THÈSE D'HABILITATION À DIRIGER DES RECHERCHES

présentée par
Benoit HUET

Étude de Contenus Multimédia: Apporter du Contexte au Contenu

Université Nice-Sophia Antipolis
Specialitée : DS9 STIC
(Sciences et Technologies de l'Information et de la Communication)

Soutenue le 3 Octobre 2012

Composition du jury:

Rapporteurs:

Prof.	Tat-Seng CHUA	National University of Singapore, Singapour
Prof.	Patrick GROS	INRIA, Rennes – France
Prof.	Alan SMEATON	Dublin City University, Dublin– Irlande

Examinateurs:

Prof.	Edwin HANCOCK	University of York, York – Royaume Uni
Prof.	Bernard MERIALDO	EURECOM, Sophia Antipolis - France
Prof.	Nicu SEBE	University of Trento, Trento – Italie

Multimedia Content Understanding: Bringing Context to Content

by

Benoit HUET

March 2012

Acknowledgement

Although this thesis describes my research activities in the field of multimedia content understanding, it is by no means the work of a single person. I am thankful to my former and current PhD students: Itheri Yahiaoui, Fabrice Souvannavong, Joakim Jiten, Eric Galmar, Rachid Benmokhtar, Marco Paleari, Stephane Turlier, Xueliang Liu and Mathilde Sahuguet for their contributions and for enabling me to achieve my research goals.

My thanks also go to my collegues at Eurecom's Multimedia Department: Prof. Bernard Merialdo, Prof. Jean-Luc Dugelay, Prof. Nick Evans and Prof. Raphael Troncy for contributing to the stimulating research atmosphere at work.

I would like to express my gratitude to my thesis committee: Prof. Tat-Seng Chua, Prof. Alan Smeaton, Prof. Patrick Gros, Prof Edwin Hancock, Prof. Bernard Merialdo and Prof. Nicu Sebe for the interresting discussions concerning my research. I particularly appreciated the insightful comments provided by Profs. Gros, Smeaton and Chua on this manuscript.

Last but not the least, I would like to thank my family for their invaluable daily support.

Contents

1 Curriculum Vitae

Benoit HUET
Assistant Professor
Multimedia Communications Department,
EURECOM,
2229 Route des Crêtes
BP 193
06904 Sophia-Antipolis
France

Phone: +33 (0)4.93.00.81.79
Fax : +33 (0)4.93.00.82.00
Email: Benoit.Huet@eurecom.fr

Date of birth: 7 February 1971 at Corbeil Essonnes
Nationality: French.
National Service completed.

1.1 Academic Qualifications

- Doctor of Philosophy (PhD): Computer Vision University of York, Computer Science Department, Computer Vision Group, 1999.
  Title: Object Recognition from Large Libraries of Line-Patterns.
  Keywords: Structural and Geometric Representation, Histograms, Relational Distance Measures and Attributed Relational Graph Matching.
  Supervisor: Prof. Edwin R. Hancock

  Three new ways of retrieving line-patterns using information concerning their geometry and structural arrangement have been devised. The first of these was based on a relational histogram, the second method used robust statistics to develop an extension of the Hausdorff distance for relational graphs, and, the final method was based on a fast graph-matching algorithm. Each of these methods was implemented and extensive experiments were devised to evaluate them on both real world and synthetic data.

- Master of Sciences: Knowledge Engineering (Artificial Intelligence) with distinction. University of Westminster (London) 1992-1993 (12 months)

  Modules: Knowledge based system design, Natural language understanding, Logic for knowledge representation, Languages for A.I. (Lisp / Prolog), Computer vision, Neural networks, Machine learning, Uncertain reasoning, Distributed artificial intelligence. Project: Recurrent Neural Networks for Temporal Sequence Recognition.

- Batchelor of Science: Enginering and Computing (first class honour) Ecole Supérieure de Technologie Electrique (Groupe E.S.I.E.E.) 1988-1992 (3 Years)

  (88-90) Electronic, Micro-electronic, Electrotechnic, Computer Science, Software Engineering. (91-92) Specialisation in Computer Science, Industrial project

- French Baccalaureate Lycee Geoffroy St Hillaire (ETAMPES 91) Serie F2 (Electronic) June 1988

1.2 Professional Experiences

- EURECOM, Multimedia Communications Department.
  Assistant Professor. 1999 – to date (since September 1999)

  Research and development of multimodal multimedia (still and moving images) indexing and retrieval techniques. Responsible for the following courses: Multimedia Technologies and Multimedia Advanced Topics. Assisting Prof. B. Merialdo for the following lecturers, tutorials and practicals: Intelligent Systems and MultiMedia Information Retrieval.

  PhD Advisor:

  - Mathilde Sahuguet, on the topic of "Web Multimedia Mining" since 2012.

- Xueliang Liu, on the topic of ”Semantic Multimodal Multimedia Mining” since 2009.
- Stephane Turlier, who received his PhD from Telecom ParisTech in 2011 for his thesis on ”Personalisation and Aggregation of Infotainment for Mobile Platforms”
- Marco Paleari, who received his PhD from Telecom ParisTech in 2009 for his thesis on ”Affective Computing; Display, Recognition and Artificial Intelligence”
- Rachid Benmokhtar, who received his PhD from Telecom ParisTech in 2009 for his thesis on ”Fusion multi-niveau pour l'indexation et la recherche multimédia par le contenu sémantique”
- Eric Galmar, who successfully defended his PhD in 2008 on ”Representation and Analysis of Video Content for Automatic Object Extraction”.

