one key frame for each individual group of similar pictures. To do so, we select one key frame for each histogram peak, as being the median image given by equation 6.22 when applied only to the frames which contributed to the peak value of the histogram;

- *rule 4* - Finally, single-mode histograms (pattern 4) are represented with two key frames. In this case, the shot presents a high variability. This is the most difficult case, because a changing content requires a large number of key frames. However, selecting many images, is not always efficient, due to the probability of capturing transition images. We use a compromise and the key frames are selected with the same strategy as for the histograms of pattern 2, thus according to equations 6.22 and 6.23 (the most common image and the most different one).

Validation results

The evaluation of video abstraction techniques is in general a subjective task, as it mainly relies on human perception, e.g., for a certain video sequence one may produce, not one, but many abstracts to cope with some quality constraints.

To test the pertinence of the abstraction approaches, we take advantage of the efficiency of user studies. We have conducted a user study involving 27 people (students, didactic personnel and several animation experts, with ages varying from 21 to 49). The tests were performed on a selection composed of 10 animated movies from CITIA [32], namely: "Casa" (6min15s), "Circuit Marine" (5min35s), "Ferrailles" (6min15s), "François le Vaillant" (8min56s), "Gazoon" (2min47s), "La Bouche Cousue" (2min48s), "La Cancion du Microsillon" (8min56s), "Le Moine et le Poisson" (6min), "Paroles en l'Air" (6min50s) and "The Buddy System" (6min19s).

The test protocol consisted in showing the participants, first, the entire movie and then, the proposed abstracts. The video summaries are presented as slideshows, 1 image/1.5 seconds. After visualizing each abstract, the participants were asked to answer several questions concerning the quality of the proposed abstracts. The answers are quantified into several degrees for which a score is assigned. For each sequence, we then, compute the average score, which provides an overall appreciation, as well as its standard deviation, which gives information about the consistency of the results.

Figure 6.14: Comparison between the proposed adaptive summary [25] (odd lines, symbol I) and the abstract obtained with one image per shot (middle frame, even lines, symbol II): shot boundaries are depicted with red vertical lines, the shot number is depicted with white (extract from the full summary of movie "Circuit Marine", temporal order from top to bottom and left to right).

Several results are illustrated in Figure 6.12, 6.13 and 6.14. Overall, the proposed video summarization strategy gives a good representation of the shot contents, key frames being selected according to the complexity of each shot.

Figure 6.12 and 6.13 present some examples of shot summaries for different patterns of cumulative histograms. For instance, multi-modal histograms conclude with one key frame for each individual group of similar pictures, e.g., the shot $[78-735]$ from the movie "The Buddy System", which contains a 3D continuous camera motion with several focuses on some interesting points of the scene, is summarized with one representative frame for each focused point, or the shot $[4847-4906]$ in which the key frames correspond to each of the two different scenes. On the other hand, visual effects and constant changes may lead to some artificial histogram modes and thus to redundant key frames, as it is the case with frame 2561 extracted from shot $[2204-3172]$, movie "Le Roman de Mon Ame". Shots containing only one group of similar frames and several visual changes are summarized with one common image and one image which captures the variability of the content, see shot $[1167-1409]$ from the movie "Le Roman de Mon Ame" or shot $[8612-8657]$ from movie "Ferrailles". Due to histogram invariance, constant shots are summarized with only one image, despite any object motion of other small movements, see shot

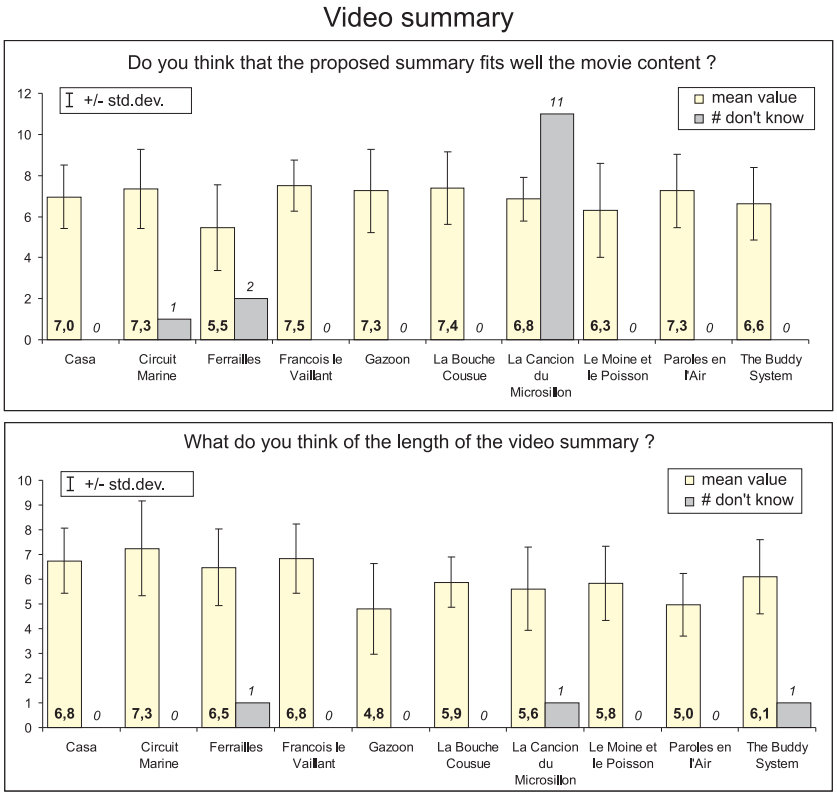


Figure 6.15: The results of the user evaluation campaign [25]: the oX axis corresponds to the tested movies while the oY axis corresponds to the movie average score, the standard deviation is depicted with vertical segments and gray bars correspond to the number of "don't know" answers.

$[3246 - 3433]$ from movie "Gazoon".

Figure 6.14 compares the proposed video summary against a standard approach, i.e., extracting one image for each shot (the middle frame). The adaptive summary provides more details for a changing content (e.g., shots 1, 6, 12, 19, 24, 32 from Figure 6.14), while only one frame is extracted from constant shots (e.g., shots 2, 27, 31, 40, 47 from Figure 6.14). Overall, this results in a picture story of the entire movie action content.

However, a less subjective evaluation is given by the user campaign. The results are presented in Figure 6.15. For the video summary, the evaluation consisted in answering two questions, thus: "*Do you think that the proposed summary fits well the movie content ?*" (score ranges from 0 to 10: 0-don't know, 1,2-not at all, 3,4-very little, 5,6-partially, 7,8-almost entirely, 9,10-entirely) and "*What do you think of the length of the video summary ?*" (score ranges also from 0 to 10: 0-don't know, 1.2-too short, 3,4- short, 5,6-appropriate, 7,8-long, 9,10-very long.)

Concerning the representation of the underling movie contents, the proposed video summary achieved an average score, over all the sequences, of 6.9 and an average standard deviation of 1.7 thus preserving *almost entirely the movie contents*, while for the length of the movie an average value of 6.1 and an average standard deviation of 1.5 thus preserving, in general, *an appropriate length*. However, the summary was less efficient for movies with a very complex content, when the number of "don't know" answers was important, e.g., 11 for the movie "La Cancion du Microsillon", or when the perception of the movie varies from one user to another one, e.g., "Le Moine et le Poisson" (high dispersion of the answers, standard deviation 2.3).

Conclusions and future work

We addressed the issue of video summarization for content-based browsing of animated movies. We proposed a storyboard-like summary, which follows the movie's events by providing each particular movie scene with one key frame. This is carried out at shot level by capturing the shot visual activity with histograms of cumulative inter-frame distances. The number of key frames is adapted to the distribution of the histogram's modes and thus suits the shot's visual content. The performance of our approach has been confirmed through several end-user studies, as well as through the manual analysis of the results. On the whole, the proposed abstracts were appreciated as being sufficiently representative of the movie's content, and were the appropriate length (neither too long nor short). One possible drawback is the length of the summary for movies with high activity content which tends to be too long. Future work will mainly address the possibility of adapting the proposed framework to the case of natural movies.

6.4.2 Video trailer

The second contribution consists of developing techniques for video skimming and in particular for creating trailer-like abstracts. These abstracts are highly compact and present only some of the most exciting action parts of the movies [25].

Contribution to state-of-the-art

The proposed trailer-like video highlight is based on determining movie action segments. This is done by analyzing the movie, both at inter-shot level (action is highlighted by selecting movie segments with a high frequency of video transitions over a certain time-window) and at inter-frame level (computing frame visual activity). Globally, the use of inter-shot analysis is not a particularly new idea, being more or

less adopted by some other approaches, e.g. [81]. The novelty of this work is rather in the efficiency of this relatively standard approach, when transposed and adapted to the specificity of the target animated movies, e.g., we deal with short animated movies (i.e., less than 15-17 minutes) which reduces the span of the action segments within the movie, color is an important feature as we deal with artistic movies in which the author's concepts are transmitted through the movie color palette, etc. Also, to solve the problem of producing video highlights for movies with a predominant static content we adapt the concept of action to the movie average rhythm. Despite the complexity of the content, in our approach we consider some simplifications, e.g., most of the movies are without dialog or commentaries, therefore, the sound is disregarded, and so is the issue of image-sound synchronization.

Approach

To produce a video highlight similar to the concept of a video trailer, we use a simple and efficient approach. It basically consists in summarizing the movie's action clips. Action detection is performed using the algorithm proposed for rhythm and action descriptors, which is presented in Section 6.3.1 (see Figure 6.5). The result is the segmentation of the movie into action and non-action segments.

As stated before, with this approach, the trailer captures the movie "most uncommon" parts. The movie trailer is computed as follows:

$$trailer = \bigcup_{m=1}^{M} \bigcup_{n=1}^{N_m} seq_{p\%}^{n} \tag{6.24}$$

where M is the number of action clips, N_m represents the number of video shots within the action clip m, and $seq_{p\%}^{n}$ is an image sequence which contains $p\%$ of the shot n frames. As the action takes place most likely in the middle of a shot, the sequence is shot-centered. Retaining a number of frames according to the shot length, provides longer shots with more details, which are more valuable as they contain more information.

The choice of parameter p is related to the histogram of cumulative inter-frame distances, $\aleph_{shot_k}^{D}$, defined in equation 6.21 for the creation of the video storyboard (see Section 6.4.1). This histogram captures the variability within the shot frames, which provides complementary information to the action analysis.

We adapt the amount of the retained shot information to its visual activity. Therefore, for shots with a cumulative histogram of pattern 1 or 2 (see also Figure 6.13), which contain similar color information, we use a smaller value of p, around 15%. On the other hand, for shots with cumulative histograms of patterns 3 or 4 (see

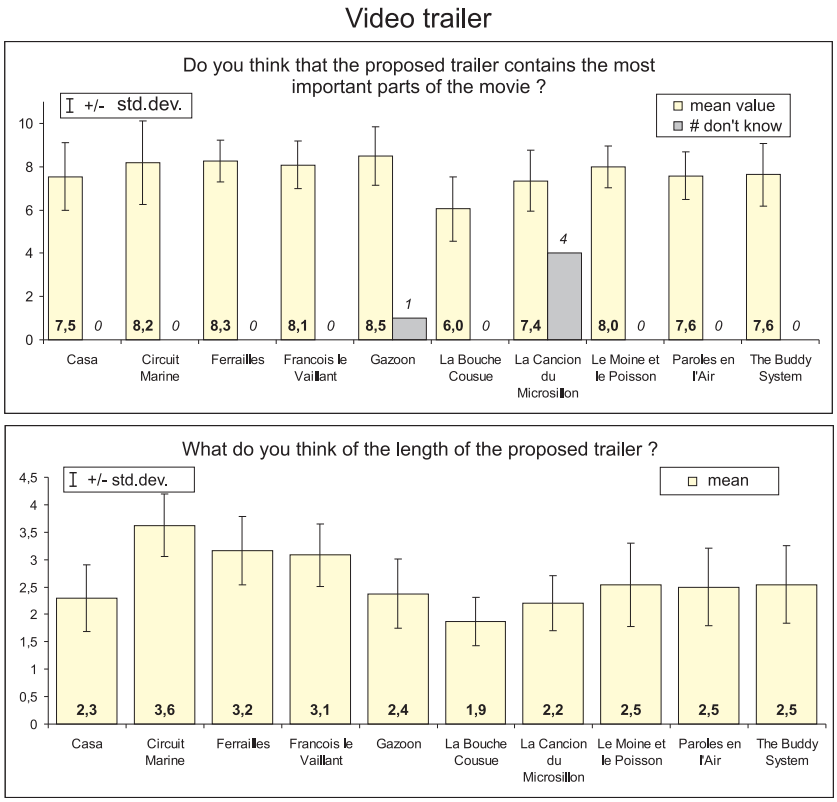


Figure 6.16: The results of the user evaluation campaign [25]: the oX axis corresponds to the tested movies while the oY axis corresponds to the movie average score, the standard deviation is depicted with vertical segments and gray bars correspond to the number of "don't know" answers.

also Figure 6.12), which contain much more action information, we use $p = 35\%$. The values of p were empirically determined after the manual analysis of several animated movies. The constraint is, first of all, to ensure the visual continuity of the trailer as well as to preserve an optimal trailer length.

Validation results

Experimental validation was conducted on the same dataset and user study presented in the previous section, Section 6.4.1, for the video storyboard evaluation.

For the trailer evaluation, the user evaluation consisted in answering two questions namely: "*Do you think that the proposed trailer contains the most important parts of the movie ?*" (score ranging from 0 to 10: 0-don't know, 1,2-not at all, 3,4-very few, 5,6-some, 7,8-almost all and 9,10-all of them) and "*What do you think of the length of the proposed trailer ?*" (score range from 0 to 4: 0-very short, 1-short,

Table 6.11: A comparative study of the achieved trailer length (action ratio = action length/movie length).

Movie	Length	Trailer length	Shot-based skim length	# Shots	Action ratio	Comp. ratio
"François le Vaillant"	8min56s	**1min25s**	2min15s	164	70%	6:1
"La Bouche Cousue"	2min48s	**16s**	42s	39	52.5%	10:1
"Ferrailles"	6min15s	**1min31s**	1min34s	138	98%	4:1
"Casa"	6min15s	**42s**	1min30s	49	87%	9:1
"Circuit Marine"	5min35s	**55s**	1min22s	125	87%	6:1
"Gazoon"	2min47s	**35s**	42s	31	89%	4:1
"La Cancion du Microsillon"	8min56s	**52s**	2min13s	97	55%	10:1
"Le Moine et le Poisson"	6min	**55s**	1min30s	99	74%	7:1
"Paroles en l'Air"	6min50s	**57s**	1min42s	63	77%	7:1
"The Buddy System"	6min19s	**1min**	1min36s	77	77%	6:1
"A Viagem"	7min32s	**1min**	1min48s	54	71%	8:1
"David"	8min12s	**23s**	1min58s	27	40%	**21:1**
"Greek Tragedy"	6min32s	**24s**	1min36s	29	48%	16:1

2-appropriate, 3-long and 4-very long). The results are depicted in Figure 6.16.

Overall, the proposed video trailer was perceived as providing *almost all the important parts of the movie*, with a global average score, over all the sequences, of 7.7 and a standard deviation of 1.3. This corresponds to our goal, as video trailers do not aim at providing all the action contents or exciting parts. Compared to the video summary, thanks to the dynamic content, the trailer was naturally more attractive for the viewers, thus the answers are less dispersed (smaller standard deviation) while the number of "don't know" answers is reduced (5 vs. 14).

The trailer length, was considered *appropriate*, with a global average score of 2.6 and a standard deviation of 0.6. However, for movies with a predominant action content, e.g., "François le Vaillant", "Ferrailles", the trailer tends to be longer.

In Table 6.11 we compare the length of the proposed trailer versus the original movie and a standard skimming approach which consists in retaining $p\%$ frames from each individual video shot (we take p=25% which is close to the average p value used with the trailer construction). One may notice that the trailer provides a good reduction of the original movie contents, with an average compression around 9:1, while the max value is up to 21:1. This is done while preserving the movie most important parts, as confirmed by the previous results. Also, compared to the

standard approach, in most cases, the trailer is more efficient. The only cases when the trailer approaches the skim length is for some of the movies with a predominant action content, that is a high action ratio (see in Table 6.11 the movies with an action ratio above 85%).

Conclusions and future work

We addressed the issue of video summarization for content-based browsing of animated movies. We proposed a trailer-like video highlight, which provides only the most interesting parts of the movie. Our approach is based on highlighting action through the analysis of the movie's rhythm (frequency of shot changes) coupled with the analysis of the shot's visual activity. We adapt the notion of action to each movie's content to solve the problem of producing trailers for movies with a more or less static content.

The performance of our approach has been confirmed through several end-user studies, as well as through the manual analysis of the results. On the whole, the proposed abstracts were appreciated as being sufficiently representative of the movie's content, and were the appropriate length. The proposed trailer achieves a good reduction of the original video content, i.e., average compression ratio 9:1, while still preserving most of the interesting parts of the movie. Naturally the trailer is attractive for the viewers, thanks to the dynamic content.

Future work will mainly consist of improving the criteria used for action detection by considering motion information. At a certain level, the visual changes caused by motion are captured with the histograms of cumulative inter-frame distances, but getting specific information about fast camera/object motion would be more valuable for the production of the video trailer.

6.5 Violent scenes detection

Video broadcasting footage (e.g., YouTube, Dailymotion) is now the largest broadband traffic category on the Internet, comprising more than a quarter of total traffic (source CISCO systems[10]). In this context, one of the emerging research areas is the *automatic filtering of video contents*. The objective is to select appropriate content for different user profiles or audiences.

A significant interest was shown for *detecting violent scenes* in movies (or affect content) which is an important requirement in various use cases related to video on demand and child protection against offensive content.

[10] `http://www.cisco.com`.

Apart from the inherent scientific challenge, solving this paradigm requires first an adequate formalisation of this highly subjective concept, i.e., violence. Another important aspect is the validation of the techniques. Before 2011 (i.e., the MediaEval Affect Task: Violent Scenes Detection [100]), there was a lack of a standard consistent and substantial evaluation framework (both from the dataset and annotations point of view). This limited significantly the reproducibility of results in the community and consequently the advances in this specific field. Each of the proposed methods tended to be tested on closed data, usually very restraint and annotated for very particular types of violence.

6.5.1 Violence classification

In this area, I have first contributed to the development of an automated technique for the detection of video segments that contain physical violence or accidents resulting in human injury or pain in typical Hollywood productions [15][11].

Contribution to state-of-the-art

In the context of the current state-of-the-art, we proposed a different perspective that exploits for the violent scenes detection the use of mid-level concepts in a multiple neural network fusing scheme. The proposed approach goes beyond the current state-of-the-art along these dimensions: by addressing a highly complex scenario where violence is considered to be any scene involving human injury or pain; thanks to the fusion of mid-level concept predictions, the method is feature-independent in the sense that it does not require the design of adapted features; violence is predicted at frame level which facilitates detecting segments of arbitrary length, not only fixed length (e.g., shots).

[11] this work was developed in cooperation with Jan Schlüter, from Intelligent Music Processing and Machine Learning Group, Austrian Research Institute for Artificial Intelligence, Vienna-Austria, Assoc. Prof. Markus Schedl, from Department of Computational Perception, Johannes Kepler University, Linz-Austria and Dr. Ionuţ Mironică, from LAPI, University Politehnica of Bucharest, Romania. The presented results were published in:

[15] *B. Ionescu*, J. Schlüter, I. Mironică, M. Schedl, "*A Naive Mid-level Concept-based Fusion Approach to Violence Detection in Hollywood Movies*", ACM International Conference on Multimedia Retrieval - ICMR 2013, Dallas, Texas, USA, April 16 - 19, 2013.

[96] J. Schlüter, *B. Ionescu*, I. Mironică, M. Schedl , "*ARF @ MediaEval 2012: An Uninformed Approach to Violence Detection in Hollywood Movies*", MediaEval Workshop, vol. 927, CEUR-WS.org, ISSN 1613-0073, Pisa, Italy, 4-5 October, 2012.

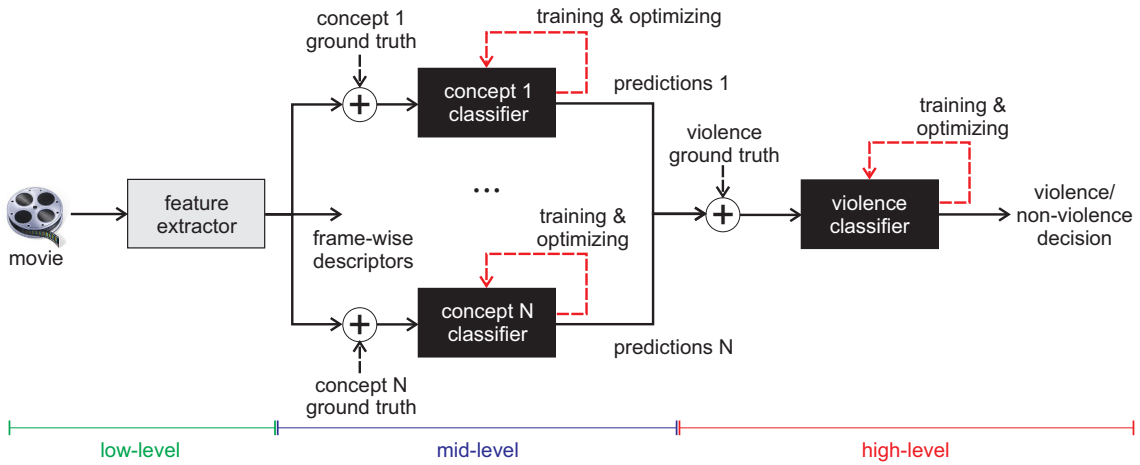


Figure 6.17: Violent scene detection diagram [15].

Approach

Given the high variability in appearance of violent scenes in movies and the low amount of training data that is usually available, training a classifier to predict violent frames directly from visual and auditory features seems rather ineffective. We propose instead to use high-level concept ground-truth obtained from manual annotation to infer mid-level concepts as a stepping stone towards the final goal. Predicting mid-level concepts from low-level features should be more feasible than directly predicting all forms of violence (highly semantic). Also, predicting violence from mid-level concepts should be easier than using directly the low-level features.

A diagram of the proposed method is shown in Figure 6.17. First, we perform feature extraction. Features are extracted at frame level. The resulting data is then fed into a multi-classifier framework that operates in two steps. The first step consists of *training* the system using ground truth data. Once we captured data characteristics we may *classify* unseen video frames into one of the two categories: "violence" and "non-violence". Violence frames are aggregated to segments.

To train the system we use ground truth data at two levels: ground truth related to concepts that are usually present in the violence scenes, such as presence of "fire", presence of "gunshots", or "gory" scenes and ground truth related to the actual violence segments. We used the data set provided with the 2012 MediaEval Affect Task: Violent Scenes Detection Task [82].

The mid-level concept detection consists of a bank of classifiers that are trained to respond to each of the target violence-related concepts. At this level, the response of the classifier is optimized for best performance. Tests are repeated for different parameter setups until the classifier yields the highest accuracy. Each classifier state

is then saved. With this step, initial features are therefore transformed into concept predictions (real valued between [0;1]).

The high-level concept detection is ensured by a final classifier that is fed with the previous concept predictions and acts as a final fusion scheme. The output of the classifier is thresholded to achieve the labeling of each frame as "violent" or "non-violent" (yes/no decision). As in the previous case, we use the violence ground truth to tune the classifier to its optimal results (e.g., setting the best threshold). The classifier state is again conserved.

To choose the right classification scheme for this particular fusion task, we conducted several preliminary experimental tests using a broad variety of classifiers, from functional-based (e.g., Support Vector Machines), decision trees to neural networks. Most of the classifiers failed in providing relevant results when coping with high amount of input data, i.e., labeling of individual frames rather than video segments (e.g., a movie has around 160,000 frames and the training data consist of million frames). The inherent parallel architecture of neural networks fitted well these requirements, in particular the use of multi-layer perceptrons. Therefore, for the concept and violence classifiers (see Figure 6.17) we employ a multi-layer perceptron with a single hidden layer of 512 logistic sigmoid units and as many output units as required for the respective concept.

Networks are trained by gradient descent on the cross-entropy error with back-propagation [83], using a recent idea by Hinton et al. [84] to improve generalization: for each presented training case, a fraction of input and hidden units is omitted from the network and the remaining weights are scaled up to compensate. This helps generalization in the following way: by randomly omitting units from the network, a higher-level unit cannot rely on all lower-level units being present and thus cannot adapt to very specific combinations of a few parent units only. Instead, it is driven to find activation patterns of a larger group of correlated units, such that dropping a fraction of them does not hinder recognizing the pattern.

Validation results

To validate our approach we use the 2012 MediaEval Affect task: Violent Scenes Detection [82] dataset. It proposes a corpus of 18 Hollywood movies of different genres, from extremely violent movies to movies without violence. Movies are divided into a development set, consisting of 15 movies: "Armageddon", "Billy Elliot", "Eragon", "Harry Potter 5", "I am Legend", "Leon", "Midnight Express", "Pirates of the Caribbean 1", "Reservoir Dogs", "Saving Private Ryan", "The Sixth Sense", "The Wicker Man", "Kill Bill 1", "The Bourne Identity", and "The Wizard of Oz" (total duration of 27h 58min, 26,108 video shots and violence duration ratio 9.39%);

and a test set consisting of 3 movies: "Dead Poets Society", "Fight Club", and "Independence Day" (total duration 6h 44min, 6,570 video shots and violence duration ratio 4.92%). Overall the entire data set contains 1,819 violence segments.

Ground truth is provided at two levels. Frames are annotated according to 10 violence related high-level concepts, namely: "presence of blood", "fights", "presence of fire", "presence of guns", "presence of cold weapons", "car chases" and "gory scenes" (for the video modality); "presence of screams", "gunshots" and "explosions" (for the audio modality) [85]; and frame segments are labeled as "violent"/"non-violent".

For experimentation, we employ the following content descriptors:

- *audio* (196 dimensions): we use a general-purpose set of audio descriptors: Linear Predictive Coefficients (LPCs), Line Spectral Pairs (LSPs), MFCCs, Zero-Crossing Rate (ZCR), and spectral centroid, flux, rolloff, and kurtosis, augmented with the variance of each feature over a window of 0.8 s centered at the current frame [86, 87];

- *color* (11 dimensions): to describe global color contents, we use the Color Naming Histogram proposed in [89]. It maps colors to 11 universal color names: "black", "blue", "brown", "grey", "green", "orange", "pink", "purple", "red", "white", and "yellow";

- *features* (81 values): we use a 81-dimensional Histogram of Oriented Gradients (HoG) [88];

- *temporal structure* (single dimension): to account for temporal information we use a measure of visual activity. We use the cut detector in [28] that measures visual discontinuity by means of difference between color histograms of consecutive frames. To account for a broader range of significant visual changes, but still rejecting small variations, we lower the threshold used for cut detection. Then, for each frame we determine the number of detections in a certain time window centered at the current frame (e.g., for violence detection good performance is obtained with 2 s windows). High values of this measure will account for important visual changes that are typically related to action.

The classifiers are trained on the development data while the actual evaluation was carried out on the testset.

To assess performance, we use classic precision and recall as defined in equation 6.5 as well as the global Fscore, see equation 6.12. Values are averaged over all experiments.

Table 6.12: Violence shot-level detection results at 2012 MediaEval Affect task: Violent Scenes Detection [82].

team	*descriptors*	*modality*	*method*	*precision*	*recall*	*Fscore*
ARF-(c)	concepts	audio-visual	**proposed**	46.14%	54.40%	49.94%
ARF-(a)	audio	audio	**proposed**	46.97%	45.59%	46.27%
ARF-(av)	audio, color, HoG, temporal	audio-visual	**proposed**	32.81%	67.69%	44.58%
Shanghai Hongkong [92]	trajectory, SIFT, STIP, MFCC	audio-visual	temp. smoothing + SVM with χ^2 kernel	41.43%	46.29%	43.73%
ARF-(avc) [96]	audio, color, HoG, temporal & concepts	audio-visual	**proposed**	31.24%	66.15%	42.44%
TEC [95]	TF-IDF B-o-AW [99], audio, color	audio-visual	fusion SVM with HIK and χ^2 kernel & Bayes Net. & Naive Bayes	31.46%	55.52%	40.16%
TUM [91]	energy & spectral audio	audio	SVM linear kernel	40.39%	32.00%	35.73%
ARF-(v)	color, HoG, temporal	visual	**proposed**	25.04%	61.95%	35.67%
LIG [93]	color, texture, SIFT, B-o-AW of MFCC	audio-visual	hierarch. fusion of SVMs & k-NNs with conceptual feedback	26.31%	42.09%	32.38%
TUB [97]	audio, B-o-AW MFCC, motion	audio-visual	SVM with RBF kernel	19.00%	62.65%	29.71%
DYNI [94]	MS-LBP texture [98]	visual	SVM with linear kernel	15.55%	63.07%	24.95%
NII [90]	concept learned from texture & color	visual	SVM with RBF kernel	11.40%	89.93%	20.24%

Notations: SIFT - Scale Invariant Features Transform, STIP - Spatial-Temporal Interest Points, MFCC - Mel-Frequency Cepstral Coefficients, SVM - Support Vector Machines, TF-IDF - Term Frequency-Inverse Document Frequency, B-o-AW - Bag-of-Audio-Words, HIK - Histogram Intersection Kernel, k-NN - k Nearest Neighbors, RBF - Radial Basis Function, MS-LBP - Multi-Scale Local Binary Pattern.

Shot-level prediction. The first experiment addressed the shot-level prediction (shot segmentation for the movies was provided by organizers [82, 85]). We assessed different feature combinations: ARF-(c) - use of only mid-level concept predictions; ARF-(a) - use of only audio descriptors (the violence classifier is trained directly on the audio features); ARF-(v) - use of only visual features; ARF-(ac) - use of only audio-visual features; ARF-(avc) - use of all concept and audio-visual features (the

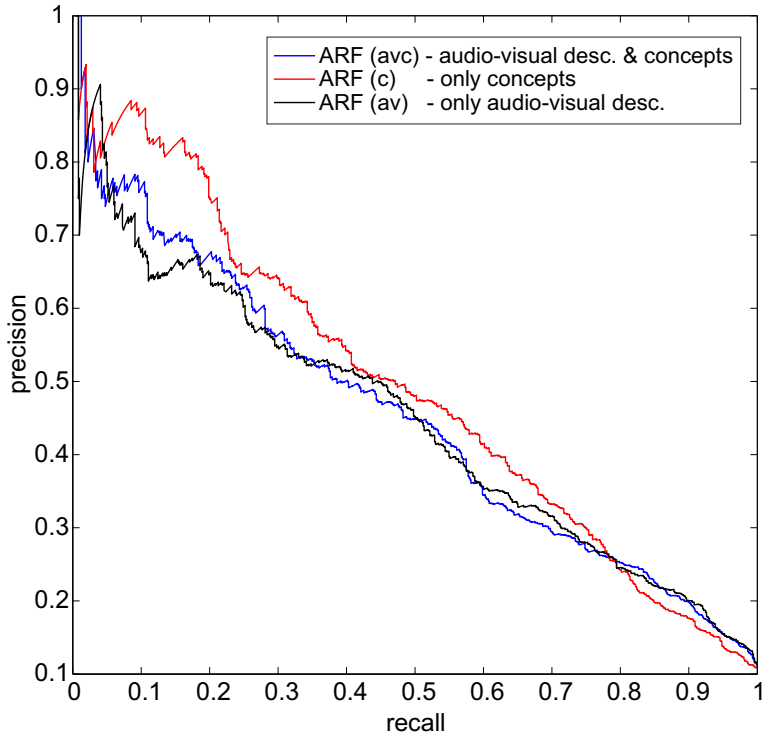


Figure 6.18: Precision-recall curves for the proposed violent scenes detection approach [15].

violence classifier is trained using the fusion of concept predictions and features).

A summary of the 2012 MediaEval best team runs is presented in Table 6.12 (results are presented in decreasing order of Fscore values). The use of mid-level concept predictions and multi-layer perceptron (see ARF-(c)) ranked first and achieved the highest $Fscore$ of 49.94%, that is an improvement of more than 6 percentage points over the other teams' best runs, i.e., team ShanghaiHongkong [92], $Fscore$ of 43.73%. For our approach, the lowest discriminative power is provided by using only the visual descriptors (see ARF-(v)), where the $Fscore$ is only 35.65%. Compared to visual features, audio features seem to show better descriptive power, providing the second best $Fscore$ of 46.27%. The combination of descriptors (early fusion) tends to reduce their efficiency and yields lower performance than the use of concepts alone, e.g., audio-visual (see ARF-(av)) yields an $Fscore$ of 44.58%, while audio-visual-concepts (see ARF-(avc)) 42.44%.

Another observation is that, despite the use of general purpose descriptors, the representation of feature information via mid-level concepts allows better performance than other, more elaborate content description approaches or classification, such as the use of SIFTs, B-o-AW of MFCC or motion information.

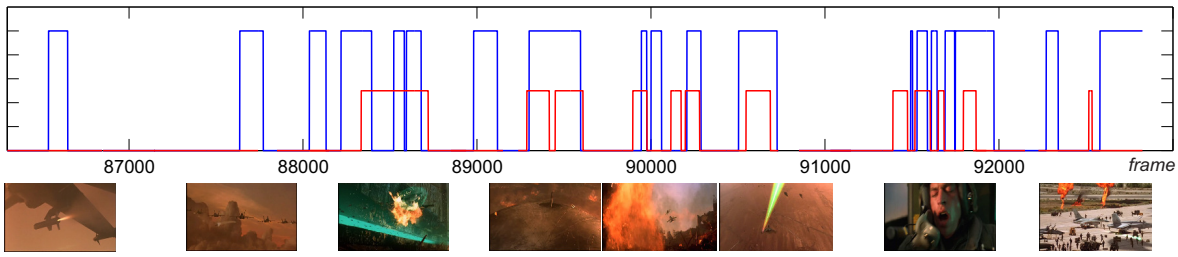


Figure 6.19: Examples of violent segment detection in movie "Independence Day" [15] (the oX axis is the time axis, the values on oY axis are arbitrary, ground truth is depicted in red while the detected segments in blue).

Figure 6.18 details the precision-recall curves for our approach. One may observe that the use of concepts alone (red line) provides significantly higher recall than the sole use of audio-visual features or the combination of all for a precision of 25% and above.

Arbitrary segment-level prediction. The final experiment is conducted at segment level. Video segments of arbitrary length are tagged as "violent"/"non-violent".

Using the mid-level concepts, we achieve average precision and recall values of 42.21% and 40.38%, respectively, while the $Fscore$ amounts to 41.27%. This yields a miss rate (at time level) of 50.69% and a very low false alarm rate of only 6%. These results are also very promising considering the difficulty of detecting precisely the exact time interval of violent scenes, but also the subjectivity of the human assessment (reflected in the ground truth). Comparison with other approaches was not possible in this case as all other teams provided only shot-level detection.

Figure 6.19 illustrates an example of violent segments detected by our approach in the movie "Independence Day". For visualization purposes, some of the segments are depicted with a small vignette of a representative frame.

In general, the method performed very well on the movie segments related to action (e.g., involving explosions, firearms, fire, screams) and tends to be less efficient for segments where violence is encoded in the meaning of human actions (e.g., fist fights or car chases). Examples of false detections are due to visual effects that share similar audio-visual signatures with the violence-related concepts. Common examples include accentuated fist hits, loud sounds or the presence of fire not related to violence (e.g., see the rocket launch or the fighter flight formation in Figure 6.19, first two images). Misdetection is in general caused by limited accuracy of the concept predictors (see last image in Figure 6.19, where some local explosions filmed from a distance have been missed).

Conclusions and future work

We presented a naive approach to the issue of violence detection in Hollywood movies. Instead of using content descriptors to learn directly how to predict violence, as most of the existing approaches do, the proposed approach relies on an intermediate step consisting of predicting mid-level violence concepts.

Content classification is performed with a multi-layer perceptron whose parallel architecture fits well the target of labeling individual video frames. The approach is naive in the sense of its simplicity. Nevertheless, its efficiency in predicting arbitrary length violence segments is remarkable. The proposed approach ranked first in the context of the 2012 Affect Task: Violent Scenes Detection at MediaEval Multimedia Benchmark (out of 36 total submissions).

However, the main limitation of the method is its dependence on a detailed annotation of violent concepts, inheriting at some level its human subjectivity. Future improvements will include exploring the use of other information sources, such as text (e.g., subtitles that are usually provided with movie DVDs).

6.5.2 Benchmarking violent scenes detection

Another contribution to this specific area was in developing a formalization for the concept of violence in Hollywood productions and in defining a common evaluation framework (annotated dataset and benchmarking), within the 2013 MediaEval Affect Task: Violent Scenes Detection [100][12]

Contribution to state-of-the-art

In the literature, violent scene detection in movies has received very little attention so far. Moreover, comparing existing results is impossible because of the different

[12]this work was developed in cooperation with Dr. Claire-Hélène Demarty, Dr. Cédric Penet, from Technicolor R&D, France, Assoc. Prof. Yu-Gang Jiang, from Fudan University, Shanghai, China, Assist. Prof. Vu Lam Quang, from MMLAB, University of Information Technology, Vietnam, Assoc. Prof. Markus Schedl, from Department of Computational Perception, Johannes Kepler University, Linz-Austria, Guillaume Gravier, from IRISA & INRIA Rennes-France, and Mohammad Soleymani, from iBUG, Imperial College London-UK. The presented results were published in:

[5] C.-H. Demarty, *B. Ionescu*, Y.-G. Jiang, V.L. Quang, M. Schedl, C. Penet, "*Benchmarking Violent Scenes Detection in Movies*", IEEE International Workshop on Content-Based Multimedia Indexing - CBMI 2014, 18-20 June, Klagenfurt, Austria, 2014.

[2] C.-H. Demarty, C. Penet, *B. Ionescu*, G. Gravier, M. Soleymani, "*Multimodal violence detection in Hollywood movies: State-of-the-art and Benchmarking*", in book "Fusion in Computer Vision - Understanding Complex Visual Content", Springer International Publishing Switzerland ACVPR - Advances in Computer Vision and Pattern Recognition, 2014.

definitions of violence adopted. As a consequence, methods suffer from a lack of standard, consistent and substantial datasets. The MediaEval Affect Task: Violent Scenes Detection (which has been run annually since 2011) constitutes a first attempt to address all these needs and establish a standard with state-of-the-art performance for future reference.

We developed the 2013 edition of the benchmarking [5], which compared the 2011 and 2012 [2] editions, formalized two distinct use case scenarios for the definition of violence (i.e., an objective and a subjective definition) with the aim to understand how different scenarios influence the systems' performance. Furthermore, a very consistent benchmarking dataset was made publicly available providing full annotations for no less than 25 Hollywood movies. Finally, the 2013 edition allowed the evaluation of systems that make use of external data (e.g., from Internet) which allows for testing open systems.

Dataset

The 2013 edition of the dataset was built on top of the 2012 edition [100]. In total, the dataset contains 25 Hollywood movies of different genres and different amounts of violence (from extremely violent movies to movies without violence), namely: "Armageddon", "Billy Elliot", "Eragon", "Harry Potter 5", "I am Legend", "Leon", "Midnight Express", "Pirates of the Caribbean 1", "Reservoir Dogs", "Saving Private Ryan", "The Sixth Sense", "The Wicker Man", "Kill Bill 1", "The Bourne Identity", "The Wizard of Oz", "Dead Poets Society", "Fight Club", "Independance Day", "Fantastic Four", "Fargo", "Forrest Gump", "Legally Blond", "Pulp Fiction", "The God Father" and "The Pianist".

The main novelty of the 2013 benchmarking consists of adopting two different definitions of violence according to two different use case scenarios:

- **objective definition** - the first use case scenario is a following of what was proposed in the previous years, where targeted violent segments are those showing "*physical violence or accident resulting in injury or pain*". Although it was designed to be as objective as possible, this definition has proven to lead to inconsistencies/ambiguities between the annotated segments and the original Technicolor use case [85], e.g., not really violent segments such as "somebody hurting himself while shaving" were considered as violent whereas segments depicting dead people but without showing the cause of death were discarded from the annotations;

- **subjective definition** - to experiment with another perspective of violence, the second use case features a more subjective definition, namely violent seg-

ments are "*those which one would not let an 8 years old child see because they contain physical violence*".

Data was annotated for both use case scenarios[13]. For the objective definition, the annotations were carried out by three human assessors using the following protocol. Firstly, two annotators labelled all the videos separately. During the annotation process, no discussions were held between them in order for the process to be totally independent. Secondly, a third master annotator merged all their annotations and reviewed the movies once again to minimize the chance of missing any violent segments. Doubtful annotations were solved via panel discussions. Each annotated violent segment contains a single action of violence whenever possible. However, if there are multiple actions in a continuous segment, the segment was annotated as a whole. The annotation granularity was decided to be at frame level.

For the subjective definition, the annotations were carried out by seven human assessors (5 regular annotators and 2 master annotators). Given the specificity of this scenario it is worth mentioning the profile of the annotators: regular annotators were graduate students (single with no children) and master annotators were lecturers (married with children). In this case the following protocol was used. Firstly, two regular annotators labelled all the movies separately. Secondly, the third regular annotator merged, reviewed and also revisited the movies to retrieve any possible missing violent segments. Once again, no discussions were held between annotators. Finally, a fourth master annotator reviewed the data from a parent perspective and refined the results. All the uncertain (borderline) cases were solved via panel discussions, involving different people from different countries and culture, to avoid cultural bias in the annotations. In contrast with the objective definition where violent segments focused on violent actions and their results, the subjective violent segments focus on the overall context of violent scenes. As a result, subjective segments tend to be slightly longer than the objective ones.

As for the previous editions of the benchmark, in addition to general violent segments annotation, a set of 10 high-level violence related concepts were annotated: "presence of blood", "fights", "presence of fire", "presence of guns", "presence of cold weapons", "car chases", "gory scenes", "gunshots", "explosions" and "screams" [2].

The dataset is divided into a development dataset intended for training the systems and a test dataset for the actual benchmarking [100]. The development set contains 32,678 shots (as obtained with automatic segmentation) from 18 movies for a total duration of 35h18min. According to the objective definition, violent shots

[13]the 2013 violent scenes detection dataset is publicly available and can be downloaded from `http://www.technicolor.com/en/innovation/research-innovation/scientific-data-sharing/violent-scenes-dataset`.

cover 12% of the shots and 9.12% of the total duration, whereas for the subjective definition, violent shots represent 21.45% of the shots and 14.74% of the duration. These figures highlight the fact that globally the subjective definition proposes segments of longer durations, and therefore covers a bigger proportion of the database. The 2013 test set consists of 7 movies (containing non violent to highly violent movies; total duration of 14h44min and 11,245 shots).

Benchmarking results

The proposed benchmarking framework was validated during the 2013 MediaEval Affect Task: Violent Scenes Detection [100]. It required participants to automatically detect violent portions of Hollywood movies by the use of multimodal features. Participants' systems were trained using the development data whereas the actual validation was carried out on the test dataset. The two definitions of violence considered were evaluated as two different sub-tasks.

For each subtask, participants were allowed to submit the following types of runs (up to 5 runs): shot-based classification without use of any external data other than the content of the DVDs (shot segmentation is provided by organizers), shot-based classification with use of external data, segment level classification without external data (participants are required to provide segment boundaries independently of the shot segmentation) and segment level classification with external data. In each case, each shot or segment has to be provided with a confidence score. For both subtasks, the required run is the run at shot level without use of external data.

In 2013, the proposed benchmarking has seen a substantial increase in participation, in total, 59 runs have been evaluated, divided between the objective (36 runs - 30 were targeting shot level prediction and 6 segment level prediction) and the subjective (23 runs - 21 runs for shot level prediction while only 2 for the segment level prediction) subtasks.

To assess performance, similar to the last years' benchmarkings, several metrics were computed, from false alarm and miss detection rates, AED-precision/recall, MediaEval cost (a function weighting false alarms and missed detections) to Detection Error Trade-off curves and Mean Average Precision. However, in 2013, the official metric was selected to be the standard Mean Average Precision (MAP), as defined with equation 6.13. In particular, in 2013, systems were optimized for a cutoff point of 100 top ranked violent segments (MAP@100).

Evaluation in terms of MAP. Table 6.13 reports the MAP@100 and MAP metrics for the best team runs for both objective and subjective use cases as well as for shot and segment level evaluation (notations: a - audio, v - visual, c - mid-level concepts, l - late fusion and e - early fusion; highest values are represented in bold).

Table 6.13: Overall MAP@100 and MAP (%) for best team runs at 2013 MediaEval Affect Task: Violent Scenes Detection [100] (according to the official metrics).

team	*objective*						*subjective*					
	shots			segments			shots			segments		
	runid	MAP @100	MAP	runid	MAP @100	MAP	runid	MAP @100	MAP	runid	MAP @100	MAP
FUDAN [101]	run5 (avc-l)	**55.3**	**51.1**	-	-	-	run5 (avc-l)	68.2	58.6	-	-	-
LIG [105]	run2 (av-l)	52.1	50.5	-	-	-	run1 (av-l)	**69**	67.3	-	-	-
FAR [102]	run1 (a)	49.6	47.6	run5 (avc-l)	35	**34.5**	-	-	-	-	-	-
TUDCL [106]	run2 (av-l)	46.9	38.7	run1 (av-l)	**42**	34.3	-	-	-	-	-	-
NII-UIT [108]	run1 (avc-l)	43.6	23.4	-	-	-	run1 (avc-l)	59.6	37.9	-	-	-
TECH-INRIA [109]	run1 (c_a)	33.8	28.8	run3 (c_{av}-l)	12.5	14.6	run1 (c_a)	53.6	44.6	run1 (c_a)	**44.8**	**35.3**
VIREO [104]	run4 (avc-l)	31.6	31.6	-	-	-	run4 (avc-l)	68.9	**67.5**	-	-	-
MTM [107]	run1 (av-e)	7.4	12.6	-	-	-	-	-	-	-	-	-
VISILAB [103]	-	-	-	run2 (v)	14.9	13.9	-	-	-	-	-	-

What is interesting to notice is that regardless of the use case scenario and the granularity of the prediction, highest performance is achieved when including mid-level information with multimodal late fusion approaches (see avc-l): objective shot-level prediction - MAP 51.1%, FUDAN run5 [101], objective segment-level - MAP 34.5%, FAR run5 [102], subjective shot-level - MAP 67.5%, VIREO run4 [104]. At modality level, visual information alone seems to provide too little discriminative power for this high level task, e.g., VISILAB run2 [103] is able to achieve a MAP of only 13.9% which is less than half the performance of the best system. Another interesting result is that in particular, using audio modality alone, it allows to achieve very good performance, e.g., for objective use case FAR run1 [102] leads to a MAP of 47.6% while TECH-INRIA run1 [109] reaches a MAP of 44.6% for the subjective scenario. This may be due to the fact that most of the violent scenes in movies tend to come with specific audio signatures.

In what concerns the granularity of the predictions, shot-based estimation is more accurate than the prediction at arbitrary length segments, e.g., TUDCL run2 [106] leads to MAP=38.7% for shot level while the same run achieves MAP=34.3%

for segments (the difference is greater for the best performing runs). Tagging directly some predefined shots is indeed a classification task, and therefore easier than the task at segment level, where a step of boundaries segmentation is involved. Furthermore, systems proposing oversegmented events at segment level will be penalized during MAP computation, as potentially a higher number of false alarms (one per segment) may be ranked in the first 100 returned results.

The predictions of the subjective use case scenario lead to significantly higher results than for the objective one, e.g., highest MAP at shot level is 51.1% (FUDAN run5 [101]) for the objective scenario while the same run achieves up to 58.6% for the subjective one. A possible explanation comes from the fact that the subjective annotations lead to longer and more unitary shots than for the objective one, where the focus was on identifying each particular individual scene.

Finally, in what concerns the cutoff point, reporting MAP@100 leads to slightly better results than the overall MAP prediction. This is useful in case the violence prediction system is considered from the perspective of retrieval where violence segments are searched within the movies. In this case, highest performing system is the one retrieving the largest number of best results at the first top ranks.

Evaluation of false and missed detections. As shown in Figure 6.20, the overall performances in terms of false and missed detections are similar for the best participants and reach 20% false alarms for 20% missed detections for objective definition at shot level, and 25% false alarms for 25% missed detections for subjective definition at shot level. For runs at segment level and for the objective scenario, performances vary from one system to another and achieve at best 40% false alarms for 25% missed detections. Recall/precision curves show that, regardless of the scenario, all systems reach high recall values (which corresponds to the targeted operating point, where one does not want to miss any violent scene) at the expense of very low precision values (between 0.1 and 0.2). The relative rarety of the events to detect in the dataset (8.28% of the duration for objective and 13.91% for subjective) partly explains these values. Last, it should be noted that the ranking of the best performing systems slightly changes while considering recall/precision or false and missed detections curves, compared to the official ranking based on MAP@100.

Conclusions and future work

We developed the 2013 MediaEval Affect Task: Violent Scenes Detection and introduced a common benchmarking framework for evaluation that included the formalization of two different perspectives of the concept of violence (objective and subjective). Judging from the results, we believe that the proposed task stands as a consistent and standardized benchmarking framework for violence detection in

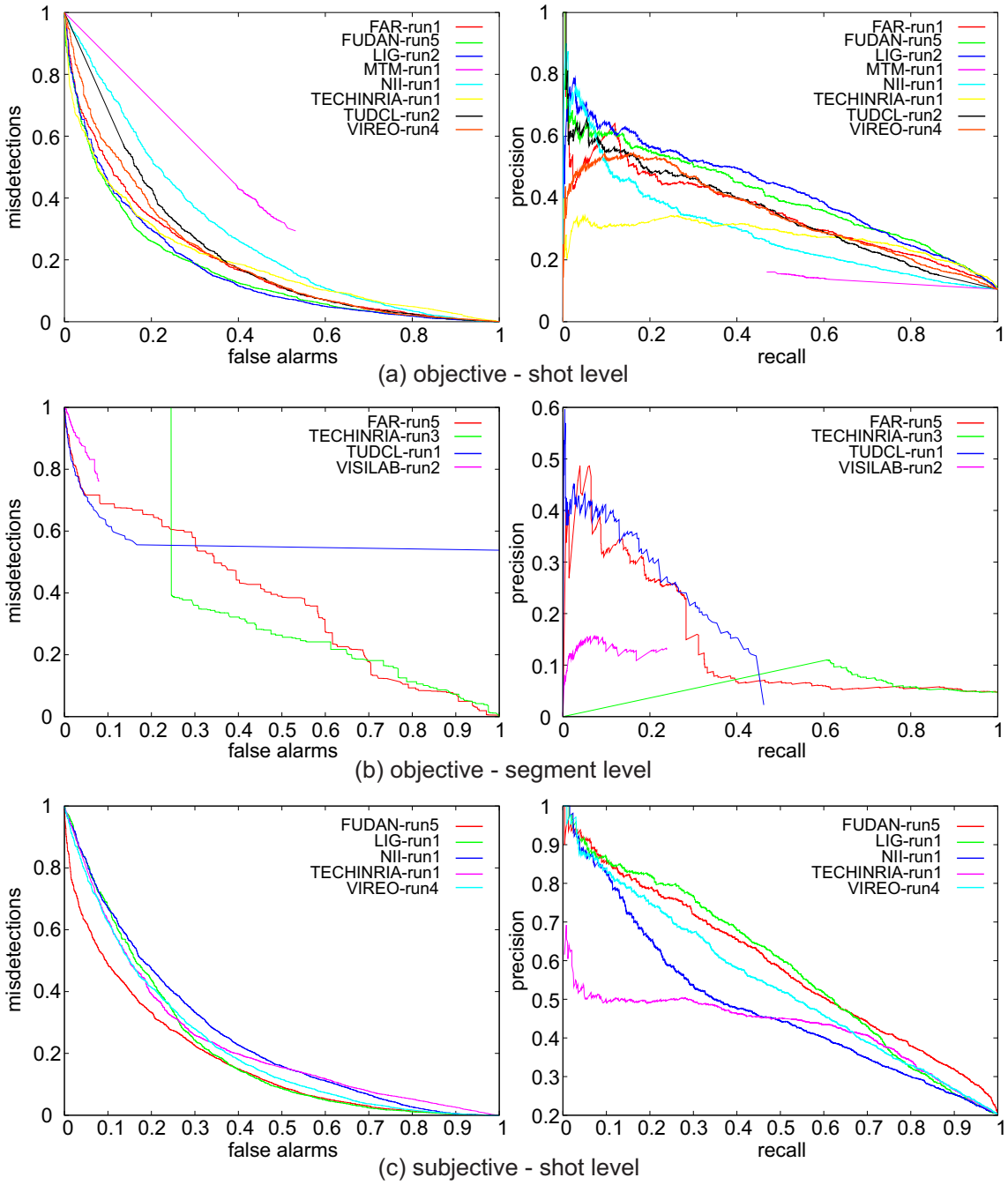


Figure 6.20: Misdetections/false alarms and precision/recall curves for best participant runs at 2013 MediaEval Affect Task: Violent Scenes Detection [100] (see also Table 6.13).

movies. Its publicly available annotated dataset provides a relevant testbed for the evaluation of a broad category of multimodal approaches.