PhD Co-advisor:

- Ithery Yahiaoui who received her PhD from Telecom Paris in October 2003 for her thesis on ”Automated Video Summary Construction”
- Fabrice Souvannavong who received his PhD from Telecom Paris in June 2005 for his thesis on ”Semantic video content indexing and retrieval”
- Joakim Jiten who received his PhD from Telecom PARISTech in 2007 for his thesis on ”Multidimensional hidden Markov model applied to image and video analysis”

Current Funded Research Projects:

- MediaMixer (EU FP7): The MediaMixer CA will address the idea of re-purposing and re-using by showing the vision of a media fragment market (the MediaMixer) to the European media production, library, TV archive, news production, e-learning and UGC portal industries.
- EventMap (EIT ICT Labs): EventMap will demonstrate the use explicit representations of events to organize the provision and exchange of information and media. A web-based semantic multimedia agenda for events associated with the role of Helsinki as the 2012 Design Capital will serve as one of multiple demonstrators.

- LinkedTV (EU FP7): Television linked to the Web provides a novel practical approach to Future Networked Media. It is based on four phases: annotation, interlinking, search, and usage (including personalization, filtering, etc.). The result will make Networked Media more useful and valuable, and it will open completely new areas of application for Multimedia information on the Web.
- ALIAS (EU/ANR): Adaptable Ambient LIving ASsistant (ALIAS) is the product development of a mobile robot system that interacts with elderly users, monitors and provides cognitive assistance in daily life, and promotes social inclusion by creating connections to people and events in the wider world.

Past Research Projects:

- K-Space: K-Space integrates leading European research teams to create a Network of Excellence in semantic inference for semi-automatic annotation and retrieval of multimedia content. The aim is to narrow the gap between content descriptors that can be computed automatically by current machines and algorithms, and the richness and subjectivity of semantics in high-level human interpretations of audiovisual media: The Semantic Gap.
- RPM2: Résumé Plurimédia, Multi-documents et Multi-opinions. Multimedia Summarisation from multiple sources.
- PorTiVity: The porTiVity project will develop and experiment a complete end-to-end platform providing Rich Media Interactive TV services for portable and mobile devices, realising direct interactivity with moving objects on handheld receivers connected to DVB-H/DMB (broadcast channel) and UMTS (unicast channel).
- Fusion for Image Classification with Orange-FT Labs: Study of fusion algorithms for low and high level fusion in the context of high level feature extraction from soccer videos.
- 3W3S: "World Wide Web Safe Surfing Service" provides a filtering agent which is able to evaluate web pages based on several methods (based on URI, on keywords or on metadata) in order to ensure that only appropriate content is displayed by the Web browser.
- SPATION: "Services Platforms and Applications for Transparent Information management in an in-hOme Network". The

  objective of this project is to find innovative solutions for the movement, organization and retrieval of information in a heterogeneous home system (PC, PVR, TV, etc...).

  - GMF4iTV: "Generic Media Framework for Interactice Television". The aim of this project is to develop and demonstrate an end-to-end platform enabling interactive services on moving objects and TV programs according to the Multimedia Home Platform standard.
  - European Patent Office: Feasability study concerning the retrieval of patent's technical drawings according to content similarity. The query "images" could either be another technical drawings (complete of subpart) or a man made sketch. The results of this preliminary study have been presented in ICIP 2001.

- National University of Singapore
  School of Computer Science Visiting Research Fellow in Prof. Tat-Seng Chua's Lab for Media Search. 2008 (4 months)
  Advising local PhD students and PostDocs with work on Multimedia Question Answering, Emotion/Affect Recognition and Large-Scale Multimedia Corpus (which led to the creation of NUS-Wide [Chua et al., 2009])

- University of York, Computer Science Department. (UK)
  Research Associate in the Computer Vision Group. 1998 – 1999 (12 Months)
  Research and development of techniques for matching multi-sensorial aerial images for DERA (Defense Evaluation and Research Agency). The technique combines histogram based segmentation and template based region correspondence matching. Learning structural description for three-dimensional object representation and recognition (EPSRC funded). The aim of this research is to provide a method for automatically learn and produce compact structural object representation from multiple object views.

- University of York, Computer Science Department. (UK)
  Tutorial Assistant. 1995 - 1998 (3 Years, max 12H/week)
  The responsibilities include preparation, marking and assessment. 1st Year: Computer Architecture and Mathematics for Computer Science. 2nd Year: Computer Systems Architecture, Formal Language and Automata, and Mathematics for Computer Science. 3rd Year: Talks for Image Analysis.
  (Teaching Experience)

- University of Westminster (UK)
  Research Assistant, Artificial Intelligence Division. 1994 - 1995 (11 Months)
  University Information System for the Modular Scheme (UIS-MS) project. The aim of the project is to generate an inter-linked, hypertext-based system that will enable all Modular Scheme related information to be made available to students and staff, allowing fast access to complete and up-to-date University of Westminster Information. Development in C and C++ of the hypertext viewer, compiler and automatic database updating tools (via e-mail and templates) for Unix platforms.
  (In depth use of Unix tools, cron job, e-mail filter...)