Several perspectives may be drawn for future extensions of this framework: we should head towards a qualitative evaluation, in addition to the quantitative metrics currently used; we should investigate strategies to expand the task to other types of video material, e.g., user-generated content, and see how the proposed systems generalize to different types of content; we will continue promoting multimodal approaches to the task by encouraging participants to use metadata (e.g., from the Internet) as a complement to the classic audiovisual features.

6.6 Search result diversification

Multimedia, such as videos and images, make for an important share of the data distributed and searched for on the Internet. Current photo search technology is mainly relying on employing text annotations, visual, or more recently on GPS information to provide users with accurate results for their queries.

Retrieval capabilities are however still below the actual needs of the common user, mainly due to the limitations of the content descriptors, e.g., text tags tend to be inaccurate (e.g., people may tag entire collections with a unique tag) and annotation might have been done with a goal in mind that is different from the searchers goals. Automatically extracted visual descriptors often fail to provide high-level understanding of the scene [110] while GPS coordinates capture the position of the photographer and not necessarily the position of the query and they can again be assigned for a large set of images regardless of exact positions.

Until recently, research focused mainly on improving the *relevance* of the results. However, an efficient information retrieval system should be able to *summarize* search results and give a global view so that it surfaces results that are both relevant and that are covering *different aspects* of a query, e.g., providing different views of a monument rather than duplicates of the same perspective showing almost identical images. Relevance was more thoroughly studied in existing literature than diversification [111, 112, 113] and even though a considerable amount of diversification literature exists (mainly in the text-retrieval community), the topic becomes more and more important, especially in multimedia [114, 115, 116, 117].

6.6.1 Mechine-crowd diversification

I have contributed to the development of a hybrid image search diversification approach that takes advantage of both machine and human computational power. Machine computation allows for a fast pre-processing of the large input data whereas

the final refinement of the results is performed by humans via crowd-sourcing[14].

Contribution to state-of-the-art

Research on automatic media analysis techniques reached the point where further improvement of retrieval performance requires the use of user expertise. We proposed a hybrid machine-human approach [17, 6] that acts as a top layer in the retrieval chain of current social media platforms.

The proposed approach goes beyond the state-of-the-art in the following directions: although the issue of results diversification was already studied in the literature, we introduce a new re-ranking scheme that allows for better selection of both relevant and diverse results; we provide a crowd-study of the diversification task which has not been yet addressed in the literature and experimentally assesses the reliability of the crowdsourcing studies for this particular task; we study the perspective and the efficiency of including humans in the computational chain by proposing a unified machine-crowd diversification approach where machine plays the role of reducing the time and resource consumption and crowdsourcing filters high quality diverse and representative images.

Approach

The proposed approach involves the following steps:

- **machine retrieval**: photos are retrieved using best current retrieval technology, e.g., using text and GPS tags on current social media platforms. These results are numerous (hundreds) and typically contain noisy and redundant information;

- **machine media analysis**: automated machine analysis is used to filter the results and reduce the time and resource consumption to allow better crowdsourcing. We designed a new approach that re-ranks the results according to

[14]this work was developed in collaboration with Anca-Livia Radu, from LAPI, University Politehnica of Bucharest, Romania, Dr. Maria Menéndez, Dr. Julian Stöttinger, Prof. Fausto Giunchiglia, Assoc. Prof. Antonella De Angeli, from the Department of Information Engineering and Computer Science, University of Trento, Italy. The presented results were published in:

[6] A.-L. Radu, *B. Ionescu*, M. Menéndez, J. Stöttinger, F. Giunchiglia, A. De Angeli, "*A Hybrid Machine-Crowd Approach to Photo Retrieval Result Diversification*", 20th International Conference on MultiMedia Modeling - MMM2014, 8-10 January, Dublin, Ireland, 2014.

[17] A.-L. Radu, J. Stöttinger, *B. Ionescu*, M. Menéndez, F. Giunchiglia, "*Representativeness and Diversity in Photos via Crowd-Sourced Media Analysis*", 10th International Workshop on Adaptive Multimedia Retrieval - AMR 2012, October 24-25, Copenhagen, Denmark, 2012.

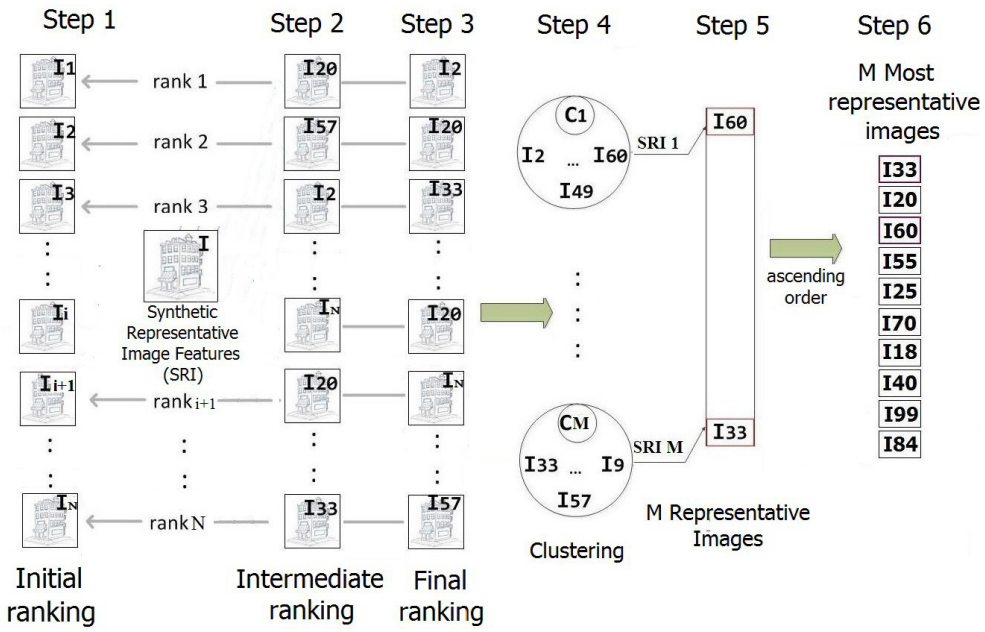


Figure 6.21: Proposed re-ranking approach [6] (I represents an image and C denotes the clusters).

the similarity to the "most common" image in the set for improving representativeness followed by a clustering mechanism that is specifically designed to ensure the diversification by retaining only the best ranked representative images in each cluster;

- **crowdsourcing analysis**: to bridge further inherent machine semantic gap, crowdsourcing is used as a final refinement step for selecting high quality diverse and representative images. To cope with the inherent crowd reliability problem, we designed two adapted studies for both representativeness and diversification. The final objective is to summarize the query with very few results which corresponds to the typical user scenario that browses only the top results.

Machine processing. The proposed method diagram is presented in Figure 6.21. It involves the following steps:

- *step 1*: for each of the N images in the initial noisy image set, S, we describe the underlying information using visual content descriptors. Then, we determine using the features of all images a Synthetic Representative Image Feature (SRI) by taking the average of the Euclidean distances between all image features;

- *step 2*: for each image, I_i, $i = 1, ..., N$, we compute the average of the Euclidean distances to the rest of the images in S which leads to a N-dimensional array. The value of SRI is subtracted from the array which is sorted in ascending order. The new position of each value will account for the intermediate new rank of its corresponding image;

- *step 3*: supposing that most representative images are among the first returned, for the final re-ranking we average the two ranks (initial rank and the intermediate one determined in step 2) which yields N average values. Values are again sorted in ascending order and images are re-arranged accordingly;

- *step 4*: all re-ranked images are clustered in M clusters using a k-means approach based on their visual content. Preliminary tests returns good performance for M around 30;

- *step 5*: a SRI_j, $j = 1, ..., M$, is computed as presented in step 1 and a new re-ranking is performed by re-iterating steps 1, 2 and 3 over the set of images inside each cluster, C_j. Each cluster's first ranked image is considered to be representative (denoted RI_j);

- *step 6*: from all RI_j images, we select only a small set of P ($P << M$) highest ranked images (ranking according to the final rank computed in step 3). This will ensure also the diversification of the results.

Crowdsourcing processing. To bridge further the inherent semantic gap of automatic machine analysis techniques, we use a crowdsourcing approach. We designed two studies that are adapted to our diversification task, where crowd is involved to refine and improve machine-analysis results with the final goal of determining a high quality set of representative and diverse images. Low monetary cost, reduced annotation time and results close to expert-based approaches [118] are some of the reported advantages of crowdsourcing that made it very suitable for solving multimedia tasks. However, crowdsourcing is not a perfect system; not every task can be crowdsourced and quality control is usually an issue [119].

For the experimentation we used the Crowdflower[15] meta crowdsourcing platform which uses an automatic quality control based on gold units (i.e., for avoiding un-relevant user input such as random responses or computer bots). Gold units are unambiguous questions for which an answer is provided by the requester. Contributors need to answer at least 4 gold questions with a minimum 70% accuracy to get their answers included in the results. In Crowdflower requesters can create jobs,

[15] `http://crowdflower.com/`.

which consist of a data file and units. Units contain the tasks to be performed and are instantiated using the data file. Before ordering, requesters can calibrate the number of judgements per unit, number of judgements per page, and worker pay per page.

The first designed study addressed the representativeness of retrieval results - experimentation was conducted on a search for representative and diverse images with monuments/locations (which stood as validation dataset for our approach). The representativeness task collected data on the variable locations (i.e., indoor, outdoor), relevance, and representativeness. In this study, relevant pictures contain, partially or entirely, the query monument. Representative images are prototypical outside views of a monument. Relevance is an objective concept indicating the presence/absence of the monument, or part of it, while representativeness is a more subjective concept that might depend on visual context or personal perception. The task was divided in two parts: first, participants were familiarized with the task and provided with a contextualized visual example of the monument query (Wikipedia entry of the monument). Secondly, contributors were asked to answer questions on location, relevance, and representativeness for a given retrieval result.

The second study addressed the diversification of the results. The diversification task collected data on perceived diversity among the set of representative pictures of the same monument that were provided by the representativeness study above. Participants assessed visual variation considering the use case scenario of constituting a monument photo album. As for the previous task, first participants were given an introduction to the task accompanied with some visual examples including a link to the Wikipedia page of the monument. Afterwards, contributors were asked to answer whether they would include the provided pictures in the photo album.

Validation results

To validate our approach we use a data set of more than 25,000 images depicting 94 Italian monument locations, from very famous ones (e.g., "Verona Arena") to lesser known to the grand public (e.g., "Basilica of San Zeno"[16]). Images were retrieved from Picasa[17], Flickr[18] and Panoramio[19] using both the name of the monument and GPS tags (with a certain radius). For each monument, when available, we retain the first 100 retrieved images per search engine, thus around 300 images in total per monument. To serve as ground truth for validation, each of the images was

[16]data set is available at `http://www.cubrikproject.eu`, FP7 CUbRIK project.
[17]`http://picasaweb.google.com/`.
[18]`http://flickr.com/`.
[19]`http://panoramio.com`.

Table 6.14: Average precision for various descriptor combinations of the proposed re-ranking approach [6].

autocorr.	CCV	hist.	c.layout	c.struct	**c.moment**	edgehist.	LBP	LTP
57.70%	58.80%	57.98%	57.70%	60.00%	**60.80%**	56.67%	56.58%	57.68%
Harris	STAR	MSER	SURF	GOOD	c.struct & GOOD	c.struct & HoG	c.moment &GOOD	all color desc.
57.84%	58.87%	59.17%	59.18%	60.39%	60.59%	59.02%	59.71%	58.43%
c.n.hist	HoG	all color desc. & GOOD						
58.33%	58.33%	58.73%						

manually labeled (as being representative or not) by several experts with extensive knowledge of monument locations.

Validation of machine media analysis. To assess performance, we use the classic precision as defined in equation 6.5. The first test consisted on determining the influence of the content descriptors on the precision of the results. We experimented with various state-of-the-art approaches:

- *MPEG-7 and related texture and color information*: we compute color autocorrelogram (autocorr.), Color Coherence Vector (CCV), color histogram (hist.), color layout (c.layout), color structure (c.struct), color moments (c.moment), edge histogram (edgehist.), Local Binary Patterns (LBP), Local Ternary Patterns (LTP) and color histogram in [89] (c.n.hist);

- *feature descriptors* that consist of Bag-of-Visual-Words representations of Histogram of oriented Gradients (HoG), Harris corner detector (Harris), STAR features, Maximally Stable Extremal Regions (MSER), Speeded Up Robust Feature (SURF) and Good Features to Track (GOOD). We use 4,000 words dictionaries, experimentally determined after testing different dictionary sizes with various descriptors.

Table 6.14 summarizes some of the results (we report the global average precision over all the results of the 94 monument queries; descriptors are fused using early fusion). Most interesting is that for this particular set-up, very low complexity texture descriptors are able to provide results very close to the use of much more complex feature point representations. Therefore, for a higher computational speed, one may go along with a simpler approach without loosing performance. The highest precision is provided by color moments, 60.8%, which are further employed for the machine-crowd integration.

Table 6.15: Performance comparison of the re-ranking [6] with other approaches (average precision).

proposed	method in [120]	*Picasa*	*Flickr*	*Panoramio*
60.8%	46.8%	39.47%	53.51%	39.85%

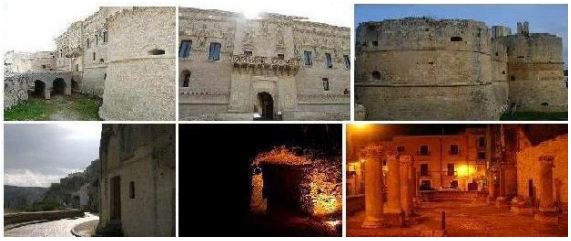
(a) - initial retrieval (first 10 images)

(b) - results after machine analysis

(c) - final results after crowd sourcing

(d) - initial retrieval (first 10 images)

(e) - results after machine analysis



(f) - final results after crowd sourcing

Figure 6.22: Output example for the proposed approach [6]: (a)(b)(c) results for query "*Rialto Bridge*" and (d)(e)(f) results for query "*Castle of Otranto*" (image sources: reference Wikipedia (image in red rectangle), others Picasa, Flickr and Panoramio).

Another interesting result is the limited representative power provided by automatic content descriptors in this context. Regardless the feature used, the disparity of the performance is within a [56%;60%] precision interval. This basically shows the limitation of automated machine analysis and the need for addressing other information sources. Nevertheless, results achieved by media analysis show significant improvement over the initial retrieval.

Table 6.15 compares these results against the initial retrieval given by the three image search engines and an approach from the literature. The average improvement

in precision over initial retrieval is more than 16%. In addition, we achieve an improvement of more than 23% over the approach in [120], that is very promising.

To have a subjective measure of performance, several example results are illustrated in Figure 6.22 (for visualization reasons, we also display a Wikipedia picture of the query - depicted with the red rectangle). One may observe that compared to the initial images, the automatic machine analysis allows for significant improvement of representativeness and diversity. However, not all the results are perfect, the limited representative power of content descriptors may lead to misclassification, e.g., in Figure 6.22.(b) some of the initial duplicates are filtered - Figure 6.22.(a) images 4 and 8, but near-duplicates may appear in the final results, e.g., Figure 6.22.(b) images 2 and 5.

Validation of crowdsourcing analysis. The representative study was conducted between 15th and 21st November 2012. Each unit contained one image, which was judged by at least three contributors. Contributors earned 0.07$ per unit. Contributors from 20 different countries located in diverse world regions were allowed to access the job. In total, 228 contributors judged 5,377 units. Due to platform's quality control method, 18% of the units were annotated more than three times. Contributors performed an average of 23.7 units (SD=25.5) with a minimum of 8 and maximum of 174 units. Most of the contributors were located in India (21%), Indonesia (18%), USA (15%), Germany (14%), Canada (9%), Italy (8%), and Morocco (6%).

Reliability analysis was calculated using Kappa statistics that measure the level of agreement among annotators discarding agreement given by chance (values from -1 to 1) [121]. As general guideline, Kappa values higher than 0.6 are considered adequate and above 0.8 are considered almost perfect [118]. For this study, fixed marginal multirater Kappa was used [121]. This variation of Cohen's Kappa is indicated when the number of categories is fixed and there are more than two raters. For analysis purposes, only images annotated exactly three times were considered. In total, 749 images were considered in the reliability analysis. Reliability among annotations achieve a Kappa value of 0.7 for location, 0.44 for relevance and 0.32 for representativeness.

In order to aggregate crowd-sourcing results, the average value among contributors' judgements for location, relevance and representativeness was calculated. First, categorical values (i.e., "Indoor", "Outdoor", "yes", "no") were coded into binary values. "Indoor" judgements for the variable location were coded with 1 and "Outdoor" judgements were coded with 0. For the variables relevance and representativeness, pictures annotated with "yes" were coded with 1 and those annotated with "no" took the value 0. Average values were calculated per image and variable, and

mapped onto a binary distribution. For location, pictures with average value equal or above 0.5 were coded as indoor locations. For relevance and representativeness, pictures with average values equal or higher than 0.5 were considered both as related and representative.

Averaged results indicate that 81% of the images depict an outdoor location, 63% contain at least a part of the monument and 57% are representative. As the definition of representativeness used in this paper implies that a picture is representative if it depicts an outdoor scenario, distributions were also calculated considering just images annotated as outdoor. In this case, results show that the percentage of images containing the monument is up to 66%, the percentage of representative pictures is 60%.

Images judged as representative were included in the diversity task, grouped by monument. The diversity study was conducted on November the 26th. In total, 499 images grouped in 82 units (i.e., monuments) were annotated. As the estimated time per unit was similar as in the representativeness study, contributors also earned 0.07$ per unit. Units were judged by at least three contributors. In total, 62 contributors participated in 262 units. Number of performed units per contributor varied from 2 to 24 units. Most of the contributors were located in Indonesia (30%), Italy (26%), India (24%), Germany (5%), Brazil (4%), and United Kingdom (4%). Results were averaged and mapped onto a binary distribution using a similar procedure as in the representativeness study.

Aggregated averaged results indicate that: 48% of the monument-grouped images contain all representative and diverse images; some 73% of the grouped images contain at least 75% representative and diverse images, 90% of the group images contain at least 50% representative and diverse images.

Crowdsourcing is a promising approach for human validation. The results of this study support existing research which identify low costs and reduced annotation time as the advantages of using crowdsourcing. However crowdsourcing is not a perfect system, issues such as quality control and reliability of results need further investigation. Automatic quality control methods, as the gold units used in this study, can be a good option since data cleaning is a time-consuming and tedious task. However, they do not always ensure high quality data; since answers can only be rejected at runtime, they may attract more spammers, malicious, and sloppy workers [119].

Reliability of results may vary depending on the level of subjectiveness of the measured variable. For example, annotations for the variable location (i.e., indoor or outdoor) achieve an adequate level of reliability, while the reliability of the annotations for representativeness is quite low. These results suggest that representative-

ness is a concept subject of individual variations (e.g., visual perception, previous experiences, level of expertise), image features (e.g., perspective, color, and scene composition), or monuments' features (e.g., popularity, distinguishable features, and kind of monument). These issues should be further investigated and considered in the development of methods for aggregation of crowd-sourcing results, since current aggregation methods may underestimate the value of heterogeneous answers [122].

Validation of the machine-crowd. The final experiment consisted in analyzing the performance of the whole machine-crowd chain. The overall average precision after the crowd step is up to 78% which is an improvement over the machine analysis and initial retrieval results (see Table 6.15). Several examples are depicted in Figures 6.22.(c) and 6.22.(f). In general, if enough representative pictures are provided after the machine analysis, the crowdsourcing step allows for increasing the diversity among them; while for the case when not enough representative pictures are available, the crowdsourcing tends to increase the relevance (this is usually due to the fact that these pictures are already highly diverse, but not that relevant).

Conclusions and future work

We addressed the problem of enforcing representativeness and diversity in noisy images sets retrieved from common social image search engines. We introduced a hybrid approach that combines a pre-filtering step carried out with an automated machine analysis with a crowdsourcing study for final refinement. The motivation of using the media analysis is also to reduce the workload in the crowdsourcing tasks for enabling better results.

Experimental validation was conducted on more than 25,000 photos in the context of the retrieval of photos with monuments. Results show that automatic media analysis reached the point where further performance improvement requires the use of human intelligence, since regardless the image descriptors use, they are limited to reach only up to 60% precision. Instead, the further use of crowdsourcing led to an improving of both representativeness and diversity in the final results. Thanks to the initial diversification of the results, after crowdsourcing some 73% of the grouped images contained a promising number of at least 75% representative and diverse images.

Overall, the use of machine and human validation allows for better performance than using solely the automated media analysis. The fact that crowdsourcing acts like a human computational machine allows its integration in the processing chain. Depending on the complexity of the content descriptors (e.g., Bag-of-Visual-Words), crowdsourcing may yield faster response than the machine analysis, but it is limited in the accuracy of the results for large data sets and therefore running it directly on

the initial data is not efficient. Crowdsourcing is not capable of providing perfect result, despite the use of human expertise. The main reason is that low monetary costs attracts people with limited experience to the task and results may be variable.

Future work will mainly consist on adapting the proposed approach to the large scale media analysis constraints.

6.6.2 Benchmarking search results diversification

Another contribution is the introduction and development of a new evaluation framework and dataset within the 2013 MediaEval Retrieving Diverse Social Images Task [123], designed to support the emerging area of information retrieval that fosters new technology for improving both the relevance and diversification of search results[20].

Contribution to state-of-the-art

Besides the scientific challenge, another critical point of the diversification approaches are the evaluation tools. In general, experimental validation is carried out on very particular and closed datasets, which limits the reproducibility of the results. Another weakness are the ground truth annotations that tend to be restrained and not enough attention is paid to their statistical significance and consequently to the statistical significance of the entire evaluation framework. A solution are the benchmarking activities that provide a framework for evaluating systems on a shared dataset and using a set of common rules. The results obtained are thus comparable and a wider community can benefit from it.

We designed and developed a new benchmarking framework with the 2013 MediaEval Retrieving Diverse Social Images Task [123, 7] and its dataset [1] (Div400), which focus on fostering new technology for improving both relevance and diversification of search results with explicit emphasis on the actual social media context.

[20]this work was developed in collaboration with Anca-Livia Radu, from LAPI, University Politehnica of Bucharest, Romania, Dr. Maria Menéndez, from the Department of Information Engineering and Computer Science, University of Trento, Italy, Prof. Henning Müller, from University of Applied Sciences Western Switzerland, Sierre, Switzerland, Dr. Adrian Popescu, from CEA-LIST, France, and Babak Loni, from Multimedia Signal Processing Group, Delft University of Technology, The Netherlands. The presented results were published in:

[1] *B. Ionescu*, A.-L. Radu, M. Menéndez, H. Müller, A. Popescu, B. Loni, "*Div400: A Social Image Retrieval Result Diversification Dataset*", ACM Multimedia Systems - MMSys, 19-21 March, Singapore, 2014.

[7] *B. Ionescu*, A. Popescu, H. Müller, M. Menéndez, A.-L. Radu, "*Benchmarking Result Diversification in Social Image Retrieval*", IEEE International Conference on Image Processing - ICIP 2014, 27-30 October, Paris, France, 2014.

These two characteristics of retrieval results are antinomic, i.e., the improvement of one of them usually results in a degradation of the other.

In the context of the current state-of-the-art we identify the following main contributions: the evaluation framework focuses on improving the current technology by using Flickr's relevance system as reference[21] (i.e., one of the state-of-the-art platforms) and addresses in particular the social dimension reflected in the nature of the data and methods devised to account for it; while smaller in size than the ImageCLEF collections [115, 124], the proposed dataset contains images that are already associated to topics by Flickr. This design choice ensures that there are many relevant images for all topics and pushes diversification into priority; unlike ImageCLEF, which worked with generic ad-hoc retrieval scenarios, a focused real-world usage scenario is set up, i.e., tourism, to disambiguate the diversification need; finally, a comparison of expert and crowdsourced ground truth production is performed to assess the potential differences between lab and real life evaluations.

Dataset

Given the important social role of geographic queries and their spatio-temporal invariance, the dataset was constructed using a retrieval of photos with landmark locations scenario. For 396 locations, up to 150 photos (with Creative Commons redistributable licenses) and associated metadata are retrieved from Flickr and ranked with Flickr's default "relevance" algorithm.

To compare different retrieval mechanisms, data was collected with both textual (i.e., location name - *keywords*) and GPS queries (*keywordsGPS*). Location metadata consists of Wikipedia links to location webpages and GPS information and photo metadata includes social data, e.g., author title and description, user tags, geotagging information, time/date of the photo, owner's name, the number of times the photo has been displayed, number of posted comments, rank, etc.

Apart from these data, to support contributions from different communities, some general purpose content descriptors are provided for the photos - visual descriptors, e.g., Histogram of oriented Gradients, Local Binary Patterns, MPEG-7 related features, etc; and textual models, e.g., probabilistic models, Term Frequency-Inverse Document Frequency (TF-IDF) weighting and social TF-IDF weighting (an adaptation to the social space) [125].

The dataset provides 43,418 photos and is divided into a *devset* of 50 locations (5,118 photos, in average 102.4/location) intended for training and a *testset* of 346

[21] `http://www.flickr.com/services/api/`.

Table 6.16: Expert and crowd annotation statistics for Div400 [1].

devset (expert)	*testset* (expert)	*testset* (crowd)
relevance (annotations - avg.Kappa - % relevant img.)		
6(6) - 0.64 - 73	3(7) - 0.8 - 65	3(175) - 0.36 - 69
diversity (annotations - avg.clusters/location - avg.img./cluster)		
1(3) - 11.6 - 6.4	1(4) - 13.1 - 5	3(33) - 4.7 - 32.5

locations (38,300 photos, in average 110.7/location) for evaluation[22].

Data are annotated for both relevance and diversity of the photos. The following definitions were adopted:

- **relevance** - a photo is relevant if it is a common photo representation of the location, e.g., different views at different times of the day/year and under different weather conditions, inside views, creative views, etc, which contain partially or entirely the target location (bad quality photos are considered irrelevant) - photos are tagged as relevant, non-relevant or with "don't know";

- **diversity** - a set of photos is considered to be diverse if it depicts complementary visual characteristics of the target location (e.g., most of the perceived visual information is different - relevant photos are clustered into visually similar groups).

Annotations were carried out mainly by experts with advanced knowledge of location characteristics. To explore differences between experts and non-experts annotations, a subset of 50 locations from the *testset* was annotated using crowdworkers (via the CrowdFlower[23] platform). In all cases, visual tools were employed to facilitate the process. Annotations were carried out by several annotators and final ground truth was determined after a lenient majority voting scheme. Table 6.16 presents the number of distinct annotations (the number of annotators is depicted in the brackets), Kappa inter-annotator agrement (*devset* reports weighted Kappa [127], *testset* reports Free-Marginal Multirater Fleiss' Kappa [126] as different parts of the data are annotated by different annotators) and cluster statistics. Expert annotations achieved a good agreement as average Kappa is above 0.6 and up to 0.8 (values above 0.6 are considered adequate and above 0.8 are considered almost perfect [127]). Only 0.04% of the photos achieved "don't know" answers. The

[22]the dataset was made publicly available [1] and can be downloaded at `http://traces.cs.umass.edu/index.php/mmsys/mmsys`.

[23]`http://crowdflower.com/`

diversity annotations lead to an average of around 12 clusters/location and 5-6 images/cluster. For the crowd annotations, the agreement is significantly lower, namely 0.36, and up to 1% of the photos achieved "don't know" answers, which reflects the variable backgrounds of the crowd (on average it leads to 4.7 clusters/location and 32.5 images/cluster).

Benchmarking results

The proposed benchmarking framework was validated during the 2013 MediaEval Retrieving Diverse Social Images Task [123]. The task required participants developing techniques that allow the refinement of the initial Flickr retrieval results by selecting a ranked list of up to 50 photos that are equally relevant and diverse representations of the query. During the competition, participants designed and trained their methods on the *devset* dataset (50 locations and 5,118 photos) while the actual benchmarking was conducted on the *testset* (346 locations and 38,300 photos).

Participants were allowed to submit the following types of runs: automated techniques that use only visual information (*run1*), automated techniques that use only text information (*run2*), automated techniques that use multimodal information fused without other resources than provided (*run3*), human-based or hybrid human-machine approaches (*run4*) and a general run where everything was allowed including using data from external sources like the Internet (*run5*). In total, the task received 38 runs from 11 participant teams.

Performance is assessed for both diversity and relevance. The main evaluation metric is Cluster Recall at X ($CR@X$) [115], defined as:

$$CR@X = \frac{N}{N_{gt}} \tag{6.25}$$

where N is the number of image clusters represented in the first X ranked images and N_{gt} is the total number of image clusters from the ground truth (N_{gt} is limited to a maximum of 20 clusters from the annotation process). Defined this way, $CR@X$ assesses how many clusters from the ground truth are represented among the top X results provided by the retrieval system. Since clusters are made up of relevant photos only, relevance of the top X results is implicitly measured by $CR@X$, along with diversity.

To get a clearer view of relevance, Precision at X ($P@X$) is also used as a secondary metric and defined as:

$$P@X = \frac{N_r}{X} \tag{6.26}$$

where N_r is the number of relevant pictures from the first X ranked results. Therefore, $P@X$ measures the number of relevant photos among the top X results. To account for an overall assessment of both diversity and precision, $F1@X$ was also reported which is the harmonic mean of $CR@X$ and $P@X$.

Evaluation was conducted for different cutoff points, X$\in${5,10,20,30,40,50}. In particular, submitted systems were optimized with respect to CR@10 (i.e., for 10 images returned) which was the official metric. CR@10 was chosen because it ensures a good approximation of the number of photos displayed on different types of screens and also in order to fit the characteristics of the dataset (at most 150 images and 20 clusters per location). It is worth mentioning that given the definition in equation 6.25, $CR@10$ is inherently limited to a highest possible value of 0.77, as on average the dataset has 13 clusters per location (see Table 6.16). We report the average values over all the locations in the dataset.

Keywords vs. keywords and GPS retrieval. Retrieval with GPS information yields more accurate results than using solely keywords, e.g., for the initial Flickr results, P@10 with keywords is 70.45% compared to 78.81% using GPS data in addition. Diversity is however slightly higher for keywords, CR@10 is 39.85% compared to 34.37% using GPS as results are sparse. In the following we focus on presenting the average overall results.

Evaluation per modality. Figure 6.23 plots overall precision against recall averages for all participant runs (38) at a cutoff at 10. For *visual approaches*, highest diversification is achieved with a Greedy optimization of VLAD+SURF descriptors, CR@10=42.91% - SocSens run1 [133], while lowest diversification is provided by a matching graph approach also with feature point information (RootSIFT and Hessian), CR@10=29.21% - ARTEMIS [138]. Using simple color information (histograms and detection of faces) still achieves high recall, e.g., CR@10=40.76% - BMEMTM run1 [132], which proves that there is no superiority of classes of descriptors, the difference in performance being mainly related to the method.

Compared to visual, *text information* tends to provide better results (see green points). Highest diversification is achieved using a re-ranking with Lucene and Greedy Min-Max optimization, CR@10=43.06% - SOTON-WAIS run2 [128] where data are represented with time-related information. On the other end, bag-of-words of TF-IDF data and web inspired ranking leads to CR@10=35.79% - UEC run2 [136]. Surprisingly, *human approaches* were less effective than the automatic ones as users tend to maximize precision at the cost of diversity, e.g., BMEMTM run4 [132] achieves P@10=89.36% but CR@10 is only 29.63%. However, human-machine integration improves also the diversity, e.g., CR@10=40.48% - SocSens run4 [133]. Overall, the best performing approach is *multimodal*, namely CR@10=43.98% -

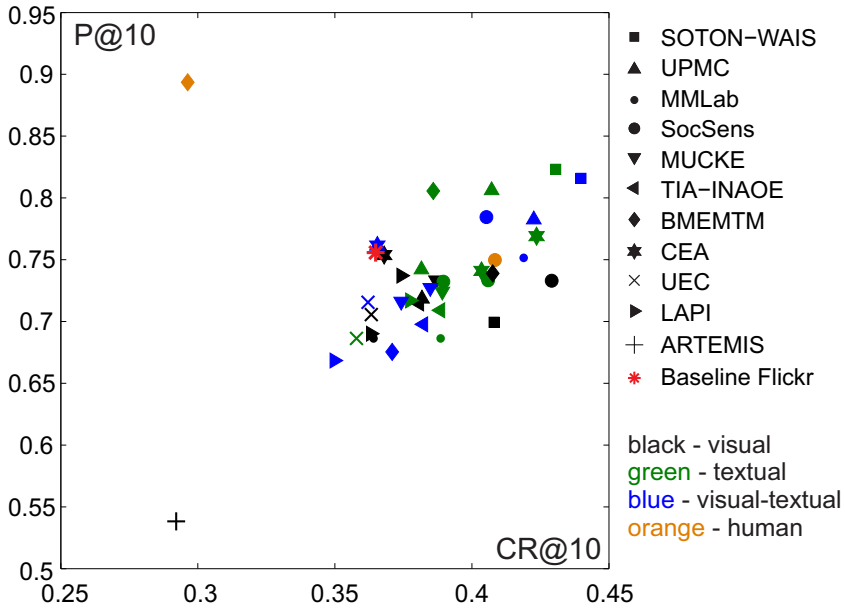


Figure 6.23: Precision vs. cluster recall averages at 10 images at the 2013 MediaEval Retrieving Diverse Social Images Task [123].

Table 6.17: Precision and cluster recall averages (%) for best team runs at the 2013 MediaEval Retrieving Diverse Social Images Task [123].

team best run	P@10	P@20	P@30	P@40	P@50	**CR@10**	CR@20	CR@30	CR@40	CR@50
SOTON-WAIS run3 [128]	**81.58**	**77.88**	74.14	70.59	66.62	**43.98**	61.97	72.16	78.44	82.43
SocSens run1 [133]	73.3	74.87	**76.03**	71.45	59.15	42.91	**63.14**	72.28	74.73	74.84
CEA run2 [137]	76.9	76.39	75.65	**74.09**	**71.53**	42.36	62.49	73.46	81.48	**86.68**
UPMC run3 [129]	78.25	73	72.54	70.99	68.91	42.26	62.68	74.7	81.54	85.4
MMLab run3 [134]	75.15	74.04	73.35	71.85	69.7	41.89	62.36	**74.92**	**82.05**	86.53
BMEMTM run1 [132]	73.89	71.64	71.82	71.15	69.27	40.76	61.39	71.84	79.35	84.4
MUCKE run2 [135]	72.43	72.28	71.83	70.8	68.84	38.92	57.49	68.77	76.84	83.06
TIA-INAOE run2 [130]	70.91	71.36	71.46	70.45	68.51	38.85	57.32	68.97	77.19	82.28
LAPI run2 [131]	71.7	71.11	68.96	64.77	57.95	37.74	57.34	68.2	74.72	77.22
baseline Flickr	75.58	72.89	71.94	70.8	68.77	36.49	53.46	65.58	74.11	79.88
UEC run1 [136]	70.56	70.92	70.76	69.48	67.52	36.33	54.48	67.43	75.72	81.54
ARTEMIS run1 [138]	53.83	33.79	22.69	17.02	13.61	29.21	33.06	33.1	33.1	33.1

SOTON-WAIS run3 [128], it improves diversification of the state-of-the-art Flickr initial results with at least one additional image class.

Table 6.17 presents the official ranking of the best team approaches for various cutoff points (highest values are in bold). In addition to the information from Figure

Table 6.18: Expert vs. crowd annotations - precision and cluster recall averages (%) for team best runs at 2013 MediaEval Retrieving Diverse Social Images Task [123].

team best run	*expert GT*		*crowd GT*	
	P@10	CR@10	P@10	CR@10
SOTON-WAIS run3 [128]	**87.55**	41.29	**77.14**	74.5
SocSens run1 [133]	79.59	41.39	72.86	76.36
CEA run2 [137]	82.65	40.81	70.82	72.87
UPMC run3 [129]	84.08	**41.51**	74.9	**78.8**
baseline Flickr	79.80	33.45	68.16	66.43

6.23, Table 6.17 shows that the precision tends to decrease with the increase of the number of images as it is more likely to obtain non-relevant pictures. On the other hand, cluster recall increases with the number of pictures as it is more likely to get pictures from additional classes.

Ranking stability analysis. To determine the statistical significance of the results and thus to examine the relevance of the dataset, a stability test was run [124]. Stability is examined by varying the number of topics which is used to compute performance. Stability tests are run with different topic subset sizes, which are compared to the results obtained with the full testset (346 topics). For each topic subset, 100 random topic samplings are performed to obtain stable averages. Spearman's rank correlation coefficient [139] is used to compare the obtained CR@10 rankings and the obtained values are 0.61, 0.86, 0.93, 0.96, 0.97, 0.98, 0.99 for subsets which contain 10, 50, 100, 150, 200, 250 and 300 topics. These results show that there is little change in the ranking when at least 100 topics are used. The size of the testset is clearly sufficient to ensure statistical significance of the ranking and therefore of the results.

Expert vs. crowd annotations. Performance assessment depends on the subjectivity of the ground truth, especially for the diversification part. The final experiment consisted of comparing both results achieved with expert and crowd annotations. Table 6.18 presents the four best team runs (highest results are depicted in bold; results on a selection of 50 locations from testset). Although precision remains more or less similar in both cases, cluster recall is significantly higher for the crowd annotations. This is due to the fact that workers tend to under-cluster the images for time reasons. Nevertheless, what is interesting is the fact that regardless of the ground truth, the improvement of the baseline is basically the same: 7.84% for experts compared to 8.07% for the crowd, which shows that results are simply translated but the relevance is still the same. Crowd annotations are an attractive

alternative to expert annotations, being fast - order of hours compared to expert ones that require weeks - while the performance is similar.

Conclusions and future work

We designed and developed a benchmarking framework for results diversification of social image retrieval and validate it during the MediaEval 2013 campaign. The strong participation in a first year (24 teams registered and 11 crossed the finish line) shows the strong interest of the research community in the topic. Similar to the strong impact of other evaluation campaigns in multimedia retrieval [140, 141, 142] an important impact can also be expected from this task. Several groups increased specific aspects of the results on the strong Flickr baseline, particularly linked to diversity. Approaches combining a large variety of modalities from manual re-ranking, GPS to visual and text attributes have the potential to improve results quality and adapt to what users may really want to obtain as results, which can be situation-dependent. Detecting objects such as faces was also used. Via the analysis of the clusters of relevant images, several categories can likely be deduced and used in connection with detectors for these aspects to optimize results. The crowdsourcing part of the relevance judgments is clearly an option as the results described in the paper show. There are differences in the results but the effort to cost ratio is an important part and crowdsourcing can likely help to create much larger resources with a limited funding. Strict quality control seems necessary to assure the crowdsources quality and this can likely also help to obtain better results.

For a continuation of the evaluation campaign it seems important to look into criteria that can stronger discriminate the runs, so basically making the task harder. More clusters are an option. A larger collection is also an option but creating diversity ground truth for large collections is tedious and expensive. Crowdsourcing could be a valid approach also for this, as the experiments show.

6.6.3 User tagging credibility estimation

Another novel contribution to image search result diversification was in developing an approach for user tagging credibility estimation of photos to improve both relevance and diversity of the retrieval[24].

[24]this work was developed in collaboration with Alexandru Gînscă, Dr. Adrian Popescu, from CEA-LIST, France, Dr. Anil Armagan, from Bilkent University, Ankara, Turkey, and Prof. Ioannis Kanellos, from TELECOM Bretagne, France. The presented results were published in:

[8] A.L. Gînscă, A. Popescu, *B. Ionescu*, A. Armagan, I. Kanellos, "*Toward Estimating User Tagging Credibility for Social Image Retrieval*", ACM International Conference on Multimedia - ACM MM 2014, 3-7 November, Orlando, Florida, USA, 2014.

Contribution to state-of-the-art

The proposed approach diversifies the retrieval results using a k-Means algorithm with user-based cluster ranking [8]. The novelty comes from relevance improvement obtained by integrating user tagging credibility. In an initial step, credibility scores are computed by probing user tag-image pairs against a large array of visual concept models learned from ImageNet [143] and by aggregating classification scores at user level. At retrieval time, images are re-ranked based on the credibility scores of the users who uploaded them. This work's contribution to the state-of-the-art is mainly in investigating user tagging credibility in the context of social multimedia mining, a problem which was marginally studied. It provides valuable findings to respond to questions such as: is it possible to automatically estimate user tagging credibility for Web 2.0 multimedia data? how should credibility be integrated in existing multimedia retrieval systems? or, what is the additional complexity of credibility estimation?

Approach

User tagging credibility is estimated by exploiting ImageNet visual models [8]. For each user - user information is provided with the retrieved images, in particular experimentation was carried out with the Div400 dataset that uses Flickr data, see Section 6.6.2 and [1] - we download at most 300 images whose textual annotations match at least one ImageNet concept. Flickr annotations are selected either from tags or from the image title and are all referred as tags hereafter. We perform multi-word detection in order to match multi-words from ImageNet. Tags are tested against corresponding ImageNet concepts to obtain individual relevance scores. User tagging credibility estimation (denoted $cred(U)$) is obtained by averaging scores from individual tag-image pairs.

Visual models are built on top of ImageNet concepts, which are often ambiguous, and tested for Flickr annotations. For instance, if an unknown image annotated with "dog" is tested, which of the different senses of "dog" should be used? An inspection of Flickr results shows that most images annotated with "dog" depict "animals" but there are some of them which depict "dog" as "food" and "dog" as "support". Our credibility estimator should be able to automatically select the right sense of "dog" for the content of the tested image.

A simple way to process ambiguity is to compare the tag-image pair to all models available for the tag and retain only the maximum classification score. Preliminary tests showed that this procedure has good behavior and it is thus used in the experiments. Beyond ambiguity, another problem is the coverage of ImageNet, with some

important senses of words not being included. For instance, "berlin" is represented as "car" but not as "city". These problems represent limitations of our method and their tackling would probably improve credibility estimations.

We propose a retrieval method which diversifies images using k-Means and improves relevance with credibility estimations. Let $L_F = \{(I_1, U_1), (I_2, U_2), (I_3, U_1), ..., (I_N, U_M)\}$ be the ranked list of Flickr images (initial retrieval results) which should be re-ranked to improve the diversity and relevance. Here (I_i, U_j) denote image-user pairs. Our retrieval method can be broken down into three steps: initial filtering, cluster ranking and image ranking:

- **image filtering**. In this step, we remove from L_F all pairs (I_i, U_j) for which I_i qualifies for face or blur removal (most likely to be non-relevant);
- **cluster ranking**. After image filtering, we perform k-Means clustering to diversify the topic representation. Let $C_F = \{C_1, C_2, ..., C_k\}$ be the clustered version of L_F. Inspired by [120], we rank clusters based on *#Users*, the number of distinct users which contribute to each cluster. Ranking based on *#Users* gives priority to clusters which show social consensus. When ties appear with *#Users*, they are broken by using the top Flickr rank among the images of the user with the highest credibility score $cred(U)$ from each cluster. As a result, we obtain $C_F^R = \{C_3, C_k, C_2, ..., C_1\}$, a list of clusters ranked using social cues. For comparison, we also rank clusters based on their raw image count (*#Images*);
- **image sorting**. We exploit credibility estimation to sort images within clusters. Let $C_c = \{(I_1, U_1), (I_3, U_5), (I_8, U_1)\}$ be a cluster with its images ranked by Flickr. Assuming that $cred(U_5) > cred(U_1)$, the sorted representation of the cluster will be $C_c^R = \{(I_3, U_5), (I_1, U_1), (I_8, U_1)\}$. In C_c^R, priority is given to images uploaded by users with higher credibility scores. The final image ranking, L_F^R, is obtained by iterating over C_F^R, the ranked list of clusters, and by selecting each time the first unseen image from C_c^R.

Validation results

Validation was carried out on the Div400 dataset [1] (see also Section 6.6.2). It consists of a development dataset (50 tourist locations with 5,118 photos) and a testing dataset (346 locations with 38,300 photos). Each location is represented with up to 150 photos and associated metadata retrieved with Flickr's default "relevance" algorithm.

Relevance and diversity annotations are available for each photo. Photos are considered relevant if they depict a common photo representation of the location.

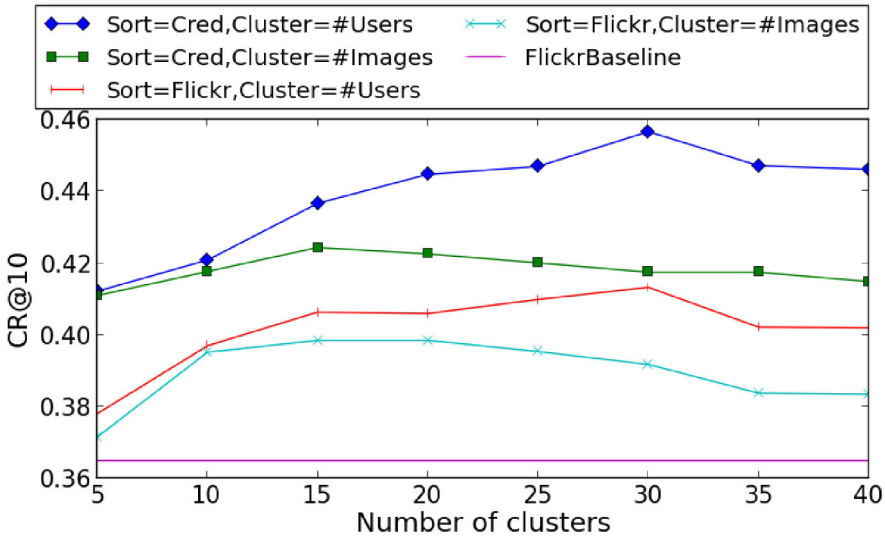


Figure 6.24: CR@10 performances with different clustering methods and different numbers of clusters on the testset of Div400 [1]. *Sort* denotes the type of image sorting used within clusters. *Cred* is a sorting based on user credibility and *Flickr* is the original Flickr ordering. "Cluster" denotes the cluster ranking method. *#Users* and *#Images* represent the user and image counts of a cluster.

A set of photos is considered to be diverse if it depicts complementary visual characteristics of the target location. Clusters are manually built from relevant images of each location. The main objective of the evaluation from Div400 [1] is diversity, which is captured with cluster recall at N (CR@N, see equation 6.25). However, since a good retrieval method should find a good compromise between relevance and diversity, we also report precision (P@N, see equation 6.26) and their harmonic mean, F1@N.

Clustering Analysis. In Figure 6.24, we illustrate the impact of the number of clusters on clustering performances. Within each cluster, *Cred*, the credibility-based image sorting outperforms the use of the initial Flickr sorting in all settings. Intuitively, the best overall results are obtained when *#Users* and *Cred* are combined for inter- and intra-cluster ranking. With 30 clusters, *Flickr* + *#Users* brings a 2 CR@10 points improvement of results compared to *Flickr* + *Images*. This result confirms the conclusions of [120], namely that the use of social cues for cluster ranking is beneficial. More importantly, the introduction of credibility estimation (*Cred* + *#Users*) further improves CR@10 by 4 points. We present results on the testset because they are obtained by averaging a larger number of topics. However, similar results are obtained on the devset and *Cred* + *#Users* with 30 clusters is used for further experiments.

Table 6.19: Comparison of retrieval results obtained with different methods on DIV400 and CR@N, P@N and F1@N metrics. SOTON-WAIS [128] and SocSens [133] are the two most efficient retrieval methods proposed at the 2013 MediaEval Retrieving Diverse Social Images Task [123]. L_F^R [8] corresponds to a setting with *Cred+#Users* and 30 clusters.

Method	*metrics*	@10	@20	@30
SOTON-WAIS [128]	P	**81.58%**	**77.88%**	74.14%
	CR	43.98%	61.97%	72.16%
	F1	54.55%	**66.07%**	70.19%
SocSens [133]	P	73.3%	74.87%	**76.03%**
	CR	42.91%	63.14%	72.28%
	F1	52.09%	65.95%	**70.87%**
proposed L_F^R	P	78.22%	71.54%	69.27%
	CR	**45.67%**	**65.82%**	**78.01%**
	F1	**55.26%**	65.9%	70.73%

Global performance. In Table 6.19, we present the results obtained with the best credibility based retrieval method proposed. It combines clustering and user credibility estimates and produces a re-ranked list of images L_F^R. For comparison, we also present results obtained by the two most efficient existing methods tested on the experimentation dataset during the 2013 MediaEval Retrieving Diverse Social Images Task [123]. To understand the impact of face and blur removal, we briefly present results obtained when we skip one of these steps. When no pre-filtering is used, CR@10 is 44.37%. The use of blur removal or of face removal augments the score to 44.76% and to 45.36%, respectively. While image filtering is beneficial, the main contribution comes from the use of credibility and of user centered clustering.

A comparison of our method to SOTON-WAIS team [128] and SocSens team [133] shows that cluster recall is improved at all cut-off points. For CR@10, the official metric of the 2013 MediaEval Retrieving Diverse Social Images Task, the improvement is close to 2 and 3 percents. Confirming other results obtained on Div400, which show that clustering hurts precision, the P@10 obtained with L_F^R is lower than those obtained in [128]. However, the F1@10 score of our method is slightly better. This comparison shows that our approach is competitive.

Conclusions and future work

We introduced the exploration of user tagging credibility estimation in image retrieval. Evaluation results show that credibility is a good complement to direct text and/or visual content analysis. A credibility model based on text-image pairs

assessment was proposed. The main limitations of our method come from: the mismatch between the background visual resource and the datasets used for retrieval, the limited amount of modeled tags available for some users, the imperfection of visual models and the exclusive use of tag-image content relations. Credibility scores were obtained by comparing user tags to ImageNet concepts with no adaptation to the evaluation dataset, which is made of tourist locations. The fact that credibility estimations are effective even in this difficult setting, accounts for their usefulness.

In the future, we will extend the background visual resource in order to narrow the gap between it and retrieval datasets. New concepts can be learned from noisy Web datasets and their availability would contribute to the reduction of the number of users for which reliable credibility scores cannot be obtained.

6.7 Relevance feedback

Content-based retrieval systems are inherently limited by the gap between the knowledge automatically extracted from the recorded data and its actual semantic meaning. Apart from the methods which attempt to improve that gap exclusively via machine learning techniques, another visited solution is to take advantage of the human expertise in the retrieval process, process known as *relevance feedback*.

Globally, a typical retrieval relevance feedback scenario can be formulated as following: for a certain retrieval query, user feedback is recorded in a very restrained result window by marking results as *relevant* or *non-relevant*. Then, the system computes a better representation of the information needed based on this ground truth and retrieval is further refined (e.g., by re-ranking the initial results). Relevance feedback can go through one or more iterations of this sort until a certain precision is achieved.

6.7.1 Hierarchical clustering relevance feedback

I have contributed to the development of a relevance feedback technique that employs an hierarchical clustering scheme for learning relevant / non-relevant results[25].

[25]this work was developed in cooperation Dr. Ionuţ Mironică and Prof. Constantin Vertan, from LAPI, University Politehnica of Bucharest, Romania. The presented results were published in:

[18] I. Mironică, *B. Ionescu*, C. Vertan, "*Hierarchical Clustering Relevance Feedback for Content-Based Image Retrieval*", IEEE/ACM 10th International Workshop on Content-Based Multimedia Indexing, 27-29 June, Annecy, France, 2012.

[20] I. Mironică, *B. Ionescu*, C. Vertan, "*The Influence of the Similarity Measure to Relevance Feedback*", 20th European Signal Processing Conference, August 27-31, Bucharest, Romania, 2012.

Contribution to state-of-the-art

The proposed relevance feedback is based on an agglomerative hierarchical clustering (HCRF) scheme [18, 19, 20]. A typical agglomerative HC strategy starts by assigning one cluster to each object in the feature space. Then, similar clusters are progressively merged based on the evaluation of a specified distance metric. By repeating this process, HC produces a dendrogram of the objects, which may be useful for displaying data and discovering data relationships. This clustering mechanism can be very valuable in solving the relevance feedback problem by providing a mechanism to refine the relevant and non-relevant clusters in the query results. A hierarchical representation of the similarity between objects in the two relevance classes allows us to select an optimal level from the dendrogram which provides a better separation of the two than the initial retrieval.