- University of Westminster (UK)
  Research Assistant, Artificial Intelligence Division. 1994 (6 Months)
  Development in C++ and X Window of an image manipulation software with graphical user interface. Among the implemented algorithms are various edge detection techniques, resizing, rotating, dithering, blurring, colour manipulation, thinning, and other image processing algorithms.
  (Low-level C++ programming, image formats and use of Xlib, Athena, Motif, Openlook)

- University of Westminster (UK)
  Part-Time Lecturer, Artificial Intelligence Division. 1994 (6 Months)
  Neural Networks for the MSc Knowledge Engineering.
  (Teaching Experience)

- University of Westminster
  Master of Science summer project. 1993 (3 Months)
  Implementation of recurrent neural networks for temporal sequence recognition. Comparison of the behaviour, efficiency and recall performances of the neural architectures on learning the contingencies implied by a finite state automaton. *(Reading, understanding and re-create experiments of research papers)*

- University of Westminster (Polytechnic of Central London)
  Research Visitor. 1992 (4 Months)
  Development in C of a real-time, graphic, multitasking software to control and analyse the behaviour of a ball and beam apparatus. The software is currently used for student tutorial at the University of

Westminster. *(Discovery of the English educational and professional environment)*

- E.T.S. Electronic, (Les Ulis 91 France)
  Third year project. 1991-1992 (7 Months)
  Software development in C++ language of a graphic interface which allows an algorithm conception. This algorithm allows industrial process control.
  (Observation of the importance of project planning, in depth study of both C and C++ Languages)

- IBM, (Corbeil Essonnes 91 France)
  Second year project. 1990 (4 Months)
  Development of an Expert System to help on Site Security Personnel to diagnose their hardware/software security system problems. Knowledge base developed on E.S.E.(Expert System Environment).
  (Importance of listening to people for a good communication, research and knowledge extraction opportunity)

- IBM, (Corbeil Essonnes 91 France)
  Summer temporary employment. 1989 (2 Months)
  Exposure and test of silicium wafers on PERKIN-ELMER Engine in ABL-MASTERSLICE.
  (Discovery of the industrial world and team work)

1.3 Additional Information

- **Member of the following societies:**
  - ACM SIGMM (since 2002),
  - IEEE Computer Society (since 1998),
  - International Society of Information Fusion (ISIF) (2006-2009).
- **Editorial Boards:**
  - Multimedia Tools and Application (Springer),
  - Multimedia Systems (Springer),
  - Guest Editor for EURASIP Journal on Image and Video Processing: selected papers from MultiMedia Modeling 2009,
  - Guest Editor for IEEE Multimedia special issue on Large Scale Multimedia Retrieval and Mining,

  - Guest Editor for IEEE Multimedia special issue on Large Scale Multimedia Data Collections,
  - Guest Editor for the Journal of Media Technology and Applications special issue on Multimedia Content Analysis,
  - Guest Editor for Multimedia Systems special issue on Social Media Mining and Knowledge Discovery.

- **Reviewer for the following international journals:**
  - ACM Multimedia Systems Journal (Springer),
  - ACM Transactions on Multimedia Computing, Communications and Applications,
  - IEEE Pattern Analysis and Machine Intelligence,
  - IEEE Multimedia Magazine,
  - IEEE Transaction on Multimedia,
  - IEEE Image Processing,
  - IEEE Transactions on Circuits and Systems for Video Technology,
  - IEEE Signal Processing,
  - Multimedia Tools and Applications (Springer),
  - IEE Vision, Image and Signal Processing,
  - EURASIP Journal on Image and Video Processing,
  - Image Communication (Eurasip/Elsevier Science),
  - International Journal on Computer Vision and Image Understanding,
  - Computer Graphics Forum (International Journal of the Eurographics Association).

- **Reviewer for the following international conferences:**
  - ACM Multimedia,
  - ACM International Conference on Multimedia Retrieval (ICMR),
  - ACM International Conference on Multimedia Image Retrieval (MIR),
  - ACM International Conference on Image and Video Retrieval (CIVR),

- IEEE International Conference on Computer Vision and Pattern Recognition (CVPR),
- International Conference on MultiMedia Modeling (MMM),
- IEEE International Conference on Multimedia and Expo (ICME),
- International Workshop on Content-Based Multimedia Indexing (CBMI),
- IEEE International Workshop on MultiMedia Signal Processing (MMSP)
- International Conference on Image Analysis and Processing (ICIAP),
- International Conference on Pattern Recognition (ICPR),
- International Conference on Image Analysis and Recognition (ICIAR),
- IS&T/SPIE Symposium Electronic Imaging Science and Technology Conference on Storage and Retrieval for Media Databases

- **Conference/Workshop Organisation/Committee:**
  - Multimedia Modeling 2013: Organizing Co-chair.
  - ACM Multimedia 2012: Area Chair (Content Processing Track)
  - Workshop on Web-scale Vision and Social Media in conjunction with ECCV 2012: Program Committee
  - LSVSM'12 CVPR Workshop 2012: Large-Scale Video Search and Mining: Program Committee.
  - ACM International Conference on Multimedia Information Retrieval 2012: Tutorial Chair
  - IEEE WIAMIS 2012: Program Committee Member
  - MUE 2012, The 6th International Conference on Multimedia and Ubiquitous Engineering: "Multimedia Modeling and Processing" Track chair.
  - ACM Multimedia 2011: Associate Program Committee Member
  - MediaEval 2011: Co-Organiser of the Social Event Detection (SED) Task.
  - ACM Multimedia 2010: Workshop Co-Chair for the 2sd Workshop on Very-Large-Scale Multimedia Corpus, Mining and Retrieval