The main advantages of the proposed approach are first with the computational efficiency of the hierarchical clustering compared to other clustering techniques used in the literature for similar tasks, such as Support Vector Machines. Furthermore, unlike most relevance feedback algorithms, e.g., FRE [144], Rocchio [145], the proposed approach does not modify the initial query or the similarity measure. The remaining retrieved items are simply clustered according to the class label. Hierarchical clustering has been previously used in relevance feedback but implemented differently. For instance, [146] proposed the QCluster algorithm for image retrieval. It generates a multi-point query to create a hierarchy of clusters followed by use of a Bayesian classification function. In our approach, we simply exploit the dendrogram representation of the two relevance classes. Experimental validation on several datasets and retrieval scenarios proved the superiority of this approach compared to the existing literature.

Approach

The proposed hierarchical clustering relevance feedback is based on the general assumption that content descriptors provide sufficient representative power that, within the first window of retrieved results, there are at least some items relevant to the query that can be used as positive feedback. This can be ensured by adjusting the size of the initial feedback window. Also, in most cases, there is at least one non-relevant item that can be used as negative feedback. The algorithm comprises three steps:

- **retrieval**. We provide an initial retrieval using a nearest-neighbor strategy. We return a ranked list of the N_{RV} items most similar to the query using the Euclidean distance between features. This constitutes the initial relevance

Algorithm 1 Hierarchical Clustering Relevance Feedback [18].

$N_{clusters} \leftarrow N_{RV}$; $clusters \leftarrow \{C_1, C_2, ..., C_{N_{clusters}}\}$;
for $i = 1 \rightarrow N_{clusters}$ **do**
  for $j = i \rightarrow N_{clusters}$ **do**
    compute $sim[i][j]$;
    $sim[j][i] \leftarrow sim[i][j]$;
  end for
end for
while $(N_{clusters} \geq \tau)$ **do**
  $\{min_i, min_j\} =$
  $argmin_{i,j}|_{C_i,C_j \in \{same\ relevance\ class\}}(sim[i][j])$;
  $N_{clusters} \leftarrow N_{clusters} - 1$;
  $C_{min} = C_{min_i} \cup C_{min_j}$;
  for $i = 1 \rightarrow N_{clusters}$ **do**
    compute $sim[i][min]$;
  end for
end while
$TP \leftarrow 0$; $current\_item \leftarrow N_{RV} + 1$;
while $((TP \leq \tau_1) \parallel (current\_item < \tau_2))$ **do**
  for $i = 1 \rightarrow N_{clusters}$ **do**
    compute $sim[i][current\_item]$;
  end for
  if ($current\_item$ is classified as relevant) **then**
    $TP \leftarrow TP + 1$;
  end if
  $current\_item \leftarrow current\_item + 1$;
end while

feedback window. Then, the user provides feedback by marking relevant results, which triggers the actual mechanism;

- **training**. The first step of the proposed algorithm consists of initializing the clusters. At this point, each cluster contains a single item from the initial relevance feedback window. Basically, we attempt to create two dendrograms, one for relevant and one for non-relevant items. To assess similarity, we compute the Euclidean distance between cluster centroids. Once we have determined the initial cluster similarity matrix, we attempt to merge progressively clusters from the same relevance class (according to user feedback) using a minimum distance criteria. The process is repeated until the number of remaining clusters becomes relevant to the categories in the retrieved window (regulated by a threshold τ);

- **updating**. After finishing the training phase, we begin to classify the next

items (outside the relevance feedback window) as relevant or non-relevant with respect to the previous clusters. A given item is classified as relevant or non-relevant if it is within the minimum centroid distance to a cluster in the relevant or non-relevant dendrogram.

Algorithm 1 summarizes the steps involved. The following notations were used: $N_{clusters}$ is the number of clusters, $sim[i][j]$ denotes the distance between clusters C_i and C_j (i.e., centroid distance), τ represents the minimum number of clusters which triggers the end of the training phase (set to a quarter of the number of items in a browsing window), τ_1 is the maximum number of searched items from the database (set to a quarter of the total number of items in the database), τ_2 is the maximum number of items that can be classified as positive (set to the size of the browsing window), TP is the number of items classified as relevant, and $current\_item$ is the index of the currently analyzed item.

Validation for image retrieval

The first validation was conducted in the context of the image retrieval systems. We experimented on the Microsoft Object Class Recognition[26] dataset, which sums up to 4,300 images distributed into 23 categories (e.g., "animals", "people", "airplanes", "cars", etc) as well as on the Caltech-101[27] dataset, which contains a total of 9,146 images, split between 101 distinct objects (including faces, watches, ants, pianos, etc) and a background category.

For image content description we tested several state-of-the-art approaches, namely: MPEG-7 image descriptors [38]: Color Histogram Descriptor, Color Layout Descriptor, Edge Histogram Descriptor and Color Structure Descriptors; classic color descriptors: Autocorrelogram, Color Coherence Vectors and Color Moments; and feature detectors: SURF, SIFT, Good Features to Track (GOOD), STAR, Accelerated Segment Test (FAST), Maximally Stable Extremal Regions (MSER) and Harris Detector available with the OpenCV library (Open Source Computer Vision[28]). Features were represented with a Bag-of-Visual-Words model.

To assess performance, we compute the overall Mean Average Precision (MAP) as defined in equation 6.13.

The evaluation consisted of systematically considering each image from the database as query image and retrieving the remainder of the database accordingly. Experiments were conducted for various retrieval browsing windows, N_{RV}, ranging from

[26]http://research.microsoft.com/en-us/projects/objectclassrecognition.
[27]http://www.vision.caltech.edu/Image_Datasets/Caltech101.
[28]http://opencv.willowgarage.com/wiki.

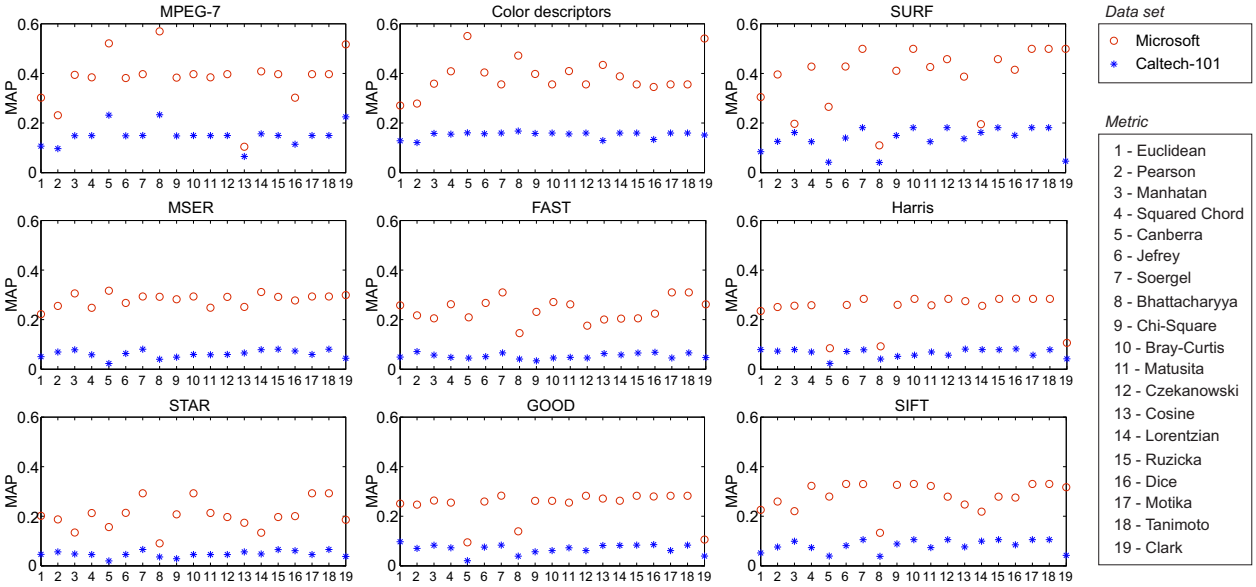


Figure 6.25: Mean Average Precision for retrieval using various descriptor set - metric combinations [18].

20 to 50 images. For brevity reasons, in the following we present only the most representative results which were obtain for $N_{RV} = 30$.

The user feedback was simulated automatically from the known class membership of each image. Compared to real user feedback, this has the advantage of providing a fast and extensive simulation framework, which otherwise could not be achieved due to physical constraints (e.g., availability of a significant number of users) and inherent human errors (e.g., indecision, misperception).

The influence of the similarity measure. In the first experiment we analyze the influence of the distance measure on the performance of a classic retrieval system. In this respect, we use the classic nearest neighbor retrieval step of the relevance feedback algorithm. Figure 6.25 presents the MAP obtained for the two data sets and the aforementioned features. Although the descriptors provide in average more or less comparable performance on same data set, results show that the distance measure plays a critical role.

In the case of the Microsoft dataset which has lowest diversity of classes, the best results are obtained with Bhattacharyya using MPEG-7 descriptors (MAP of 57%) followed by Canberra and Clark using classic color descriptors (MAP of 55% and 54%, respectively) which is an improvement of around 18% above the average descriptor value. Results are significantly decreasing on the Caltech-101 dataset which contains five times more categories. The highest accuracy is achieved again

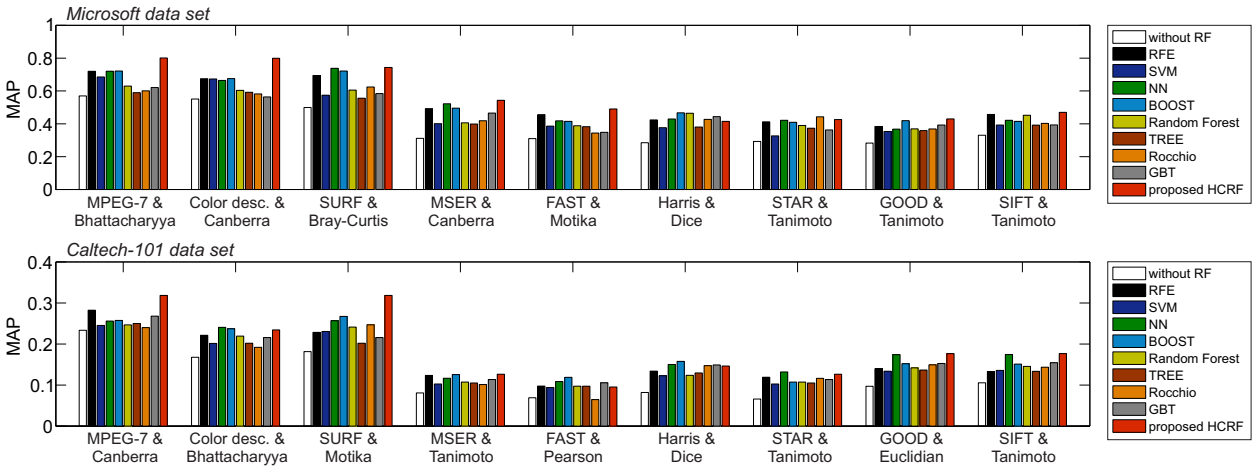


Figure 6.26: Mean Average Precision for retrieval with relevance feedback using various descriptors and metrics [18].

for the classic MPEG-7 descriptors using Bhattacharyya and Canberra distances (MAP of 23.4% and 23.2%, respectively). In this case, the improvement is of at least 5% above the average descriptor value. However, it should be considered the fact that Bhattacharyya is one of the most computational expensive solutions.

It can be noted that some distance measures can be less adapted to the structure of the descriptors, an example are Bhattacharyya and Canberra which performed significantly worst with Bag-of-Visual-Words representation of feature descriptors (see SURF, SIFT, Harris and GOOD in Figure 6.25). Another interesting result is that the classic Euclidean distance, despite its popularity, proves to have poor discriminant power in most of the cases.

Relevance feedback evaluation. Form the previous experiment it can be seen that the retrieval performance is relatively low with either of the methods. In this experiment we test the advantage of using relevance feedback. We compare the proposed approach against other validated methods from the literature: Rocchio algorithm [145], Relevance Feature Estimation (RFE) [144], Support Vector Machines (SVM) [147], Decision Trees (TREE) [148], AdaBoost (BOOST) [149], Random Forest [150], Gradient Boosted Trees (GBT) [151] and Nearest Neighbor (NN) [152]. As for the previous experiment, we assess similarity using various distance measures. Each experiment is conducted using only one feedback iteration. Some of the results are presented in Figure 6.26. For brevity reasons, we depict only the results obtained with the distance measure providing the highest performance.

From the relevance feedback point of view, globally, all the strategies provide significant improvement in retrieval performance compared to the retrieval without

Table 6.20: Improvement achieved by the proposed hierarchical clustering relevance feedback (HCRF) [18].

Microsoft dataset			
descriptor	*1st MAP*	*2nd MAP*	*3rd MAP*
MPEG-7	HCRF - 80%	BOOST - 72%	NN - 72%
Color desc.	HCRF - 80%	RFE - 68%	BOOST - 68%
Caltech-101 dataset			
descriptor	*1st MAP*	*2nd MAP*	*3rd MAP*
MPEG-7	HCRF - 32%	RFE - 28%	GBT - 27%
SURF	HCRF - 32%	BOOST - 27%	NN - 26%

relevance feedback. Better performance is naturally obtained when targeting a more reduced number of image categories. For instance, on the Microsoft data set (23 classes) relevance feedback MAP is up to 80% (proposed method HCRF), compared to only 57% without relevance feedback (improvement of 23%). On Caltech-101 (102 classes) the highest MAP is 32% (proposed HCRF) compared to 23% without relevance feedback (improvement of 9%).

The proposed approach tends to provide better retrieval performance in most of the cases. Table 6.20 summarizes some of these results. For the Microsoft data set, the highest increase in performance is achieved for MPEG-7 descriptors, namely 8% compared to BOOST; while for Caltech-101, the highest increase is of 5% for SURF compared also to BOOST. Less accurate results are obtained for descriptors such as FAST, STAR or MSER due to their limited discriminant power for this particular task. From the distance point of view, results show that there is no general preference for a certain distance metric. As expected, the choice of distance is dependent on the type of content descriptors. Nevertheless, Canberra and Bhattacharyya distances prove to be more reliable for use with classic numeric content descriptors, such as MPEG-7 and color descriptors, while Tanimoto provided better performance for Bag-of-Visual-Words approaches.

Validation for video retrieval

Video retrieval validation was conducted on the 2011 MediaEval Video Genre Tagging Task dataset [48], which contains 2,375 web sequences labeled according to 26 genre categories (e.g., "'art", "autos", "comedy", etc). Each sequence was represented with the audio-visual descriptors proposed in Section 6.3.1. As for the previous experiment, user feedback was simulated automatically from the known class membership of each video (i.e., the genre labels). In addition to MAP, we

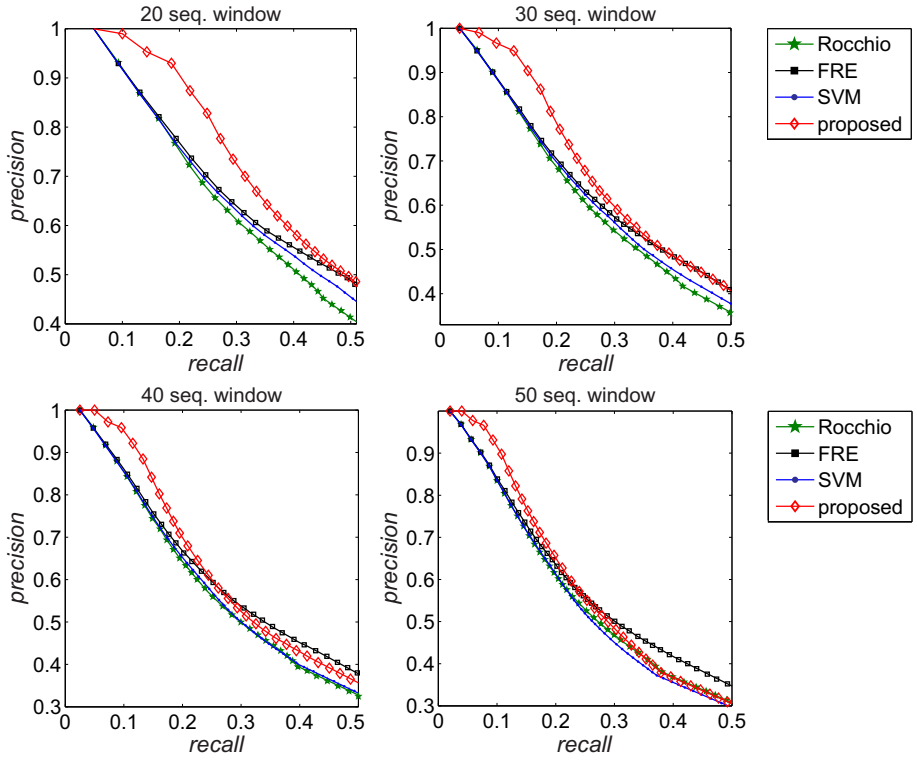


Figure 6.27: Precision - recall curves obtained with the proposed hierarchical clustering relevance feedback (HCRF) [4] for different size browsing windows.

Table 6.21: MAP obtained with the proposed hierarchical clustering relevance feedback (HCRF) [4].

RF method	*20 seq. window*	*30 seq. window*	*40 seq. window*	*50 seq. window*
Rocchio [145]	46.8%	43.84%	42.05%	40.73%
FRE [144]	48.45%	45.27%	43.67%	42.12%
SVM [147]	47.73%	44.44%	42.17%	40.26%
proposed	**51.27%**	**46.79%**	**43.96%**	**41.84%**

report also precision and recall (see equation 6.5).

Figure 6.27 compares the precision - recall curves obtained with the proposed approach, with those of several other approaches, namely Rocchio [145], Feature Relevance Estimation (FRE) [144] and Support Vector Machines (SVM) [147]. The proposed approach (HCRF) provides an improvement in retrieval, particularly for small browsing windows (e.g., 20, 30 video sequences, see the red line in Figure 6.27). With increasing window size, all methods tend to converge at some point to similar results. Table 6.21 summarizes the overall retrieval MAP. For the proposed approach, the MAP ranges from 41.8% to 51.3%, which is an improvement over the

other methods of at least a few percents.

Conclusions and future work

We proposed a relevance feedback approach that uses the hierarchical clustering of the query results. Experimental testing performed on several standard databases using state-of-the-art descriptors and distance measures proved its efficiency.

Although descriptors provided more or less comparable retrieval results, the choice of the distance measure proves to be highly critical for the performance. Distances such as Canberra and Bhattacharyya proved to be more reliable for use with classic numeric descriptors, such as MPEG-7 and color descriptors, while metrics such as Tanimoto provided better performance for Bag-of-Visual-Words approaches. The proposed approach was validated in both image and video retrieval scenarios outperforming other approaches from the literature which makes it valuable for solving the retrieval paradigm. Future work will mainly involve considering the constraints of large-scale indexing.

6.7.2 Fisher Kernel-based relevance feedback

Another approach to relevance feedback was from the perspective of Fisher Kernel representation[29] (see also Section 6.3.2).

Contribution to state-of-the-art

We proposed a new relevance feedback approach adapted in particular to video retrieval that uses a combination of Fisher Kernels (FK) with Support Vector Machine Classifiers (SVM) [13]. FKs are a powerful framework which combines the advantages of a generative algorithms with the strengths of discriminative approaches. The main idea of FK is to describe a signal with a gradient vector derived from a generative probability model, e.g., Gaussian Mixture Model - GMM, and then train this representation with a discriminative classifier, i.e., in most of the cases SVMs. FKs have been successfully applied to many fields, from image categorization [153], to audio indexing [155] and handwritten word-spotting [154], but to our knowledge, it never been used in relevance feedback or in video classification (more information is presented in Section 6.3.2).

In order to describe a document, most of the relevance feedback strategies use a single feature vector, but, video documents can be considered as a sequence of

[29]this work was developed in collaboration with Dr. Ionuţ Mironică, from LAPI, University Politehnica of Bucharest, Romania, Dr. Jasper Uijlings, and Prof. Nicu Sebe, from MHUG, University of Trento, Italy. The presented results were published in:

Algorithm 2 The Fisher Kernels Relevance Feedback Algorithm [13].

initial parameters:
Labeled sample set: X_i *and labels* Y_i;
Unlabeled sample set X_r;
SVM classifier parameters (C, γ*)*;
n*: the window size*;
start:
do 10% PCA reduction for all multimodal features;
altering features step:
compute GMM centroids for X_i;
for $x \subset Xi$ **do**
  $FK(x) = FK(x, GMM)$;
  normalize $FK(x)$;
end for
training - re-ranking step:
train SVM(C, γ*) using* FK *features*;
for $x \subset Xr$ **do**
  $FK(x) = FK(x, GMM)$;
  normalize $FK(x)$;
  compute $h(x)$ = *SvmConfidenceLevel(*$FK(x)$*)*;
end for
sort $h(x)$ *values*;
show new ranked list according to $h(x)$ *values*;

scenes, and the features from each scene can be used to model and retrieve the video content. Because we use the FK framework, we can integrate the spatial relationship of the video scenes in our relevance feedback approach. In addition, the proposed approach allows a fast implementation similar to a classical SVM relevance feedback strategy, but with a higher increase of performance.

Approach

We introduced the following relevance feedback approach (see Algorithm 2). Using a single video as query, we rank all videos using a nearest-neighbor strategy. Then the user selects from the top n videos which ones are relevant and which ones are not, where n is typically small (e.g., 20 in our experiments). We learn a generative Gaussian Mixture Model (GMM) from first n retrieved documents. Then we re-represent the top k videos using a Fisher Kernel (FK) representation with respect to this GMM (as defined in Section 6.3.2), where k is typically large (e.g., $2,000$ in

[13] I. Mironică, *B. Ionescu*, J. Uijlings, N. Sebe, "*Fisher Kernel based Relevance Feedback for Multimodal Video Retrieval*", ACM International Conference on Multimedia Retrieval - ICMR 2013, Dallas, Texas, USA, April 16 - 19, 2013.

our experiments). We only consider the top k as it is unlikely that relevant videos are ranked lower in the initial ranking. Afterwards, we train an SVM on the FK vectors of the top n user labeled results. We apply this SVM on the top k videos to obtain a final ranking. The algorithm involves the following steps:

- **altering features after user's feedback**. After the initial query, using nearest-neighbor search, we train a Gaussian Mixture Model on the features of the top n videos, regardless of their true relevance. In a practical application this would allow the training of the Gaussian Mixture Model in the background during the time that the user is giving feedback. For optimization reasons we initialize the centroids with a k-means output.

  To make the Fisher Kernel computationally feasible, we first apply Principal Component Analysis (PCA) on the original feature vectors of the documents. We compute PCA individually on each feature type and reduce the dimensions by 10%. After having obtained the mixture model, we convert the original features of the top k videos into the FK representation according to equations 6.14 and 6.15. For both the GMM clustering and the Fisher projection we use the software from [153].

  Finally, we perform normalization on the FK vectors as this will significantly increase performance.

- **training - re-ranking step**. We use the FK representations of the top n videos along with the labels obtained using feedback from the user to train a two-class SVM classifier. SVMs are appropriate for relevance feedback as they are relatively robust to the situation in which only few training examples are available.

Validation experiments

For testing, we use the 2012 MediaEval Video Genre Tagging dataset [62] consisting of around 15,000 sequences (up to 2,000 hours of video footage), labeled according to 26 video genre categories (e.g., "art", "autos and vehicles", "business", etc - see the complete list in Section 6.3.3 with the Experimental results). Video description is carried out with the following type of descriptors:

- *audio features*:
  - block-based audio features (11,242 values) [40] - capture the temporal properties of the audio signal. We use the audio descriptors introduced in Section 6.3.1;

  - standard audio features (196 values) [86] - we used a set of general-purpose audio descriptors: Linear Predictive Coefficients (LPCs), Line Spectral Pairs (LSPs), MFCCs, Zero-Crossing Rate (ZCR), and spectral centroid, flux, rolloff and kurtosis, augmented with the variance of each feature over a certain window (we used the common setup for capturing enough local context that is equal to 1.28 s). For a clip, we take the mean and standard deviation over all frames.

- *visual descriptors*:

  - MPEG-7 related descriptors (1,009 values) [38] - we adopted standard color and texture-based descriptors such as: Local Binary Pattern (LBP), autocorrelogram, Color Coherence Vector (CCV), Color Layout Pattern (CLD), Edge Histogram (EHD), Scalable Color Descriptor (SCD), classic color histogram (hist) and color moments. For each sequence, we aggregate the features by taking the mean, dispersion, skewness, kurtosis, median and root mean square statistics over all frames;
  - global HoG (81 values) [88] - from this category, we compute global Histogram of oriented Gradients (HoG) over all frames;
  - structural descriptors (1,430 values) - the structural description is based on a characterization of geometric attributes for each individual contour, e.g. degree of curvature, angularity, circularity, symmetry and "wiggliness", as proposed in [39] and introduced in Section 6.3.1.

- *text descriptors*: Term Frequency - Inverse Document Frequency (TF-IDF) extracted with automated speech recognition algorithms (ASR), with 3,466 values (ASR is provided with the dataset [62]) - we use the standard TF-IDF approach. First, we filter the input ASR text by removing the terms with a document frequency less than 5%-percentile of the frequency distribution. We reduce further the term space by keeping only those terms that discriminate best between genres according to the 2-test. We generate a global list by retaining for each genre class, the m terms (e.g., m = 150) with the highest 2 values that occur more frequently than in complement classes. This results in a vector representation for each document.

All the visual and audio descriptors are normalized to L_∞ norm, and text descriptors to cosine normalization.

User feedback is automatically simulated from the known class membership of each video document (ground truth is provided with the databases). This approach

Table 6.22: Comparison between feature MAP using different metrics without relevance feedback [13].

Feature	*Manhatan*	*Euclidian*	*Mahalanobis*	*Cosinus*	*Bray Curtis*	*Chi Square*	*Canberra*
HoG	17.02%	**17.18**%	17.07%	17.00%	17.10%	17.07%	16.67%
structural desc.	10.87%	10.55%	11.14%	2.18%	10.92%	11.58%	**14.82**%
MPEG-7 related	12.37%	10.85%	21.14%	08.69%	13.34%	13.34%	**25.97**%
standard audio desc.	7.76%	7.78%	**29.26**%	15.28%	7.78%	8.04%	1.58%
block-based audio desc.	19.33%	19.58%	20.21%	**21.23**%	19.71%	19.99%	20.37%
text TF-IDF ASR	8.32%	7.15%	5.39%	17.64%	**20.40**%	9.83%	9.68%

allows a fast and extensive simulation which is necessary to evaluate different methods and parameter settings.

To assess performance, we use classic precision and recall (see equation 6.5) and Mean Average Precision (MAP) (see equation 6.13). In our evaluation we systematically consider each document from the database as a query document and retrieve the remainder of the database accordingly. Precision, recall and MAP are averaged over all retrieval experiments. Experiments were conducted for various browsing top n, ranging from 10 to 30 documents. For space and brevity reasons, in the following we shall present only the results using the top 20 videos per window. The general observations in this paper hold for all values of n.

Evaluating feature metrics. Some distance measures are better adapted to the structure of the descriptors than others [13]. We tested a broad variety of metrics. Their performance is presented in Table 6.22. We conclude that each feature has its own preferred metric. In the rest of our experiments we use for each feature its best metric as indicated in Table 6.22 with bold.

Relevance feedback using Fisher Kernels. In this experiment we study the influence of Fisher Kernel parameters on the system's performance. In the first experiment we analyze the influence of the number of Gaussian centroids. Figure 6.28 presents the variation of MAP using a different number of Gaussian centroids. It can be observed that the best results are obtained using only a single Gaussian centroid. In this case the size of Fisher Kernel descriptors will be 2 times bigger than the document descriptor.

The second experiment presents the influence of Fisher normalization strategies on system performance. In [153], it was demonstrated that some normalization strategies can improve the performance of Fisher Kernels. We have tested four nor-

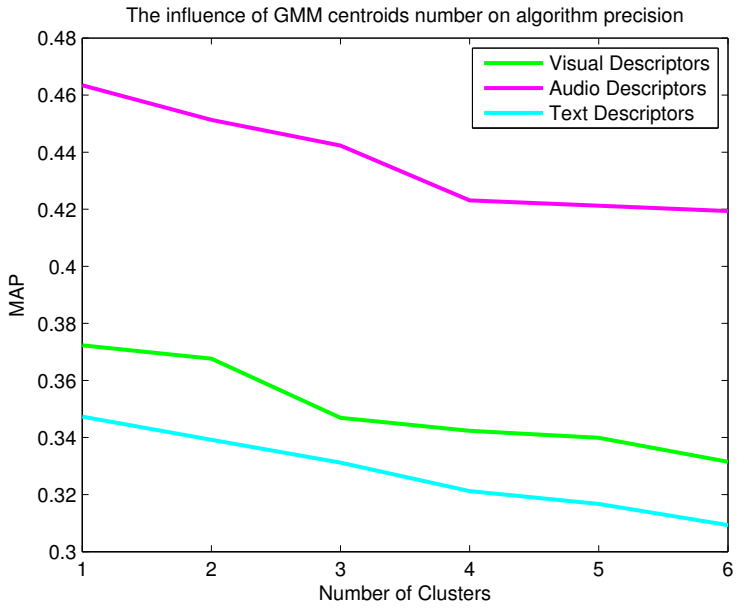


Figure 6.28: The influence of GMM centroids on system performance [13].

Table 6.23: The influence of different normalization strategies on relevance feedback performance (MAP) [13].

Normalization/Features	*Visual*	*Audio*	*Text*
without normalization	37.25%	38.68%	31.13%
L1	36.82%	37.97%	29.83%
L2	39.22%	41.94%	30.51%
Log norm.	38.61%	42.01%	**35.07**%
PN	38.51%	41.37%	34.93%
PN + L2 norm.	39.20%	42.98%	30.12%
PN + L1 norm.	**39.46**%	**43.23**%	31.71%

malization algorithms and some combination of them: L1 normalization, L2 normalization, power normalization ($f(x) = sign(x)\sqrt{\alpha|x|}$) and logarithmic normalization ($f(x) = sign(x)log(1 + \alpha|x|)$). The results are presented in Table 6.23.

It can be observed that using the combination of L1 normalization - alpha normalization we obtained the best results for visual and audio features, while the highest performance for text features is obtained with logarithmic normalization. Another observation is that using the L1 normalization alone, the results are lower than in the case when L1 is used in combination with other normalizations.

In order to compare our algorithm with other relevance feedback approaches, we have selected the settings that provide the greatest improvement in performance:

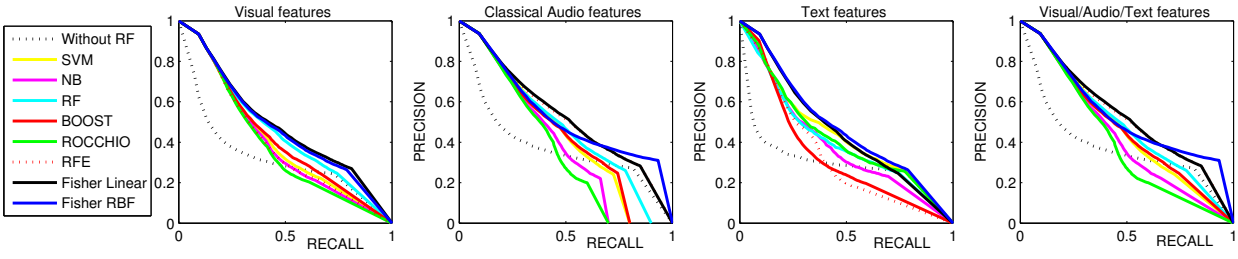


Figure 6.29: Precision-recall curves for different content descriptors and combinations: 1 - combination of all visual descriptors, 2 - standard audio features, 3 - text descriptors and 4 - combination of all features [13].

Table 6.24: Comparison of the proposed FK relevance feedback with state-of-the-art algorithms in terms of MAP (%) [13].

Feature	*without RF*	*Rocchio*	*NB*	*BOOST*	*SVM*	*RF*	*RFE*	*FKRF Linear*	*FKRF RBF*
HoG	17.18	25.57	24.18	26.72	26.49	26.89	27.5	**29.46**	**29.59**
structural	14.82	21.96	23.73	23.63	24.62	24.69	23.91	**26.28**	**23.96**
MPEG-7	25.97	30.88	34.09	32.55	32.90	36.85	31.93	**40.50**	**40.80**
all visual	26.11	32.76	34.15	35.76	35.88	**39.08**	32.43	38.01	**38.23**
standard audio	29.26	32.71	34.88	32.88	38.58	40.46	44.32	**44.80**	**46.34**
block-based audio	21.23	35.39	35.22	39.87	31.46	33.41	31.96	**43.96**	**43.69**
text	20.40	32.55	26.91	26.93	34.70	34.70	25.82	**34.84**	**35.14**
all desc.	30.29	37.91	39.88	38.88	40.93	45.31	44.93	**45.43**	**45.80**

one GMM centroid, L1 normalization with alpha normalization for audio and visual descriptors and logarithmic normalization for text descriptors. We also used 2 SVM kernels: a linear SVM classifier and a nonlinear RBF kernel.

Comparison to state-of-the-art techniques. In the following, we compare our approach against other validated algorithms from the literature, namely: the Rocchio's algorithm [145], Relevance Feature Estimation (RFE) [144], Support Vector Machines (SVM) [147], AdaBoost (BOOST) [149], Random Forests (RF) [150] and Nearest Neighbor (NB) [152].

Figure 6.29 presents the precision-recall curves after one iteration of feedback for different descriptor categories. Globally, all relevance feedback strategies provide significant improvement in retrieval performance compared to the retrieval without relevance feedback (see the dashed black and blue lines in Figure 6.29). Better performance is naturally obtained with audio descriptors, while text and visual descriptors have similar performance. The highest increase in system performance

Table 6.25: Comparison between the proposed Fisher Kernel relevance feedback with RBF kernel on all data and RF RBF (MAP) [13].

Feature	*proposed FK for all data*	*proposed FK with RBF kernel*
visual desc.	34.02%	**38.23**%
standard audio desc.	38.25%	**46.34**%
text desc.	32.37%	**35.14**%

is obtained using standard audio descriptors, increase of MAP from 29.35% (without RF) to 46.34% and with all combined features from 30.29% to 45.80%.

We present in Table 6.24 the MAP values for different features combinations. The proposed approach (FKRF) has the highest values for most of the cases, except for the combination of all visual descriptors, where the Random Trees relevance feedback achieves the highest performance values. The highest increase in performance is obtained using MPEG-7 descriptors, increase of 4 MAP percents (from 40.80% using proposed FKRF and RBF kernel to 36.85% with Random Forests) and block based audio (from 43.96% using proposed FKRF with linear kernel to 39.87% using Boost relevance feedback).

In most of the cases, RFE and Random Forests provide good results, but still less than our approach. We conclude that the proposed approach improves the retrieval performance, outperforming some other existing approaches.

Fisher representation on all data. We could also generate a Fisher Kernel representation by learning a GMM on *all* the data. A valid question is therefore: do we obtain good results because the Fisher Kernel representation in general is more powerful than our initial features, or are our performance improvements caused by altering the features with respect to the top n results? In the former case, we can just alter the features once off-line, which would speed up computation. Yet, if this is the case we would just prove that the Fisher representation is more powerful than our initial features, independent of our relevance feedback setting.

To test this, we train a GMM on all the feature vectors of the whole dataset, and represent all videos as Fisher vectors with respect to this global mixture model. We use these features in the SVM relevance feedback framework and compare this with our proposed FK relevance feedback. Notice that the only differences between these two systems are on what data the GMM is learned and when the features are changed to the Fisher representation. GMM on *all* drastically reduce the system performance. The results are presented in Table 6.25.

We conclude that altering the data based on the top n videos is crucial for obtaining good performance. This validates our claim that the Fisher Kernel is particularly suited for use in a relevance feedback application.

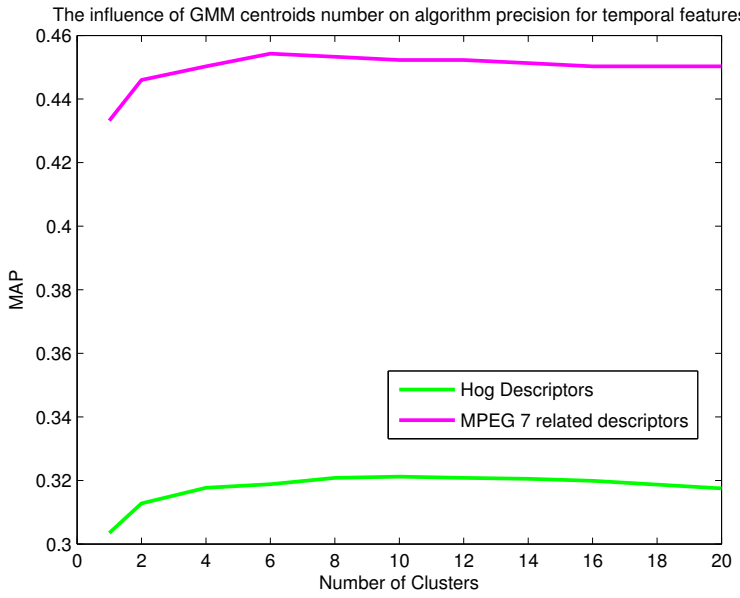


Figure 6.30: The influence of GMM centroids on system performance using video chunks [13].

Table 6.26: Comparison between the proposed global FKRF and temporal FKRF (MAP) [13].

Feature	*FKRF Linear*	*FKRF RBF*	*Temporal FKRF Linear*	*Temporal FKRF RBF*
HoG	29.46%	29.59%	32.12%	**32.87**%
MPEG-7 related	40.50%	40.80%	44.69%	**45.43**%

Temporal Fisher representation for relevance feedback. In the following, we prove the superiority of the FK relevance feedback approach when we use more then one feature per video document. We experimented with two types of visual descriptors: HoG descriptors and MPEG-7 related descriptors. We extract a small collection of salient frames using [156], and compute a visual feature for each image. Now, every frame models the parameters of the existing model, and the feature space becomes more complex. In this case, we estimate that the number of centroids used by the Fisher vectors is higher than one. Figure 6.30 presents the variation of MAP using a different number of Gaussian centroids. It can be observed that the best results are obtained using 5 to 10 number of centroids.

We present in Table 6.26 a comparison between the MAP values of the proposed global FKRF and temporal FKRF. The temporal Fisher representation for relevance

feedback tends to provide better retrieval performance in all cases with more than 4 percents increase of performance (from 29.59% to 32.87% for HoG features and from 40.80% to 45.43% for MPEG-7 related descriptors).

Conclusions and future work

We proposed a new relevance feedback approach that uses Fisher Kernels (FK). Tested on a large scale video database (MediaEval 2012) and using several multimodal descriptor schemes (i.e., visual, video and textual), the proposed approach improves the retrieval performance outperforming some other existing approaches. We also present a novel method to train the relevance feedback algorithm, using more than one feature per video. The experiments with visual descriptors showed that using more features vectors to describe a video document, instead of only one, the performance is drastically improved. We proved that we don't need large vocabularies to train the FK framework, we achieve the best performance with only 5-10 words. This makes the proposed approach implementable for a real time RF approach. Future improvements will mainly consist in adapting the method to address a higher diversity of video categories (use of the Internet). Another improvement is the usage of more elaborated spatio-temporal features.

6.8 Multimedia browsing

Apart from the retrieval mechanism, another important component of a content-based retrieval system is the *browsing system* (see Figure 6.1). Approaching this task requires defining an environment for browsing multimedia information, from the software architecture point of view. I have contributed to the development of two multimedia browsing interfaces[30].

[30]this work was developed in cooperation with Ioan Chera, Vlad Dima and Alexandru Marin, from LAPI, University Politehnica of Bucharest, Romania. The presented results were published in the following articles and thesis:

[21] *B. Ionescu*, K. Seyerlehner, C. Rasche, C. Vertan, P. Lambert, "*Content-based Video Description for Automatic Video Genre Categorization*", 18th International Conference on MultiMedia Modeling - MMM 2012, 4-6 January, Klagenfurt, Austria, 2012.

[22] *B. Ionescu*, P. Lambert, D. Coquin, A. Marin, C. Vertan, "*Analyzing Animated Movie Contents for Automatic Video Indexing*", in Machine Learning Techniques for Adaptive Multimedia Retrieval: Technologies Applications and Perspectives, IGI - Global Printing House, Eds. Chia-Hung Wei et al., pp. 228-260, (ISBN 978-1-61692-859-9), 2011.

[] Vlad Dima: "A 3D System for Navigating through Multimedia Databases", dissertation thesis, University Politehnica of Bucharest, Romania.

[] Ioan Chera: "Multimedia Browsing Interfaces", dissertation thesis, University Politehnica of Bucharest, Romania.

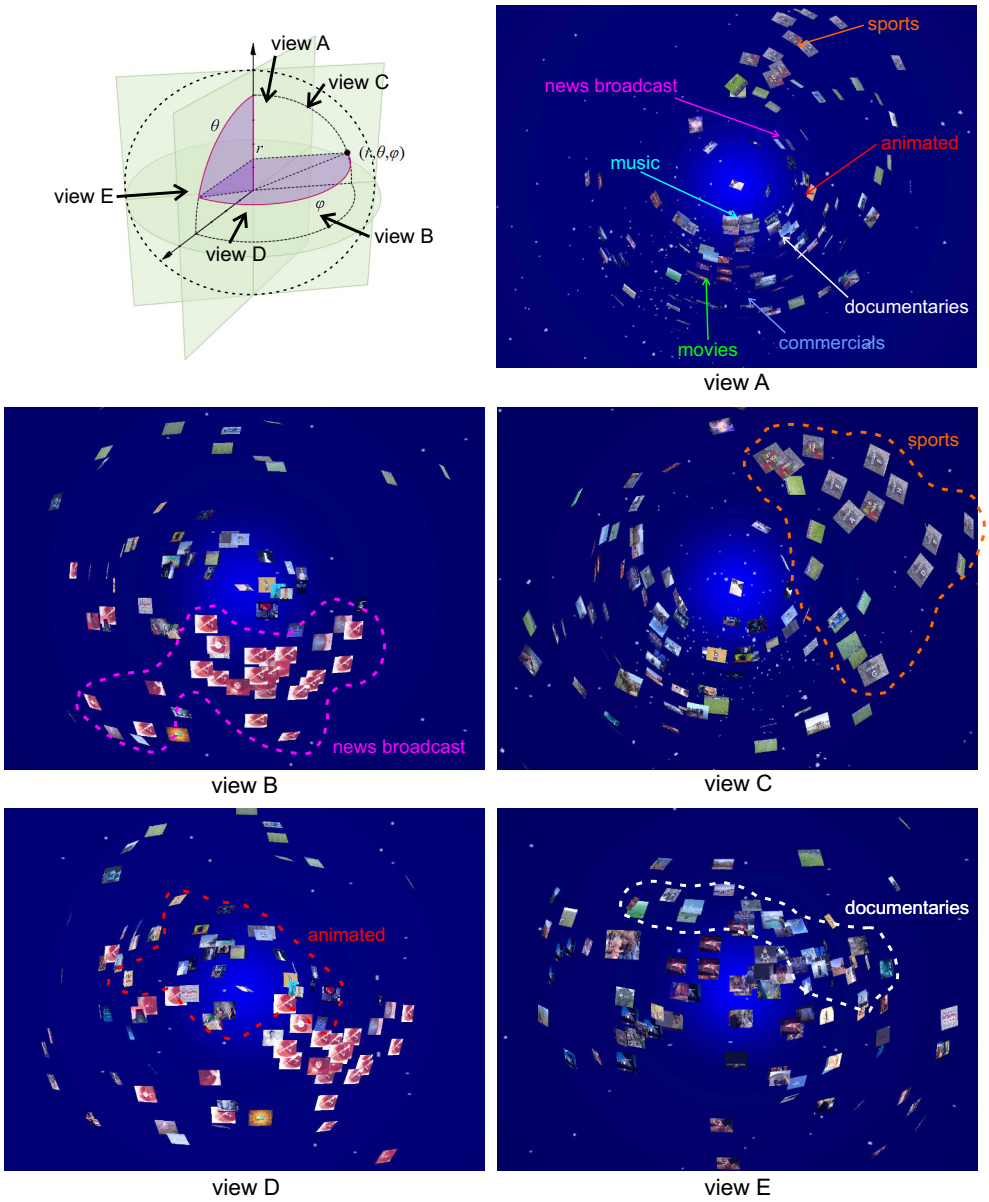


Figure 6.31: MovieGlobe feature-based 3D movie representation in a spherical coordinate system (inclination-θ, azimuth-φ, radius-r). Each movie from the data set is represented by a point with which we associate an image vignette. Views A to E are screenshots taken from different perspectives (the points of view used are shown in the chart). In views A-E, representative genres are annotated.

Experiment 1 - MovieGlobe - a 3D image/video browsing interface. The first experiment consisted in simulating a image/video browsing environment in which items are represented via content descriptors. We have developed a client-server architecture which provides a virtual 3D browsing environment for image/video databases

[22, 21]. Items are displayed in a spherical coordinate system, and each item is represented by one key frame. The user interface resembles that of Google Earth[31] (by which we were inspired): the user flies virtually through a "world of images/videos". User interaction consists of moving freely in the 3D environment, zooming on the items as well as visualizing items' properties (e.g., description, watching a movie, etc)[32].

Figure 6.31 exemplifies our interface by presenting several screen shots achieved when displaying a small video database according to descriptor coordinates, namely the first three principal components of the descriptors proposed in Section 6.3.1. Due to similarity in content and structure, movies re-group according to genre, e.g., the most clearly grouped genres are news (see view B) and sports (see view C).

The main drawback of this platform was the used technology, being programmed in Java 3D. This technology limited in the first place its portability, both on different platforms and in time, as well as its processing capabilities - all the data has to be uploaded in the RAM memory which limits significantly the number of displayed items (e.g., to hundreds of videos, or a few thousand images).

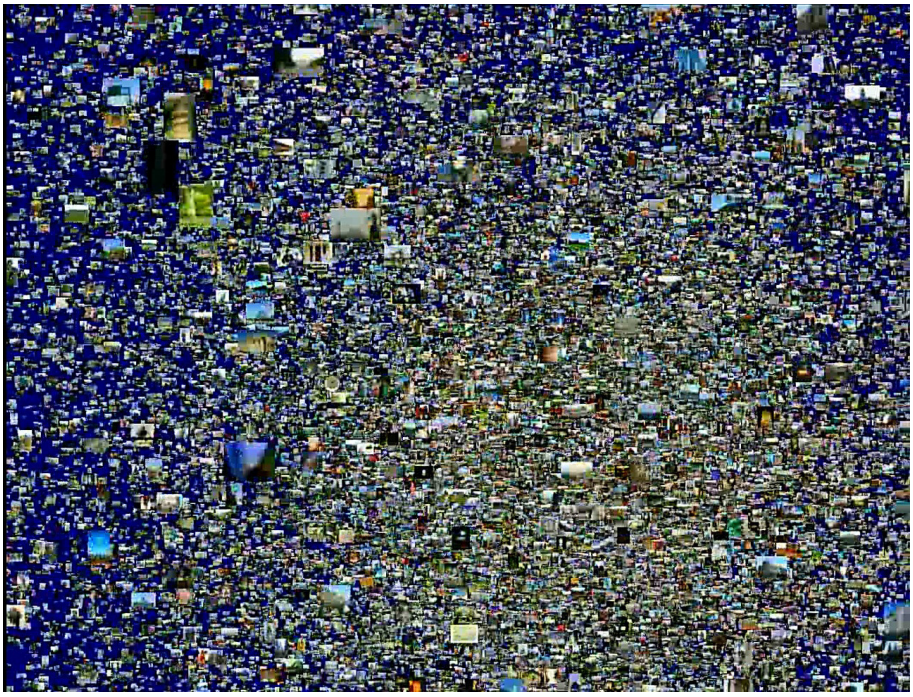
Experiment 2 - MediaView - a 3D multimedia browsing interface. Motivated by the success of the first interface, we proceeded with developing of a more complex browsing system. The new interface inherits the idea of displaying multimedia items on a sphere and use Google Earth like interaction. However, this application focused on providing a more general framework addressing not only image/videos but all type of multimedia including audio.

The system was developed using Visual C++ which makes it highly portable and easily upgradable in the future (it can be compiled on any machine). Then, the key feature is in the capability of displaying in real-time large datasets of all types of media information, from image, video to audio data. This was possible by an efficient allocation and pre-caching of resources. Tests conducted with more than 150,000 items (image-video-audio) show real-time browsing capabilities even when run on a regular laptop computer.

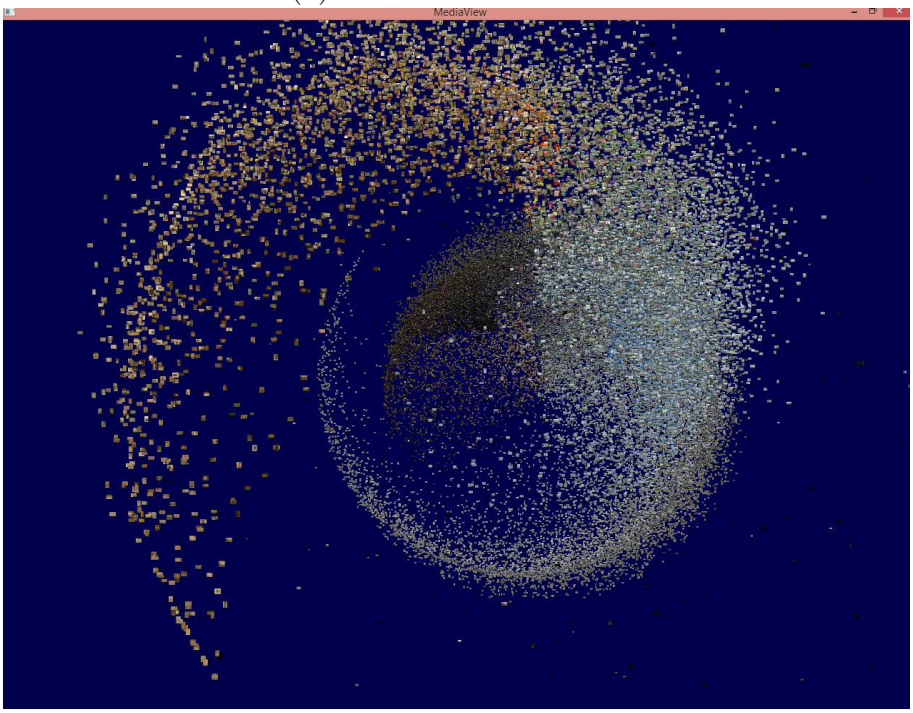
Apart from displaying multimedia items according to user specified coordinates, the system provides also a basic searching system which uses a k-nearest neighbor approach. Resources for displayed items, e.g., content descriptors, have been externalized in order to increase its generality. Basically, item coordinates can be generated with any existing technology, such as Matlab environment or OpenCV. The interface with the system is as general as possible and uses XML (Extensible Markup Language) descriptions.

[31] `https://www.google.com/earth/`.

[32] a demo is available at `http://imag.pub.ro/~bionescu/index_files/MovieGlobe.avi`.



(a) Random coordinates



(b) Color histogram descriptors (PCA)

Figure 6.32: Screenshots of the proposed MediaView (representation in a spherical coordinate system - each multimedia item is represented with an image vignette).

Figure 6.32 presents two screenshots of the proposed interface when displaying a 120,000 image database (image source Flickr). In Figure 6.32.(a), images are displayed using randomly generated coordinates. In Figure 6.32.(b) each image is positioned according to the first three principal components of a global color histogram descriptors. One can observe that images are scattered according to their predominant colors.

Preliminary tests show the efficiency of this platform in displaying large scale datasets. Future work consists mainly in integrating this browsing environment with the previous approaches and coming up with a final content-based multimedia information retrieval system.

Part II

Evolution and development of professional career

Chapter 7

Evolution and development of professional career

The future evolution and development of my professional career will follow naturally the directions presented in the previous parts of the habilitation thesis. I identify the following strategic directions whose schedule is synthesized with the Gantt diagram in Table 7.1:

- conducting fundamental and applicative research;
- increasing research visibility;
- increasing publications' impact;
- increasing participation to international projects;
- teaching and student coordination;
- coordination of a research group;
- technological transfer.