- ACM Multimedia 2010: Associate Program Committee Member (Content Processing Track)
- ACM International Conference on Multimedia Information Retrieval 2010: Program Committee Member
- IEEE International Conference and Multimedia Expo 2010: Program Committee Member
- CVIDS'10 ICME Workshop 2010: Visual Content Identification and Search: Program Committee
- Multimedia Modeling 2010: Publicity Chair
- International Workshop on Content-Based Multimedia Indexing,CMBI'09: Technical Program Committee
- ACM International Conference on Image and Video Retrieval 2009: Program Committee Member
- First International Workshop on Content-Based Audio/Video Analysis for Novel TV Services: Program Committee Member
- ACM Multimedia 2009: Doctoral Symposium Chair,
- ACM Multimedia 2009: Workshop Co-Chair for the 1st Workshop on Web-Scale Multimedia Corpus,
- International Conference on MultiMedia Modeling 2009 (MMM'09): **General Chair**
- ACM International Conference on Image and Video Retrieval 2009: Program Committee Member
- International Workshop on Content-Based Multimedia Indexing,CMBI'09: Technical Program Committee
- IEEE ICETE-SIGMAP 2008: Program Committee Member
- International Workshop on Content-Based Multimedia Indexing,CMBI'08: Technical Program Committee
- ACM Multimedia 2007: Tutorial Chair
- ACM Multimedia 06: Associate Program Committee Member (Content Processing Track)
- IEEE MultiMedia Modeling Conference 2006: Program Committee Member
- ACM Multimedia 05: Associate Program Committee Member (Content Processing Track)
- Coresa'05: Program Committee Member

- Fourth International Workshop on Content-Based Multimedia Indexing,CMBI'05: Technical Program Committee
- IEEE ICME'2005: Technical Program Committee
- ACM Multimedia 04: Associate Program Committee Member (Content Processing Track)
- Third International Workshop on Content-Based Multimedia Indexing, CBMI'03: Program Committee Member
- ACM Multimedia 2002: **Local Arrangements Chair and Treasurer**
- EMMCVPR 1999 (Second International Workshop on Energy Minimisation Methods in Computer vision and Pattern Recognition: Local Arrangements).

- **Project Evaluation/Expertise**
  - International Reviewer for the Singapore Ministry of Research.
  - International Reviewer for COST-Action Project Proposal (Switzerland)
  - European Commission, Information Society and Media (FP6 and FP7): Independent Expert and Reviewer.
  - RIAM: French national network on Research and Innovation in Audiovisual and Multimedia.
  - OSEO - CNC - Direction de l'innovation, de la vidéo et des industries techniques: Expert for the French Innovation Directorate for video and industrial techniques.
- **Military Service (1990-1991)** Regiment de Marche du TCHAD. Analyst Programmer in the Computing Science Department.

1.4 Publications

- **Books and Book Chapters**
  1. Raphaël Troncy, Benoit Huet, Simon Schenk, "Multimedia semantics: metadata, analysis and interaction" Wiley-Blackwell, July 2011, ISBN: 978-0470747001, pp 1-328
  2. Rachid Benmokhtar, Benoit Huet, Gaël Richard, Slim Essid, "Feature extraction for multimedia analysis", Book Chapter no. 4 in "Multimedia Semantics: Metadata, Analysis and Interaction", Wiley, July 2011, ISBN: 978-0-470-74700-1 , pp 35-58

3. Slim Essid, Marine Campedel, Gaël Richard, Tomas Piatrik, Rachid Benmokhtar, Benoit Huet, "Machine learning techniques for multimedia analysis" Book Chapter no. 5 in "Multimedia Semantics: Metadata, Analysis and Interaction", Wiley, July 2011, ISBN: 978-0-470-74700-1 , pp 59-80
4. Benoit Huet, Alan F. Smeaton, Ketan Mayer-Patel , Yannis Avrithis; Advances in Multimedia Modeling Springer : Lecture Notes in Computer Science, Subseries: Information Systems and Applications, incl. Internet/Web, and HCI , Vol. 5371, ISBN: 978-3-540-92891-1
5. Benoit Huet and Bernard Mérialdo, "Automatic video summarization", Chapter in "Interactive Video, Algorithms and Technologies" by Hammoud, Riad (Ed.), 2006, XVI, 250 p, ISBN: 3-540-33214-6 , pp 27-41.