Each of these points is developed in the following.

7.1 Fundamental and applicative research

In what concerns my current and future research, it will mainly target the aspects related to the processing and analysis of multimedia information, multimedia retrieval and its large-/very large- scale and real-time implications, domain which gains more and more popularity due to the explosion of Internet and social media.

Whether more than one decade ago the availability of multimedia contents was very low, today we can witness a "multimedia revolution". Accessing multimedia content, i.e., images, sound, text or video, became part of our daily routine.

Table 7.1: Gantt diagram of the long term strategy for the development of the professional career.

action	**<2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**
Fundamental & applicative research	video: temp. segm., color, action, motion, and summarization.			multimedia: visual-audio-text, fusion, classification/clustering, machine learning, 3D interfaces, crowdsourcing.							multimedia: large scale retrieval system, database management, intelligent browsing interfaces.					
Increasing research visibility						EUSIPCO, CBMI, MM, MediaEval.		organizing benchmarking campaigns (MediaEval), conference committees: ICIP, EUVIP, IPTTA, MM, ICMR, tehnical programe committees, reviewer, editor.								
Increasing publications' impact	2 ISI	1 ISI		1 ISI		2 ISI	1 ISI	articles in high impact journals (> 12 ISI articles)								
Increasing part. to int. projects	Grant RP-2						Grant SCOUTER			participation to international projects, participation to strategic programmes, (4-6 projects proposals per year).						
Teaching & student coordination	PhD	Lecturer						Assoc. Prof.			Professor					
Coordination of a research group	Coord. of the video group with LAPI									Coord. of the multimedia lab with CAMPUS						
Technological transfer					UTI Grup video surveillance			Attracting companies in the field: Adobe, Google, Facebook, Samsung, Yahoo, etc.								

The technological advances of the acquisition and processing systems (e.g., mobile terminals, computational systems, data storage, sensing devices, etc) and information transmission protocols (e.g., wireless telecommunications, high-speed local area networks, multimedia telecommunication standards such as 3G and 4G, etc) led to an exponential increase of the transitioned multimedia data. An addition to this is the spreading of the Internet to more and diverse social medias together with the huge success and popularity of the on-line social platforms (e.g., Facebook, Twitter, Linkedln, Google+) and multimedia platforms (e.g., YouTube, DailyMotion, Picasa, Flickr, and so on).

Apart from the production of, lets say, commercial multimedia contents (i.e., produced by companies for commercial reasons, such as entertainment movie production), a new type of multimedia arrived, namely the personal multimedia data. The development of the on-line social media platforms eased the possibility for people around the world to share personal (amateur) data, e.g., sharing personal photos, private collection movies, video reports, video blogging and so on. These sources are actually significant multimedia content providers, e.g., the Facebook[1] social network sums in 2014 more that 1.3 billion users that upload and share various type of multimedia contents every day.

The data exchanging dynamics on the Internet is actually astonishing, everything taking place now in "real-time" from every multimedia terminal, mobile (e.g., smartphone) or fixed (e.g., personal computer). By a simple touch of a button video recordings or pictures are uploaded in the on-line environment. The following statistics are edificatory for the current scale of the multimedia "consumption" over the Internet: over 100 hours of video are uploaded every minute on the YouTube[2] platform, more than 600 years of videos from YouTube are being watched daily on Facebook, more than 900 video recordings from YouTube are shared every minute on the Twitter[3] platform. From the transitioned multimedia information, videos are the most frequently used over the Internet. It is estimated that by 2015 more than 1 million minutes of videos (674 days) will transit the Internet each second (source CISCO Systems[4]).

In this context, the actual issue is not the lack of information, but on the contrary, the difficulty to deal with this huge amalgam of media in order to provide the user with only the relevant data for its needs. We reached the point where this task cannot be performed manually by humans and it is imperatively necessary to address this in an automatic manner, e.g., by using computers.

[1] `https://www.facebook.com/`.
[2] `http://www.youtube.com/`.
[3] `https://twitter.com/`.
[4] `http://www.cisco.com`.

Figure 7.1: Content-based search example on Google Image Search for a "water lily" (image on the left). Images on the right are the first most relevant results by decreasing similarity (order from top to bottom and left to right).

This emerging area is know as *multimedia information retrieval* and it is situated at the confluence of domains such as signal processing, computer vision, pattern recognition and machine learning. Current research in the field targets mainly the development of algorithms and technologies for replicating, in an automatic manner, the *human understanding of multimedia contents*. This knowledge is to be used for achieving data searching capabilities that provide results close to the ones humans would select.

A potential solution to the problem of multimedia information retrieval was discussed earlier in the context of content-based retrieval of images and was relying on automatic indexing techniques. Transposed to the actual context, these techniques have now to adapt, on one hand, to the processing of a huge amount of data, e.g., 1 minute of video is the equivalent of 1,500 static images and therefore only one video sequence is the equivalent of an entire image collection; and on the other hand to deal with changing contents (temporal data) and multi-modal information (image-audio-text). Despite the current technology that allows for huge computational capabilities (e.g., a regular mobile phone features now several CPUs), the complexity of the multimedia information retrieval requires optimization and parallel computing to be able to perform with decent results. However, dealing with large-scale data is still an open issue.

To have an idea of the current retrieval capabilities, I take for example the image retrieval systems that have now more than 2 decades of existence (e.g., the Query By Image Content - QBIC system was released by IBM in 1995). Each of us searched over the Internet for images and noticed that the true content-based search capabilities, e.g., searching for images similar with an example image or the ability of specifying the image contents, are still far from the way a human would solve the problem (see the example in Figure 7.1). Multimedia retrieval capabilities, although mostly missing, are obviously significantly less developed and are mainly

limited to be extensions of the image retrieval ones (e.g., for video, image search is extended at individual frames without taking into account the temporal context or moving information).

Currently, there are no multimedia retrieval systems publicly available, the only existing ones being experimental, closed, adapted to specific domains (e.g., sports videos, news broadcasting, Hollywood productions, etc) and capable to perform only off-line due to the high computational complexity.

The existing web multimedia platforms are actually relying for performing the information retrieval solely on text data, such as content descriptors provided by users when uploading the data. This information has its limitations as it cannot be determined automatically (requiring user's input), is not always available, tends to be noisy and is not reflecting the actual data contents, e.g., users may tag an entire collection with the same description.

In this context, my long term research will target both fundamental and applicative research in this high impact emerging area. The final outcome is the possibility of developing a multimedia information retrieval system able to cope with large-/very large-scale multimedia data specific to Internet. As presented in the previous sections, results so far that addressed punctual aspects of such a system, starting from data pre-processing, multi-modal content description, indexing mechanisms and intelligent browsing interfaces, show very encouraging and promising perspectives. My research will focus mainly on bridging the sensor and semantic gaps of this system by developing:

- techniques for **spatio-temporal content description** of multimedia with specific emphasis on very low computational complexity and real-time performance;
- techniques for **multi-modal data fusion** of various information sources: visual, temporal, audio, text;
- techniques for **machine learning** adapted to multimedia such as exploration of the new paradigm of deep learning;
- hardware and software solutions for **real-time computation** such as parallel computing and distributed systems;
- high precision retrieval mechanisms that exploit the **human in the loop** concept which makes use of the human computational power via social platforms (e.g., crowdsourcing and gamification);
- intelligent **browsing interfaces** for multi-modal data visualization and search;

- a fully functional **multimedia retrieval system**.

These are high impact research directions and constitute the future of the multimedia information retrieval domain.

7.2 Increasing research visibility

Attendance to international forums is an important asset for maximizing the impact of the research in the scientific community and therefore contributing to the global progress. My strategy to strengthen and improve the impact of my research as well as to increase the institutional visibility at international level consist of:

- **benchmarking**: maintaining a healthy involvement with coordinating and organizing of international benchmarking campaigns, activity started in 2013 with the co-organizing of the Affect Task: Violent Scenes Detection and organizing of the Retrieving Diverse Social Images Task within the prestigious MediaEval - Benchmarking Initiative for Multimedia Evaluation. Benchmarking activities provide a framework for evaluating systems on a shared dataset and using a set of common rules. The results obtained are thus comparable and a wider community can benefit from it which widens significantly the impact of the research;

- **technical committees**: participating in the technical committees and organization of the main scientific conference and events from the field, e.g., ACM MM International Conference on Multimedia, IEEE ICIP International Conference on Image Processing, ACM ICMR International Conference on Multimedia Retrieval, IEEE/ACM CBMI International Workshop on Content-Based Multimedia Indexing, etc;

- **reviewing committees**: participating in the reviewing committees of the main journals from the field, e.g., IEEE Transactions on Multimedia, Multimedia Tools and Applications, IEEE Transactions on Image Processing. This will ensure a permanent contact with the latest innovations and current advances;

- **conference participation**: disseminating the main research results by attending the major conferences in the specific area of multimedia information retrieval.

7.3 Increasing publications' impact

One of the main institutional evaluation criteria is the number of scientific publications and their estimated impact factor in the field (e.g., Thomson Reuters ISI Impact Factor). In this respect, my short term objectives are:

- **high impact journal publications**: significantly increasing the number of publications in high ranked journals, some publications are currently under development, e.g., Elsevier Computer Vision and Image Understanding, Elsevier Image and Vision Computing, whereas some are already submitted being currently under review, e.g., IEEE Transactions on Circuits and Systems for Video Technology, Multimedia Tools and Applications. I target to achieve the acceptance of more than 2 articles each year;

- **number of citations**: increasing the impact of the research which is visible through the number of citations, i.e., the number of times the research findings were used by other researchers in their work. Maintaining a permanent dissemination and contact with the international community will contribute to significantly increase the number of citations. According to Google Scholar, my citations increased from 2009 with more than 86% (from 48 to 340) which corresponds to the period after my PhD when I was actively involved in the community. Given the impact of my current research, I expect that in the following 5 years to reach more than 1,000 citations.

7.4 Increasing participation to international projects

My participation and coordination of national research grants was more or less achieved at a constant rate. The critical point is the increase of the participation to international programmes such as the freshly started Horizon 2020 funding programme for which I am currently a technical program expert (for the Romanian commission).

My long term objective is to be able to coordinate such projects. My research activity so far allowed me to establish a vast network of collaborators around the world, see Section 3.5, which keeps me in a permanent contact with existing international projects (some of my students are developing their master or PhD theses within European projects) and project proposal perspectives. Each year I am involved with submitting 3-4 international project proposals which offers promising perspectives.

In parallel with the research projects, an important role is played also by the strategic programmes, such as the ones funded under European Structural Funds

(ESF) (e.g., infrastructure investments, human resources investments, etc). In this respect, my short and long term objective is to contribute to the development of the institution human resources and infrastructure and increase the absorbtion of ESF funds. A concrete example is my involvement with the managing team for the construction of the new research center of the University Politehnica of Bucharest, CAMPUS - Research Center for Advanced Materials, Products and Processes[5], funded under POS-CCE ESF axis, which is expected to be finalized in 2015.

7.5 Teaching and student coordination

In my opinion, research and didactic activities must me conducted and developed in close relation as these two are strongly connected. Valuable research achievements should be constantly included with the university curricula (e.g., courses, applications) to substantiate new research directions; whereas research should be funded on strong fundamental and theoretical basis using existing validated knowledge.

The development of my didactic career addresses in priority the transfer of the relevant research results with the curricula of the Faculty of Electronics, Telecommunications and Information Technology, by developing new laboratory platforms, scientific seminars, courses in the areas related to multimedia and information retrieval, student projects (semester projects, year projects, license and master thesis) as well as improving the existing ones; activities which I continuously carry over since the finalization of my PhD in 2007 (see Chapter 2).

At long term, given the high importance of this field, I consider the possibility of introducing a new master degree programme focusing on all the aspects of multimedia retrieval, starting form pre-processing, content description, retrieval to coding and database design. This is in-line with the industrial sectors where research on multimedia becomes more and more important with domains such as surveillance, entertainment, telecommunications, medical, military, etc.

7.6 Coordination of a research group

Research on multimedia information retrieval and its related areas is a major topic worldwide. There are currently many research groups in the international community dedicated to this field. Unfortunately, at national level, research on this topic is still developing. In this context, my strategy is to have the basis for a research group on multimedia information retrieval with the University Politehnica of Bucharest.

[5] `http://www.campus.upb.ro/`.

The prerogatives for this action were already created back in 2009 when thanks to the CNCSIS RP-2 research grant I developed a new research group on video processing and its infrastructure[6] as part of The Image Processing and Analysis Laboratory, Faculty of Electronics, Telecommunications and Information Technology. This initiative will continue with the construction of the University Politehnica of Bucharest's new research center, CAMPUS - Research Center for Advanced Materials, Products and Processes (to be finalized in 2015). The center will contain 41 research laboratories in various areas of engineering. In particular, I am in charge with the new Multimedia Content Processing and Analysis Laboratory which provides the perspective of leading an independent research group.

The ability of conducting PhD research plays a critical role in my strategy as PhD research constitutes the basis research unit for technological progress. Currently, I am involved with the coordination of PhD research for several students at the university as well as officially PhD co-advisor at the University of Trento, Italy. Obtaining the habilitation will allow me to take lead of PhD coordination and have the basis for a future research laboratory, attracting funding and research projects.

7.7 Technological transfer

The main objective of the didactic and research activities is the training of the new generations for the active life as well as pushing forward technology via technological transfer. Therefore, an important aspect is the permanent connection of the research with the industry, which is the potential beneficiary of the results.

In this respect, my strategy is to involve as much as possible companies from the field to stimulate potential technological transfer and innovation. Currently I am running a product-oriented innovation project with UTI Grup (2013-2015), Romanian leading company in the fields of security and video surveillance. I plan to extend this type of cooperation to other potential companies, e.g., Adobe Romania, Samsung Romania, Facebook Romania, Yahoo, etc and therefore to close a research-development-product cycle. This is even more important in the context of the new research laboratory I will coordinate in the Research Center for Advanced Materials, Products and Processes of the University Politehnica of Bucharest (see also the previous section).

[6] `http://imag.pub.ro/VideoIndexingRP2/`.

Bibliography

[1] B. Ionescu, A.-L. Radu, M. Menéndez, H. Müller, A. Popescu, B. Loni, "Div400: A Social Image Retrieval Result Diversification Dataset", ACM Multimedia Systems, 19-21 March, Singapore, 2014.

[2] C.-H. Demarty, C. Penet, B. Ionescu, G. Gravier, M. Soleymani, "Multimodal violence detection in Hollywood movies: State-of-the-art and Benchmarking", in book "Fusion in Computer Vision - Understanding Complex Visual Content", Springer International Publishing Switzerland ACVPR - Advances in Computer Vision and Pattern Recognition, 2014.

[3] C.A. Mitrea, I. Mironică, B. Ionescu, R. Dogaru, "Multiple Instance-based Object Retrieval in Video Surveillance: Dataset and Evaluation", IEEE International Conference on Intelligent Computer Communication and Processing, September 4-6, Cluj-Napoca, Romania, 2014.

[4] B. Ionescu, K. Seyerlehner, I. Mironică, C. Vertan, P. Lambert, "An Audio-Visual Approach to Web Video Categorization", Multimedia Tools and Applications, 70(2), pp 1007-1032 (DOI:10.1007/s11042-012-1097-x) 2014.

[5] C.-H. Demarty, B. Ionescu, Y.-G. Jiang, V.L. Quang, M. Schedl, C. Penet, "Benchmarking Violent Scenes Detection in Movies", IEEE International Workshop on Content-Based Multimedia Indexing - CBMI 2014, 18-20 June, Klagenfurt, Austria, 2014.

[6] A.-L. Radu, B. Ionescu, M. Menéndez, J. Stöttinger, F. Giunchiglia, A. De Angeli, "A Hybrid Machine-Crowd Approach to Photo Retrieval Result Diversification", 20th International Conference on MultiMedia Modeling - MMM2014, 8-10 January, Dublin, Ireland, 2014.

[7] B. Ionescu, A. Popescu, H. Müller, M. Menéndez, A.-L. Radu, "Benchmarking Result Diversification in Social Image Retrieval", IEEE International Conference on Image Processing - ICIP 2014, 27-30 October, Paris, France, 2014.

[8] A.L. Gînscă, A. Popescu, B. Ionescu, A. Armagan, I. Kanellos, "Toward Estimating User Tagging Credibility for Social Image Retrieval", ACM International Conference on Multimedia - ACM MM 2014, 3-7 November, Orlando, Florida, USA, 2014.

[9] B. Ionescu, I. Mironică, "The Content-Based Indexing Paradigm in the Context of Multimodal Data" (in romanian - "Conceptul de Indexare Automată dupaă Conţinut în Contextul Datelor Multimedia"), Publishing House "Editura MartrixRom", 2013.

[10] C. Florea, B. Ionescu, C. Vertan, "Computer Vision - Techniques for Digital Camera Calibration and Analysis of Visual Information" (in romanian - "Computer Vision - Tehnici de Calibrare a Camerei Digitale şi Analizei Informaţiei Vizuale"), Publishing House "Editura MatrixRom", ISBN 978-973-755-942-5, 2013.

[11] B. Ionescu, P. Lambert, "An Intensity-Driven Dissolve Detection Adapted to Synthetic Video Contents", SPIE - Journal of Electronic Imaging, 22(2), 023011, 2013.

[12] I. Mironică, J. Uijlings, N. Rostamzadeh, B. Ionescu, N. Sebe, "Time Matters! Capturing Variation in Time in Video using Fisher Kernels", ACM Multimedia, 21-25 October, Barcelona, Spain, 2013.

[13] I. Mironică, B. Ionescu, J. Uijlings, N. Sebe, "Fisher Kernel based Relevance Feedback for Multimodal Video Retrieval", ACM International Conference on Multimedia Retrieval - ICMR 2013, Dallas, Texas, USA, April 16 - 19, 2013.

[14] I. Mironică, B. Ionescu, P. Knees, P. Lambert, "An In-Depth Evaluation of Multimodal Video Genre Categorization", IEEE International Workshop on Content-Based Multimedia Indexing - CBMI 2013, 17-19 June, Veszprém, Hungary, 2013.

[15] B. Ionescu, J. Schlüter, I. Mironică, M. Schedl, "A Naive Mid-level Concept-based Fusion Approach to Violence Detection in Hollywood Movies", ACM International Conference on Multimedia Retrieval - ICMR 2013, Dallas, Texas, USA, April 16 - 19, 2013.

[16] B. Ionescu, K. Seyerlehner, C. Rasche, C. Vertan, P. Lambert, "Video Genre Categorization and Representation using Audio-Visual Information", SPIE - Journal of Electronic Imaging, 21(2), (DOI:10.1117/1.JEI.21.2.023017), 2012.

[17] A.-L. Radu, J. Stöttinger, B. Ionescu, M. Menéndez, F. Giunchiglia, "Representativeness and Diversity in Photos via Crowd-Sourced Media Analysis", 10th International Workshop on Adaptive Multimedia Retrieval - AMR 2012, October 24-25, Copenhagen, Denmark, 2012.

[18] I. Mironică, B. Ionescu, C. Vertan, "Hierarchical Clustering Relevance Feedback for Content-Based Image Retrieval", IEEE/ACM 10th International Workshop on Content-Based Multimedia Indexing - CBMI 2012, 27-29 June, Annecy, France, 2012.

[19] B. Ionescu, K. Seyerlehner, I. Mironică, C. Vertan, P. Lambert, "Automatic Web Video Categorization using Audio-Visual Information and Hierarchical Clustering Relevance Feedback", 20th European Signal Processing Conference - EUSIPCO 2012, August 27-31, Bucharest, Romania, 2012.

[20] I. Mironică, B. Ionescu, C. Vertan, "The Influence of the Similarity Measure to Relevance Feedback", 20th European Signal Processing Conference - EUSIPCO 2012, August 27-31, Bucharest, Romania, 2012.

[21] B. Ionescu, K. Seyerlehner, C. Rasche, C. Vertan, P. Lambert, "Content-based Video Description for Automatic Video Genre Categorization", 18th International Conference on MultiMedia Modeling - MMM 2012, 4-6 January, Klagenfurt, Austria, 2012.

[22] B. Ionescu, P. Lambert, D. Coquin, A. Marin, C. Vertan, "Analyzing Animated Movie Contents for Automatic Video Indexing", in Machine Learning Techniques for Adaptive Multimedia Retrieval: Technologies Applications and Perspectives, IGI - Global Printing House, Eds. Chia-Hung Wei et al., pp. 228-260, (ISBN 978-1-61692-859-9), 2011.

[23] B. Ionescu, C. Vertan, P. Lambert, "Dissolve Detection in Abstract Video Contents", IEEE ICASSP - International Conference on Acoustic, Speech and Signal Processing, Prague, Czech Republic, 22-27 May, 2011.

[24] B. Ionescu, C. Vertan, P. Lambert, A. Benoit, "A Color-Action Perceptual Approach to the Classification of Animated Movies", ACM International Conference on Multimedia Retrieval, Trento, Italy, 17-20 April, 2011.

[25] B. Ionescu, L. Ott, P. Lambert, D. Coquin, A. Pacureanu, V. Buzuloiu, "Tackling Action - Based Video Abstraction of Animated Movies for Video Browsing", SPIE - Journal of Electronic Imaging, 19(3), 2010.

[26] B. Ionescu, "The Analysis and Processing of Video Sequences: Automatic Content-Based Indexing" (in romanian - "Analiza şi Prelucrarea Secvenţelor Video: Indexarea Automată după Conţinut"), Publishing House "Editura Tehnică Bucureşti", ISBN 978-973-31-2354-5, 2009.

[27] B. Ionescu, D. Coquin, P. Lambert, V. Buzuloiu, "A Fuzzy Color-Based Approach for Understanding Animated Movies Content in the Indexing Task", Eurasip Journal on Image and Video Processing, doi:10.1155/2008/849625, 2008.

[28] B. Ionescu and V. Buzuloiu and P. Lambert and D. Coquin, "Improved Cut Detection for the Segmentation of Animation Movies", IEEE International Conference on Acoustic, Speech and Signal Processing, Toulouse, France, 2006.

[29] B. Han and X. Gao and H. Ji, "A Unified Framework for Shot Boundary Detection", Springer LNCS Pattern Recognition, 3801, pp. 997-1002, 2005.

[30] Cees G.M. Snoek and M. Worring, "Multimodal Video Indexing: A Review of the State-of-the-art", Multimedia Tools and Applications, 25(1), pp. 5-35, 2005.

[31] A. F. Smeaton and P. Over and W. Kraaij, “High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements”, Multimedia Content Analysis, Theory and Applications, Springer Verlag, ISBN 978-0-387-76567-9, pp. 151-174, 2009.

[32] CITIA - City of Moving Images, http://www.citia.info/, http://www.citia.info/.

[33] B.T. Truong and C. Dorai and S. Venkatesh, “New Enhancements to Cut, Fade, and Dissolve Detection Processes in Video Segmentation”, ACM Multimedia; Los Angeles, CA, USA, pp. 219-227, 2000.

[34] R. Zabih and J. Miller and K. Mai, “A Feature-Based Algorithm for Detecting and Classification Production Effects”, Multimedia Systems, 7, pp. 119-128, 1999.

[35] C.W. Su and H.-Y.M. Liao and H.R. Tyan and K.C. Fan and L.H. Chen, “A Motion-Tolerant Dissolve Detection Algorithm”, IEEE Transactions on Multimedia, 7(6), pp. 1106-1113, 2005.

[36] X. Yuan, W. Lai, T. Mei, X.-S. Hua, X.-Q. Wu, S. Li, “Automatic Video Genre Categorization using Hierarchical SVM”, IEEE International Conference on Image Processing, pp. 2905-2908, 2006.

[37] D. Brezeale, D.J. Cook, “Using Closed Captions and Visual Features to Classify Movies by Genre”, International Workshop on Multimedia Data Mining, 2006.

[38] T. Sikora, “The MPEG-7 Visual Standard for Content Description - An Overview”, IEEE Transactions on Circuits and Systems for Video Technology, 11(6), pp. 696-702, 2001.

[39] C. Rasche, “An Approach to the Parameterization of Structure for Fast Categorization”, International Journal of Computer Vision, 87(3), pp. 337-356, 2010.

[40] K. Seyerlehner, G. Widmer, T. Pohle, “Fusing Block-level Features for Music Similarity Estimation”, 13th International Conference on Digital Audio Effects, Graz, Austria, 2010.

[41] R. Lienhart, “Reliable Transition Detection in Videos: A Survey and Practitioners Guide”, International Journal of Image and Graphics, 1(3), pp. 469-486, 2001.

[42] W.A.C. Fernando, C.N. Canagarajah, D.R. Bull, “Fade and Dissolve Detection in Uncompressed and Compressed Video Sequence”, IEEE International Conference on Image Processing, Kobe, Japan, pp. 299-303, 1999.

[43] P. Kay, T. Regier, “Resolving the Question of Color Naming Universals”, Proceedings of the National Academy of Sciences of the United States of America, 100(15), pp. 9085-9089, 2003.

[44] J. Itten, "The Art of Color: The Subjective Experience and Objective Rationale of Color", Reinhold, New York, NY, USA, 1961.

[45] A. Hauptmann, R. Yan, Y. Qi, R. Jin, M. Christel, M. Derthick, M.- Y. Chen, R. Baron, W.-H. Lin, T. D. Ng, "Video Classification and Retrieval with the Informedia Digital Video Library System", Text Retrieval Conference, 2002.

[46] J. Canny, "A Computational Approach To Edge Detection", IEEE Trans. on Pattern Analysis and Machine Intelligence, 8(6), pp. 679-698, 1986.

[47] I. H. Witten, E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques", Second Edition, Eds. Morgan Kaufmann, ISBN 0-12-088407-0, 2005.

[48] M. Larson, M. Eskevich, R. Ordelman, C. Kofler, S. Schmiedeke, G.J.F. Jones, "Overview of MediaEval 2011 Rich Speech Retrieval Task and Genre Tagging Task", Working Notes Proceedings of the MediaEval 2011 Workshop at Interspeech 2011, vol. 807, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-807/, Pisa, Italy, September 1-2, 2011.

[49] Weka Data Mining with Open Source Machine Learning Software in Java, University of Waikato, http://www.cs.waikato.ac.nz/ml/weka/, 2014.

[50] M. Montagnuolo, A. Messina, "Parallel Neural Networks for Multimodal Video Genre Classification", Multimedia Tools and Applications, 41(1), pp. 125-159, 2009.

[51] M. Rouvier, G. Linares, "LIA @ MediaEval 2011: Compact representation of heterogeneous descriptors for video genre classification", Working Notes Proceedings of the MediaEval 2011 Workshop at Interspeech 2011, vol. 807, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-807/Rouvier_LIA_Genre_me11wn.pdf, Pisa, Italy, September 1-2, 2011.

[52] J. Perea-Ortega, A. Montejo-Raez, M. Diaz-Galiano, M.T. Martin-Valdivia, "Genre tagging of videos based on information retrieval and semantic similarity using WordNet", Working Notes Proceedings of the MediaEval 2011 Workshop at Interspeech 2011, vol. 807, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-807/Perea-Ortega_SINAI_Genre_me11wn.pdf, Pisa, Italy, September 1-2, 2011.

[53] R. Tiwari, C. Zhang, M. Montes, "UAB at MediaEval 2011: Genre Tagging Task", Working Notes Proceedings of the MediaEval 2011 Workshop at Interspeech 2011, vol. 807, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-807/Tiwari_UAB_Genre_me11wn.pdf, Pisa, Italy, September 1-2, 2011.

[54] J.M. Cigarran Recuero, V. Fresno Fernandez, A. Garcia-Serrano, D. Hernandez Aranda, R. Granados, "UNED at MediaEval 2011: Can Delicious help us to improve automatic video tagging?", Working Notes Proceedings

of the MediaEval 2011 Workshop at Interspeech 2011, vol. 807, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-807/Cigarran_UNED2011_Genre_me11wn.pdf, Pisa, Italy, September 1-2, 2011.

[55] S. Rudinac, M. Larson, A. Hanjalic, "TUD-MIR at MediaEval 2011 Genre Tagging Task: Query expansion from a limited number of labeled videos", Working Notes Proceedings of the MediaEval 2011 Workshop at Interspeech 2011, vol. 807, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-807/Rudinac_TUD_Genre_me11wn.pdf, Pisa, Italy, September 1-2, 2011.

[56] T. Semela, H.K. Ekenel, "KIT at MediaEval 2011 - Content-based genre classification on web-videos", Working Notes Proceedings of the MediaEval 2011 Workshop at Interspeech 2011, vol. 807, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-807/Semela_KIT_Genre_me11wn.pdf, Pisa, Italy, September 1-2, 2011.

[57] B. Ionescu, K. Seyerlehner, C. Vertan, P. Lambert, "Audio-Visual content description for video genre classification in the context of social media", Working Notes Proceedings of the MediaEval 2011 Workshop at Interspeech 2011, vol. 807, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-807/Ionescu_RAF_Genre_me11wn.pdf, Pisa, Italy, September 1-2, 2011.

[58] S. Schmiedeke, P. Kelm, T. Sikora, "TUB @ MediaEval 2011 Genre Tagging Task: Prediction using bag-of–words approaches", Working Notes Proceedings of the MediaEval 2011 Workshop at Interspeech 2011, vol. 807, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-807/Schmiedeke_TUB_Genre_me11wn.pdf, Pisa, Italy, September 1-2, 2011.

[59] T. Jaakkola, D. Haussler, "Exploiting generative models in discriminative classifiers", NIPS, 1999.

[60] F. Perronnin, J. Sanchez, T. Mensink, "Improving the Fisher Kernel for Large-Scale Image Classification", European Conference on Computer Vision, 2010.

[61] G. Csurka, C.R. Dance, L. Fan, J. Willamowski, C. Bray, "Visual Categorization with Bags of Keypoints", ECCV Workshop on Statistical Learning in CV, 2004.

[62] S. Schmiedeke, C. Kofler, I. Ferrané, "Overview of the MediaEval 2012 Tagging Task", Working Notes Proceedings of the MediaEval Workshop, vol. 927, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-927/, Pisa, Italy, 4-5 October, 2012.

[63] B. Ionescu, I. Mironică, K. Seyerlehner, P. Knees, J. Schlüter, M. Schedl, H. Cucu, A. Buzo, P. Lambert, "ARF @ MediaEval 2012: Multimodal Video Classification", Working Notes Proceedings of the MediaEval Workshop, vol. 927, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-927/mediaeval2012_submission_7.pdf, Pisa, Italy, 4-5 October, 2012.

[64] T. Semela, M. Tapaswi, H.K. Ekenel, R. Stiefelhagen, "KIT at MediaEval 2012 - Content-based Genre Classification with Visual Cues", Working Notes Proceedings of the MediaEval Workshop, vol. 927, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-927/mediaeval2012_submission_10.pdf, Pisa, Italy, 4-5 October, 2012.

[65] S. Schmiedeke, P. Kelm, T. Sikora, "TUB @ MediaEval 2012 Tagging Task: Feature Selection Methods for Bag-of-(visual)-Words Approaches", Working Notes Proceedings of the MediaEval Workshop, vol. 927, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-927/mediaeval2012_submission_9.pdf, Pisa, Italy, 4-5 October, 2012.

[66] Y. Shi, M.A. Larson, P. Wiggers, C.M. Jonker, "MediaEval 2012 Tagging Task: Prediction based on One Best List and Confusion Networks", Working Notes Proceedings of the MediaEval Workshop, vol. 927, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-927/mediaeval2012_submission_12.pdf, Pisa, Italy, 4-5 October, 2012.

[67] J. Almeida, T. Salles, E. Martins, O. Penatti, R. da S. Torres, M. Gonçalves, J. Almeida, "UNICAMP-UFMG at MediaEval 2012: Genre Tagging Task", Working Notes Proceedings of the MediaEval Workshop, vol. 927, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-927/mediaeval2012_submission_13.pdf, Pisa, Italy, 4-5 October, 2012.

[68] P. Xu, Y. Shi, M.A. Larson, "TUD at MediaEval 2012 genre tagging task: Multimodality video categorization with one-vs-all classifiers", Working Notes Proceedings of the MediaEval Workshop, vol. 927, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-927/mediaeval2012_submission_11.pdf, Pisa, Italy, 4-5 October, 2012.

[69] K. Reddy, M. Shah, "Recognizing 50 human action categories of web videos", Machine Vision and Applications Journal, 2012.

[70] O. Kliper-Gross, Y. Gurovich, T. Hassner, L. Wolf, "Motion Interchange Patterns for Action Recognition in Unconstrained Videos", European Conference on Computer Visions, 2012.

[71] B. Solmaz, S. Modiri Assari, M. Shah, "Classifying web videos using a global video descriptor", Machine Vision and Applications Journal, 2012.

[72] I. Everts, J. van Gemert, T. Gevers, "Evaluation of Color STIPs for Human Action Recognition", IEEE International Conference on Computer Vision and Pattern Recognition, 2013.

[73] R. Messing, C. Pal, H. Kautz, "Activity recognition using the velocity histories of tracked keypoints", International Conference on Computer Vision, 2009.

[74] M. Rohrbach, S. Amin, M. Andriluka, B. Schiele, "A database for fine grained activity detection of cooking activities", IEEE International Conference on Computer Vision and Pattern Recognition, 2012.

[75] Y. Yang, D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts", IEEE International Conference on Computer Vision and Pattern Recognition, 2011.

[76] J. Wang, Z. Chen, Y. Wu, "Action Recognition with Multiscale Spatio-Temporal Contexts", IEEE International Conference on Computer Vision and Pattern Recognition, 2011.

[77] Z. Lin, Z. Jiang, L. Davis, "Recognizing actions by shape-motion prototype trees", International Conference on Computer Vision, 2009.

[78] J.R.R. Uijlings, A.W.M. Smeulders, R.J.H. Scha, "Real-Time Visual Concept Classification", IEEE Transactions on Multimedia, 99, 2010.

[79] A.G. Money, H. Agius, "Video Summarisation: A Conceptual Framework and Survey of the State of the Art", International Journal of Visual Communication and Image Representation, 19, pp. 121-143, 2008.

[80] B.T. Truong, S. Venkatesh, "Video abstraction: A systematic review and classification", ACM Transactions on Multimedia Computing Communications and Applications, 3(1), pp. 3, 2007.

[81] H.W. Chen, J.-H. Kuo, W.-T. Chu, J.-L. Wu, "Action Movies Segmentation and Summarization based on Tempo Analysis", ACM International Workshop on Multimedia Information Retrieval, pp. 251-258, New York, 2004.

[82] C.-H. Demarty, C. Penet, G. Gravier, M. Soleymani, "The MediaEval 2012 Affect Task: Violent Scenes Detection in Hollywood Movies", Working Notes Proceedings of the MediaEval Workshop, vol. 927, CEUR-WS.org, ISSN 1613-0073, `http://ceur-ws.org/Vol-927/mediaeval2012_submission_3.pdf`, Pisa, Italy, 4-5 October, 2012.

[83] D. E. Rumelhart, G. E. Hinton, R. J. Williams, "Learning Representations by Back-Propagating Errors", Nature, 323, pp. 533–536, 1986.

[84] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, "Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors", arXiv.org, `http://arxiv.org/abs/1207.0580`, 2012.

[85] Technicolor, http://www.technicolor.com.

[86] Yaafe core features, http://yaafe.sourceforge.net/.

[87] C. Liu, L. Xie, H. Meng, "Classification of Music and Speech in Mandarin News Broadcasts", Int. Conf. on Man-Machine Speech Communication, China, 2007.

[88] O. Ludwig, D. Delgado, V. Goncalves, U. Nunes, "Trainable Classifier-Fusion Schemes: An Application To Pedestrian Detection", IEEE Int. Conf. On Intelligent Transportation Systems, 1, pp. 432-437, St. Louis, 2009.

[89] J. Van de Weijer, C. Schmid, J. Verbeek, D. Larlus, "Learning color names for real-world applications", IEEE Transactions on Image Processing, 18(7), pp. 1512-1523, 2009.

[90] V. Lam, D.-D. Le, S.-P. Le, Shin'ichi Satoh, D.A. Duong, "NII Japan at MediaEval 2012 Violent Scenes Detection Affect Task", Working Notes Proceedings of the MediaEval Workshop, vol. 927, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-927/mediaeval2012_submission_21.pdf, Pisa, Italy, 4-5 October, 2012.

[91] F. Eyben, F. Weninger, N. Lehment, G. Rigoll, B. Schuller, "Violent Scenes Detection with Large, Brute-forced Acoustic and Visual Feature Sets", Working Notes Proceedings of the MediaEval Workshop, vol. 927, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-927/mediaeval2012_submission_25.pdf, Pisa, Italy, 4-5 October, 2012.

[92] Y.-G. Jiang, Q. Dai, C.C. Tan, X. Xue, C.-W. Ngo, "The Shanghai-Hongkong Team at MediaEval2012: Violent Scene Detection Using Trajectory-based Features", Working Notes Proceedings of the MediaEval Workshop, vol. 927, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-927/mediaeval2012_submission_28.pdf, Pisa, Italy, 4-5 October, 2012.

[93] N. Derbas, F. Thollard, B. Safadi, G. Quénot, "LIG at MediaEval 2012 Affect Task: use of a Generic Method", Working Notes Proceedings of the MediaEval Workshop, vol. 927, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-927/mediaeval2012_submission_39.pdf, Pisa, Italy, 4-5 October, 2012.

[94] V. Martin, H. Glotin, S. Paris, X. Halkias, J.-M. Prevot, "Violence Detection in Video by Large Scale Multi-Scale Local Binary Pattern Dynamics", Working Notes Proceedings of the MediaEval Workshop, vol. 927, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-927/mediaeval2012_submission_43.pdf, Pisa, Italy, 4-5 October, 2012.

[95] C. Penet, C.-H. Demarty, M. Soleymani, G. Gravier, P. Gros , "Technicolor/INRIA/Imperial College London at the MediaEval 2012 Violent Scene Detection Task", Working Notes Proceedings of the MediaEval Workshop, vol. 927, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-927/mediaeval2012_submission_26.pdf, Pisa, Italy, 4-5 October, 2012.

[96] J. Schlüter, B. Ionescu, I. Mironică, M. Schedl , "ARF @ MediaEval 2012: An Uninformed Approach to Violence Detection in Hollywood Movies", Working Notes Proceedings of the MediaEval Workshop, vol. 927, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-927/mediaeval2012_submission_36.pdf, Pisa, Italy, 4-5 October, 2012.

[97] E. Açar, S. Albayrak, "DAI Lab at MediaEval 2012 Affect Task: The Detection of Violent Scenes using Affective Features", Working Notes Proceedings of the MediaEval Workshop, vol. 927, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-927/mediaeval2012_submission_33.pdf, Pisa, Italy, 4-5 October, 2012.

[98] S. Paris, H. Glotin, "Pyramidal Multi-level Features for the Robot Vision @ICPR 2010 Challenge", 20th Int. Conf. on Pattern Recognition, pp. 2949 - 2952, Marseille, France, 2010.

[99] C. Penet, C.-H. Demarty, G. Gravier, P. Gros, "Multimodal Information Fusion and Temporal Integration for Violence Detection in Movies", IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Kyoto, 2012.

[100] C.-H. Demarty, C. Penet, M. Schedl, B. Ionescu, V.L. Quang, Y.-G. Jiang, "The MediaEval 2013 Affect Task: Violent Scenes Detection", Working Notes Proceedings of the MediaEval Workshop, vol. 1043, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_4.pdf, Barcelona, Spain, October 18-19, 2013.

[101] Q. Dai, J. Tu, Z. Shi, Y.-G. Jiang, X. Xue, "Fudan at MediaEval 2013: Violent Scenes Detection using Motion Features and Part-Level Attributes", Working Notes Proceedings of the MediaEval Workshop, vol. 1043, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_6.pdf, Barcelona, Spain, October 18-19, 2013.

[102] M. Sjöberg, J. Schlüter, B. Ionescu, M. Schedl, "FAR at MediaEval 2013 Violent Scenes Detection: Concept-based Violent Scenes Detection in Movies", Working Notes Proceedings of the MediaEval Workshop, vol. 1043, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_10.pdf, Barcelona, Spain, October 18-19, 2013.

[103] I. Serrano, O. Déniz, G. Bueno, "VISILAB at MediaEval 2013: Fight Detection", Working Notes Proceedings of the MediaEval Workshop, vol. 1043, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_11.pdf, Barcelona, Spain, October 18-19, 2013.

[104] C.C. Tan, C.-W. Ngo, "The Vireo Team at MediaEval 2013: Violent Scenes Detection by Mid-level Concepts Learnt from Youtube", Working Notes Proceedings of the MediaEval Workshop, vol. 1043, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_12.pdf, Barcelona, Spain, October 18-19, 2013.

[105] N. Derbas, B. Safadi, G. Quénot, "LIG at MediaEval 2013 Affect Task: Use of a Generic Method and Joint Audio-Visual Words", Working Notes Proceedings of the MediaEval Workshop, vol. 1043, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_13.pdf, Barcelona, Spain, October 18-19, 2013.

[106] S. Goto, T. Aoki, "TUDCL at MediaEval 2013 Violent Scenes Detection: Training with Multimodal Features by MKL", Working Notes Proceedings of the MediaEval Workshop, vol. 1043, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_14.pdf, Barcelona, Spain, October 18-19, 2013.

[107] B.D.N. Teixeira, "MTM at MediaEval 2013 Violent Scenes Detection: Through Acoustic-visual Transform", Working Notes Proceedings of the MediaEval Workshop, vol. 1043, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_15.pdf, Barcelona, Spain, October 18-19, 2013.

[108] V. Lam, D.-D. Le, S. Phan, S. Satoh, D.A. Duong, "NII-UIT at MediaEval 2013 Violent Scenes Detection Affect Task", Working Notes Proceedings of the MediaEval Workshop, vol. 1043, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_27.pdf, Barcelona, Spain, October 18-19, 2013.

[109] C. Penet, C.-H. Demarty, G. Gravier, P. Gros, "Technicolor/INRIA Team at the MediaEval 2013 Violent Scenes Detection Task", Working Notes Proceedings of the MediaEval Workshop, vol. 1043, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_31.pdf, Barcelona, Spain, October 18-19, 2013.

[110] A. García Seco de Herrera, J. Kalpathy-Cramer, D. Demner Fushman, S. Antani, H. Müller, Overview of the ImageCLEF 2013 medical tasks, Working Notes of CLEF 2013 (Cross Language Evaluation Forum), Valencia, Spain, 2013.

[111] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, "Content-based image retrieval at the end of the early years", IEEE Transaction on Pattern Analysis and Machine Intelligence, 22(12), pp. 1349 - 1380, 2000.

[112] R. Datta, D. Joshi, J. Li, J.Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age", ACM Comput. Surv., 40(2), pp. 1-60, 2008.

[113] R. Priyatharshini, S. Chitrakala, "Association Based Image Retrieval: A Survey", Mobile Communication and Power Engineering, Springer Communications in Computer and Information Science, 296, pp 17-26, 2013.

[114] R.H. van Leuken, L. Garcia, X. Olivares, R. van Zwol, "Visual Diversification of Image Search Results", ACM World Wide Web, pp. 341-350, 2009.

[115] M.L. Paramita, M. Sanderson, P. Clough, "Diversity in Photo Retrieval: Overview of the ImageCLEF Photo Task 2009", ImageCLEF 2009.

[116] B. Taneva, M. Kacimi, G. Weikum, "Gathering and Ranking Photos of Named Entities with High Precision, High Recall, and Diversity", ACM Web Search and Data Mining, pp. 431-440, 2010.

[117] S. Rudinac, A. Hanjalic, M.A. Larson, "Generating Visual Summaries of Geographic Areas Using Community-Contributed Images", IEEE Transactions on Multimedia, 15(4), pp. 921-932, 2013.

[118] S. Nowak and S. Rüger, "How reliable are annotations via crowd-sourcing? a study about inter-annotator agreement for multi-label image annotation", International Conference on Multimedia Information Retrieval, 2010.

[119] A. Kittur, E. H. Chi and and B. Suh, "Crowd-sourcing user studies with Mechanical Turk", SIGCHI Conference on Human Factors in Computing Systems, pages 453–456, Italy, 2008.

[120] L. S. Kennedy and M. Naaman, "Generating diverse and representative image search results for landmarks", International Conference on World Wide Web, pages 297-306, China, 2008.

[121] J. J. Randolph, R. Bednarik and N. Myller, "Author Note: Free-Marginal Multirater Kappa (multirater kfree): An Alternative to Fleiss' Fixed - Marginal Multirater Kappa", `http://files.eric.ed.gov/fulltext/ED490661.pdf`, last accessed on 10-08-2014.

[122] J. A. Noble, "Minority voices of crowd-sourcing: why we should pay attention to every member of the crowd", ACM Conference on Computer Supported Cooperative Work Companion, pages 179-182, USA, 2012.

[123] B. Ionescu, M. Menéndez, H. Müller, A. Popescu, "Retrieving Diverse Social Images at MediaEval 2013: Objectives, Dataset and Evaluation", Working Notes Proceedings of the MediaEval Workshop, vol. 1043, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_3.pdf, Barcelona, Spain, October 18-19, 2013.

[124] T. Tsikrika, J. Kludas, A. Popescu, "Building Reliable and Reusable Test Collections for Image Retrieval: The Wikipedia Task at ImageCLEF", IEEE Multimedia, 19(3), pp. 24-33, 2012.

[125] A. Popescu, G. Grefenstette, "Social Media Driven Image Retrieval", ACM ICMR, April 17-20, Trento, Italy, 2011.

[126] J.J. Randolph, "Free-Marginal Multirater Kappa (multirater κfree): an Alternative to Fleiss Fixed-Marginal Multirater Kappa", Joensuu Learning and Instruction Symposium, 2005.

[127] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit", Psychological Bulletin, Vol. 70(4), pp. 213-220, 1968.

[128] N. Jain, J. Hare, S. Samangooei, J. Preston, J. Davies, D. Dupplaw, P. Lewis, "Experiments in Diversifying Flickr Result Sets", Working Notes Proceedings of the MediaEval Workshop, vol. 1043, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_18.pdf, Barcelona, Spain, October 18-19, 2013.

[129] C. Kuoman, S. Tollari, M. Detyniecki, "UPMC at MediaEval 2013: Relevance by Text and Diversity by Visual Clustering", Working Notes Proceedings of the MediaEval Workshop, vol. 1043, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_19.pdf, Barcelona, Spain, October 18-19, 2013.

[130] H.J. Escalante, A. Morales-Reyes, "TIA-INAOE's Approach for the 2013 Retrieving Diverse Social Images Task", Working Notes Proceedings of the MediaEval Workshop, vol. 1043, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_20.pdf, Barcelona, Spain, October 18-19, 2013.

[131] A-L. Radu, B. Boteanu, O. Pleş, B. Ionescu, "LAPI @ Retrieving Diverse Social Images Task 2013: Qualitative Photo Retrieval using Multimedia Content", Working Notes Proceedings of the MediaEval Workshop, vol. 1043, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_21.pdf, Barcelona, Spain, October 18-19, 2013.

[132] G. Szűcs, Z. Paróczi, D.M. Vincz, "BMEMTM at MediaEval 2013 Retrieving Diverse Social Images Task: Analysis of Text and Visual Information", Working Notes Proceedings of the MediaEval Workshop, vol. 1043, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_23.pdf, Barcelona, Spain, October 18-19, 2013.

[133] D. Corney, C. Martin, A. Göker, E. Spyromitros-Xioufis, S. Papadopoulos, Y. Kompatsiaris, L. Aiello, B. Thomee, "SocialSensor: Finding Diverse Images at MediaEval 2013", Working Notes Proceedings of the MediaEval Workshop, vol. 1043, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_24.pdf, Barcelona, Spain, October 18-19, 2013.

[134] B. Vandersmissen, A. Tomar, F. Godin, W. De Neve, R. Van de Walle, "Ghent University-iMinds at MediaEval 2013 Diverse Images: Relevance-Based Hierarchical Clustering", Working Notes Proceedings of the MediaEval Workshop, vol. 1043, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_25.pdf, Barcelona, Spain, October 18-19, 2013.

[135] A. Armagan, A. Popescu, P. Duygulu, "MUCKE Participation at Retrieving Diverse Social Images Task of MediaEval 2013", Working Notes Proceedings of the MediaEval Workshop, vol. 1043, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_29.pdf, Barcelona, Spain, October 18-19, 2013.

[136] K. Yanai, D.H. Nga, "UEC, Tokyo at MediaEval 2013 Retrieving Diverse Social Images Task", Working Notes Proceedings of the MediaEval Workshop, vol. 1043, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_30.pdf, Barcelona, Spain, October 18-19, 2013.

[137] A. Popescu, "CEA LIST's Participation at the MediaEval 2013 Retrieving Diverse Social Images Task", Working Notes Proceedings of the MediaEval Workshop, vol. 1043, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_43.pdf, Barcelona, Spain, October 18-19, 2013.

[138] A. Bursuc, T. Zaharia, "ARTEMIS @ MediaEval 2013: A Content-Based Image Clustering Method for Public Image Repositories", Working Notes Proceedings of the MediaEval Workshop, vol. 1043, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_48.pdf, Barcelona, Spain, October 18-19, 2013.

[139] A. Lehman, "Jmp For Basic Univariate And Multivariate Statistics: A Step-by-step Guide", Cary, NC: SAS Press. p. 123. ISBN 1-59047-576-3, 2005.

[140] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A.F. Smeaton, G. Quéenot, "TRECVID 2013 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics", Proceedings of TRECVID 2013, `http://www-nlpir.nist.gov/projects/tvpubs/tv13.papers/tv13overview.pdf`, NIST, USA, 2013.

[141] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results", `http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html`, 2012.

[142] T. Tsikrika, Theodora, A. García Seco de Herrera, H. Müller, "Assessing the Scholarly Impact of ImageCLEF", Springer Lecture Notes in Computer Science (LNCS), pages 95-106, 2011.

[143] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database", IEEE International Conference on Computer Vision and Patern Recognition, 2009.

[144] Y. Rui, T. S. Huang, M. Ortega, M. Mehrotra, S. Beckman, "Relevance Feedback: a Power Tool for Interactive Content-Based Image Retrieval", IEEE Transactions on Circuits and Video Technology, 8(5), pp. 644-655, 1998.

[145] N. V. Nguyen, J.-M. Ogier, S. Tabbone, A. Boucher, "Text Retrieval Relevance Feedback Techniques for Bag-of-Words Model in CBIR", International Conference on Machine Learning and Pattern Recognition, 2009.

[146] D.-H. Kim and C.-W. Chung, "Qcluster: Relevance Feedback Using Adaptive Clustering for Content Based Image Retrieval", ACM Conference on Management of Data, 2003.

[147] S. Liang, Z. Sun, "Sketch Retrieval and Relevance Feedback with Biased SVM Classification", Pattern Recognition Letters, 29, pp. 1733-1741, 2008.

[148] S.D. MacArthur, C.E. Brodley, C.-R. Shyu, "Interactive Content-Based Image Retrieval Using Relevance Feedback", Computer Vision and Image Understanding, 88(2), pp. 55-75, 2002.

[149] S.H. Huang, Q.J Wu, S.H. Lu, "Improved AdaBoost-Based Image Retrieval with Relevance Feedback via Paired Feature Learning", ACM Multimedia Systems, 12(1), pp. 14-26, 2006.

[150] Y. Wu, A. Zhang, "Interactive Pattern Analysis for Relevance Feedback in Multimedia Information Retrieval", Journal on Multimedia Systems, 10(1), pp. 41-55, 2004.

[151] J. Ye, J. Chow, J. Chen, Z. Zheng, "Stochastic Gradient Boosted Distributed Decision Trees", ACM Conference on Information and Knowledge Management, 2009.

[152] G. Giacinto, "A Nearest-Neighbor Approach to Relevance Feedback in Content-Based Image Retrieval", ACM International Conference on Image and Video Retrieval, 2007.

[153] F. Perronnin, J. Sanchez, T. Mensink, "Improving the Fisher Kernel for Large-Scale Image Classification", European Conference on Computer Vision, 2010.

[154] F. Perronnin, J.A. Rodriguez-Serrano, "Fisher Kernels for Handwritten Word-spotting", International Conference on Document Analysis and Recognition, 2009.

[155] P. Moreno, R. Rifkin, "Using the Fisher kernel method for web audio classification", International Conference on Acoustics, Speech and Signal Processing, pp. 2417-2420, 2000.

[156] P. Kelm, S. Schmiedeke, T. Sikora, "Feature-based video key frame extraction for low quality video sequences", Workshop on Image Analysis for Multimedia Interactive Services, 2009.