- **Journals**

1. Benoit Huet, Tat-Seng Chua and Alexander Hauptmann, "Large-Scale Multimedia Data Collections", to appear in IEEE Multimedia, 2012.
2. Rachid Benmokhtar and Benoit Huet, "An ontology-based evidential framework for video indexing using high-level multimodal fusion", Multimedia Tools Application, Springer, December 2011 , pp 1-27
3. Rong Yan, Benoit Huet, Rahul Sukthankar, "Large-scale multimedia retrieval and mining", IEEE Multimedia, Vol 18, No. 1, January-March 2011
4. Benoit Huet, Alan F. Smeaton, Ketan Mayer-Patel, Yannis Avrithis, "Selected papers from multimedia modeling conference 2009", EURASIP Journal on Image and Video Processing Volume 2010, Article ID 792567
5. Fabrice Souvannavong, Lukas Hohl, Bernard Merialdo and Benoit Huet, "Structurally Enhanced Latent Semantic Analysis for Video Object Retrieval ", Special Issue of the IEE Proceedings on Vision, Image and Signal Processing , Volume 152, No. 6, 9 December 2005 , pp 859-867.
6. Fabrice Souvannavong, Bernard Merialdo and Benoit Huet, "Partition sampling: an active learning selection strategy for large database annotation", Special Issue of the IEE Proceedings on

Vision, Image and Signal Processing ,Volume 152 No. 3, May 2005, Special section on Technologies for interactive multimedia services , pp 347-355.

7. Ithery Yahiaoui, Bernard Merialdo and Benoit Huet, "Comparison of multi-episode video summarisation algorithms", EURASIP Journal on Applied Signal Processing, Special issue on Multimedia Signal Processing, Vol. 2003, No. 1, page 48-55, January 2003.
8. Huet B. and E. R. Hancock, "Relational Object Recognition from Large Structural Libraries", Pattern Recognition, Vol. 35, No. 9, page 1895-1915, Sept 2002.
9. Huet B. and E. R. Hancock, "Line Pattern Retrieval Using Relational Histograms", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 21, No. 12, page 1363-1370, December 1999.
10. Huet B., A.D.J. Cross and E.R. Hancock, "Shape Recognition from Large Image Libraries by Inexact Graph Matching", Pattern Recognition in Practice VI, June 2-4 1999, Vlieland, The Netherlands. Appeared in a special issue of Pattern Recognition Letters, 20, page 1259-1269, December 1999.
11. Huet B. and E.R. Hancock, "Object Recognition from Large Structural Libraries", Advances in Pattern Recognition: Lecture Notes in Computer Science (SSPR98), Springer-Verlag, 1451, August 1998.

- **International Conferences and Workshops**

  1. Xueliang Liu and Benoit Huet, "Social Event Visual Modeling from Web Media Data", ACM Multimedia'12 Workshop on Socially-Aware Multimedia, Nara, Japan, 2012.
  2. Xueliang Liu and Benoit Huet, "Social Event Discovery by Topic Inference", WIAMIS 2012, 13th International Workshop on Image Analysis for Multimedia Interactive Services, 23-25 May 2012, Dublin City University, Ireland , Dublin, Ireland.
  3. Xueliang Liu, Raphaël Troncy and Benoit Huet, "Using social media to identify events" WSM'11, ACM Multimedia 3rd Workshop on Social Media, November 18-December 1st, 2011, Scottsdale, Arizona, USA

4. Symeon Papadopoulos, Raphaël Troncy, Vasileios Mezaris, Benoit Huet, Ioannis Kompatsiaris, "Social event detection at MediaEval 2011: Challenges, dataset and evaluation", MediaEval 2011, MediaEval Benchmarking Initiative for Multimedia Evaluation, September 1-2, 2011, Pisa, Italy
5. Xueliang Liu, Raphaël Troncy and Benoit Huet, " EURECOM @ MediaEval 2011 social event detection task" MediaEval 2011, MediaEval Benchmarking Initiative for Multimedia Evaluation, September 1-2, 2011, Pisa, Italy
6. Xueliang Liu, Raphaël Troncy and Benoit Huet, "Finding media illustrating events", ICMR'11, 1st ACM International Conference on Multimedia Retrieval, April 17-20, 2011, Trento, Italy
7. Marco Paleari, Ryad Chellali and Benoit Huet, "Bimodal emotion recognition", ICSR'10, International Conference on Social Robotics, November 23-24, 2010, Singapore - Also published as LNCS Volume 6414/2010 , pp 305-314
8. Xueliang Liu and Benoit Huet, "Concept detector refinement using social videos", VLS-MCMR'10, International workshop on Very-large-scale multimedia corpus, mining and retrieval, October 29, 2010, Firenze, Italy , pp 19-24
9. Benoit Huet, Tat-Seng Chua and Alexander Hauptmann, "ACM international workshop on very-large-scale multimedia corpus, mining and retrieval", ACMMM'10, ACM Multimedia 2010, October 25-29, 2010, Firenze, Italy , pp 1769-1770
10. Xueliang Liu, Benoit Huet, "Automatic concept detector refinement for large-scale video semantic annotation", ICSC'10, IEEE 4th International Conference on Semantic Computing, September 22-24, 2010, Pittsburgh, PA, USA , pp 97-100
11. Marco Paleari, Benoit Huet, Ryad Chellali, "Towards multimodal emotion recognition : A new approach", CIVR 2010, ACM International Conference on Image and Video Retrieval, July 5-7, Xi'an, China , pp 174-181
12. Marco Paleari, Ryad Chellali, Benoit Huet, "Features for multimodal emotion recognition : An extensive study", CIS'10, IEEE International Conference on Cybernetics and Intelligent Systems, June 28-30, 2010, Singapore , pp 90-95
13. Marco Paleari, Vivek Singh, Benoit Huet, Ramesh Jain, "Toward environment-to-environment (E2E) affective sensitive communi-

cation systems", MTDL'09, Proceedings of the 1st ACM International Workshop on Multimedia Technologies for Distance Learning at ACM Multimedia, October 23rd, 2009, Beijing, China , pp 19-26

14. Benoit Huet, Jinhui Tang, Alex Hauptmann, ACM SIGMM the first workshop on web-scale multimedia corpus MM'09 : Proceedings of the seventeen ACM international conference on Multimedia, October 19-24, 2009, Beijing, China , pp 1163-1164
15. Marco Paleari, Carmelo Velardo, Benoit Huet, Jean-Luc Dugelay, "Face dynamics for biometric people recognition" MMSP'09, IEEE International Workshop on Multimedia Signal Processing, October 5-7, 2009, Rio de Janeiro, Brazil
16. Rachid Benmokhtar and Benoit Huet, "Hierarchical ontology-based robust video shots indexing using global MPEG-7 visual descriptors", CBMI 2009, 7th International Workshop on Content-Based Multimedia Indexing, June 3-5, 2009, Chania, Crete Island, Greece
17. Rachid Benmokhtar and Benoit Huet, "Ontological reranking approach for hybrid concept similarity-based video shots indexing", WIAMIS 2009, 10th International Workshop on Image Analysis for Multimedia Interactive Services, May 6-8, 2009, London, UK
18. Marco Paleari, Rachid Benmokhtar and Benoit Huet, "Evidence theory based multimodal emotion recognition", MMM 2009, 15th International MultiMedia Modeling Conference, January 7-9, 2009, Sophia Antipolis, France , pp 435-446
19. Thanos Athanasiadis, Nikolaos Simou, Georgios Th. Papadopoulos, Rachid Benmokhtar, Krishna Chandramouli, Vassilis Tzouvaras, Vasileios Mezaris, Marios Phiniketos, Yannis Avrithis, Yiannis Kompatsiaris, Benoit Huet, Ebroul Izquierdo, "Integrating image segmentation and classification for fuzzy knowledge-based multimedia indexing" MMM 2009, 15th International MultiMedia Modeling Conference, January 7-9, 2009, Sophia Antipolis, France
20. Rachid Benmokhtar, Eric Galmar and Benoit Huet, "K-Space at TRECVid 2008" TRECVid'08, 12th International Workshop on Video Retrieval Evaluation, November 17-18, 2008, Gaithersburg, USA

21. Rachid Benmokhtar and Benoit Huet, "Perplexity-based evidential neural network classifier fusion using MPEG-7 low-level visual features", MIR 2008, ACM International Conference on Multimedia Information Retrieval 2008, October 27- November 01, 2008, Vancouver, BC, Canada , pp 336-341
22. L. Goldmann, T. Adamek, P. Vajda, M. Karaman, R. Mörzinger, E. Galmar, T. Sikora, N. O'Connor, T. Ha-Minh, T. Ebrahimi, P. Schallauer, B. Huet, "Towards fully automatic image segmentation evaluation" ACIVS 2008, Advanced Concepts for Intelligent Vision Systems, October 20-24, 2008, Juan-les-Pins, France
23. Eric Galmar and Benoit Huet, "Spatiotemporal modeling and matching of video shots", 1st ICIP Workshop on Multimedia Information Retrieval : New Trends and Challenges, October 12-15, 2008, San Diego, California, USA , pp 5-8
24. Marco Paleari, Benoit Huet, Antony Schutz and Dirk T. M. A. Slock, "A multimodal approach to music transcription", 1st ICIP Workshop on Multimedia Information Retrieval : New Trends and Challenges, October 12-15, 2008, San Diego, USA , pp 93-96
25. Eric Galmar, Thanos Athanasiadis, Benoit Huet, Yannis Avrithis, "Spatiotemporal semantic video segmentation" MMSP 2008, 10th IEEE International Workshop on MultiMedia Signal Processing, October 8-10, 2008, Cairns, Queensland, Australia , pp 574-579
26. Stéphane Turlier, Benoit Huet, Thomas Helbig, Hans-Jörg Vögel, "Aggregation and personalization of infotainment, an architecture illustrated with a collaborative scenario" 8th International Conference on Knowledge Management and Knowledge Technologies, September 4th, 2008, Graz, Austria
27. Marco Paleari, Benoit Huet, Antony Schutz and Dirk T. M. A. Slock, "Audio-visual guitar transcription", Jamboree 2008 : Workshop By and For KSpace PhD Students, July, 25 2008, Paris, France
28. Rachid Benmokhtar, Benoit Huet and Sid-Ahmed Berrani, "Low-level feature fusion models for soccer scene classification", 2008 IEEE International Conference on Multimedia & Expo, June 23-26, 2008, Hannover, Germany
29. Marco Paleari, Benoit Huet, "Toward emotion indexing of multimedia excerpts" CBMI 2008, 6th International Workshop on

Content Based Multimedia Indexing, June, 18-20th 2008, London, UK [**Best student paper award**]

30. Marco Paleari, Benoit Huet, Brian Duffy, "SAMMI, Semantic affect-enhanced multimedia indexing", SAMT 2007, 2nd International Conference on Semantic and Digital Media Technologies, 5-7 December 2007, Genoa, Italy

31. Rachid Benmokhtar, Eric Galmar and Benoit Huet, "Eurecom at TRECVid 2007: Extraction of high level features", TRECVid'07, 11th International Workshop on Video Retrieval Evaluation, November 2007, Gaithersburg, USA

32. Rachid Benmokhtar, Eric Galmar and Benoit Huet, ,"K-Space at TRECVid 2007", TRECVid'07, 11th International Workshop on Video Retrieval Evaluation, November 2007, Gaithersburg, USA

33. Marco Paleari, Brian Duffy and Benoit Huet, "ALICIA, an architecture for intelligent affective agents", IVA 2007 7th International Conference on Intelligent Virtual Agents, 17th - 19th September 2007 Paris, France — Also published in LNAI Volume 4722 , pp 397-398

34. Marco Paleari, Brian Duffy and Benoit Huet, "Using emotions to tag media", Jamboree 2007: Workshop By and For KSpace PhD Students, September, 15th 2007, Berlin, Germany

35. Eric Galmar and Benoit Huet, "Analysis of vector space model and spatiotemporal segmentation for video indexing and retrieval", CIVR 2007, ACM International Conférence on Image and Video Retrieved, July 9-11 2007, Amsterdam, The Netherlands

36. Rachid Benmokhtar, Benoit Huet, Sid-Ahmed Berrani, Patrick Lechat, "Video shots key-frames indexing and retrieval through pattern analysis and fusion techniques", FUSION'07, 10th International Conference on Information Fusion, July 9-12 2007, Quebec, Canada

37. Rachid Benmokhtar and Benoit Huet, "Multi-level fusion for semantic indexing video content", AMR'07, International Workshop on Adaptive Multimedia Retrieval, June 5-6 2007, Paris, France

38. Rachid Benmokhtar and Benoit Huet, "Performance analysis of multiple classifier fusion for semantic video content indexing and

retrieval", MMM'07, International MultiMedia Modeling Conference, January 9-12 2007, Singapore - Also published as LNCS Volume 4351 , pp 517-526

39. Rachid Benmokhtar and Benoit Huet, "Neural network combining classifier based on Dempster-Shafer theory for semantic indexing in video content", MMM'07, International MultiMedia Modeling Conference, January 9-12 2007, Singapore - Also published as LNCS Volume 4351 , pp 196-205
40. Rachid Benmokhtar, Emilie Dumont, Bernard Mérialdo and Benoit Huet, "Eurecom in TrecVid 2006: high level features extractions and rushes study", TrecVid 2006, 10th International Workshop on Video Retrieval Evaluation, November 2006, Gaithersburg, USA
41. Peter Wilkins, Tomasz Adamek, Paul Ferguson, Mark Hughes, Gareth J F Jones, Gordon Keenan, Kevin McGuinness, Jovanka Malobabic, Noel E. O'Connor, David Sadlier, Alan F. Smeaton, Rachid Benmokhtar, Emilie Dumont, Benoit Huet, Bernard Mérialdo, Evaggelos Spyrou, George Koumoulos, Yannis Avrithis, R. Moerzinger, P. Schallauer, W. Bailer, Qianni Zhang, Tomas Piatrik, Krishna Chandramouli, Ebroul Izquierdo, Lutz Goldmann, Martin Haller, Thomas Sikora, Pavel Praks, Jana Urban, Xavier Hilaire and Joemon M. Jose, "K-Space at TRECVid 2006", TrecVid 2006, 10th International Workshop on Video Retrieval Evaluation, November 2006, Gaithersburg, USA
42. Isao Echizen, Stephan Singh, Takaaki Yamada, Koichi Tanimoto, Satoru Tezuka and Benoit Huet, "Integrity verification system for video content by using digital watermarking", ICSSSM'06, IEEE International Conference on Services Systems and Services Management, 25-27 October 2006, Troyes, France
43. Eric Galmar and Benoit Huet, "Graph-based spatio-temporal region extraction", ICIAR 2006, 3rd International Conference on Image Analysis and Recognition, September 18-20, 2006, Póvoa de Varzim, Portugal — Also published as Lecture Notes in Computer Science (LNCS) Volume 4141 , pp 236–247
44. Rachid Benmokhtar and Benoit Huet, "Classifier fusion : combination methods for semantic indexing in video content", ICANN 2006, International Conference on Artificial Neural Networks, 10-14 September 2006, Athens, Greece - also published as LNCS Volume 4132 , pp 65-74

45. Bernard Mérialdo, Joakim Jiten, Eric Galmar and Benoit Huet, "A new approach to probabilistic image modeling with multidimensional hidden Markov models", AMR 2006, 4th International Workshop on Adaptive Multimedia Retrieval , 27-28 July 2006, Geneva, Switzerland —Also published as LNCS Volume 4398

46. Fabrice Souvannavong and Benoit Huet, "Continuous behaviour knowledge space for semantic indexing of video content", Fusion 2006, 9th International Conference on Information Fusion, 10-13 July 2006, Florence Italy

47. Benoit Huet and Bernard Mérialdo, "Automatic video summarization", Chapter in "Interactive Video, Algorithms and Technologies" by Hammoud, Riad (Ed.), 2006, XVI, 250 p, ISBN: 3-540-33214-6 , pp 27-41

48. Joakim Jiten, Bernard Mérialdo and Benoit Huet, "Multi-dimensional dependency-tree hidden Markov models", ICASSP 2006, 31st IEEE International Conference on Acoustics, Speech, and Signal Processing, May 14-19, 2006, Toulouse, France

49. Joakim Jiten, Benoit Huet and Bernard Mérialdo, "Semantic feature extraction with multidimensional hidden Markov model", SPIE Conference on Multimedia Content Analysis, Management and Retrieval 2006, January 17-19, 2006 - San Jose, USA - SPIE proceedings Volume 6073 Volume 6073 , pp 211-221

50. Joakim Jiten, Fabrice Souvannavong, Bernard Mérialdo and Benoit Huet, "Eurecom at TRECVid 2005: extraction of high-level features", TRECVid 2005, TREC Video Retrieval Evaluation, November 14, 2005, USA

51. Benoit Huet, Joakim Jiten, Bernard Merialdo, "Personalization of hyperlinked video in interactive television", IEEE International Conference on Multimedia & Expo July 6-8, 2005, Amsterdam, The Netherlands.

52. B. Cardoso, F. de Carvalho, L. Carvalho, G. Fernàndez, P. Gouveia, B. Huet, J. Jiten, A.López, B. Merialdo, A. Navarro, H. Neuschmied, M. Noé, R. Salgado, G. Thallinger, "Hyperlinked video with moving object in digital television", IEEE International Conference on Multimedia & Expo, July 6-8, 2005, Amsterdam, The Netherlands.

53. F. Souvannavong, B. Merialdo and B. Huet, "Region-based video content indexing and retrieval", Fourth International Workshop

on Content-Based Multimedia Indexing (CBMI'05), June 21-23, 2005 Riga, Latvia.

54. F. Souvannavong, B. Merialdo and B. Huet, "Multi-modal classifier fusion for video shot content", 6th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'05), Montreux, Switzerland, April 2005.
55. Fabrice Souvannavong, L. Hohl, B. Merialdo and B. Huet, "Enhancing Latent Semantic Analysis Video Object Retrieval with Structural Information", IEEE International Conference on Image Processing, October 24-27, 2004 Singapore.
56. Fabrice Souvannavong, B. Merialdo and B. Huet, "Latent Semantic Analysis For An Effective Region Based Video Shot Retrieval System", 6th ACM SIGMM International Workshop on Multimedia Information Retrieval, held in conjunction with ACM Multimedia 2004, October 15-16, 2004, New York, NY USA.
57. Fabrice Souvannavong, B. Merialdo and B. Huet, "Eurecom at Video-TREC 2004: Feature Extraction Task ", NIST Special Publication, The 13th Text Retrieval Conference (TREC 2004 Video Track).
58. Bernardo Cardoso and Fausto de Carvalho and Gabriel Fernandez and Benoit Huet and Joakim Jiten and Alejandro Lopez and Bernard Merialdo and Helmut Neuschmied and Miquel Noe and David Serras Pereira and Georg Thallinger. "Personalization of Interactive Objects in the GMF4iTV project ". Proceedings of TV'04: the 4th Workshop on Personalization in Future TV held in conjunction with Adaptive Hypermedia 2004 ,Eindhoven, The Netherlands, August 23, 2004.
59. Fabrice Souvannavong, L. Hohl, B. Merialdo and B. Huet, "Using Structure for Video Object Retrieval", International Conference on Image and Video Retrieval, July 21-23, 2004, Dublin City University, Ireland .
60. Fabrice Souvannavong, B. Merialdo and B. Huet, "Improved Video Content Indexing By Multiple Latent Semantic Analysis", International Conference on Image and Video Retrieval, July 21-23, 2004, Dublin City University, Ireland .
61. Fabrice Souvannavong, B. Merialdo and B. Huet, "Latent Semantic Indexing For Semantic Content Detection Of Video Shots", IEEE International Conference on Multimedia and Expo (ICME'2004), June 27th – 30th, 2004, Taipei, Taiwan.

62. Fabrice Souvannavong, B. Merialdo and B. Huet, ”Partition Sampling for Active Video Database Annotation”, 5th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS’04), April 21-23, 2004, Instituto Superior Técnico, Lisboa, Portugal.
63. Fabrice Souvannavong, B. Merialdo and B. Huet, ”Latent Semantic Indexing for Video Content Modeling and Analysis”, NIST Special Publication, The 12th Text Retrieval Conference (TREC 2003 Video Track).
64. Fabrice Souvannavong, B. Merialdo and B. Huet, ”Video Content Structuration With Latent Semantic Analysis”, Third International Workshop on Content-Based Multimedia Indexing, CBMI 2003, 22-24 Septembre 2003, Rennes, France.
65. Fabrice Souvannavong, B. Merialdo and B. Huet, ”Semantic Feature Extraction using Mpeg Macro-block Classification”, NIST Special Publication: SP 500-251, The Eleventh Text Retrieval Conference (TREC 2002 Video Track).
66. Gerhard Mekenkamp, Mauro Barbieri, Benoit Huet, Itheri Yahiaoui, Bernard Merialdo, Riccardo Leonardi and Michael Rose, ”Generating TV Summaries for CE Devices”, ACM Multimedia 2002, December 3-5 2002, Juan Les Pins, France.
67. Benoit Huet, Itheri Yahiaoui, Bernard Merialdo, ”Image Similarity for Automatic Video Summarization”, EUSIPCO 2002 - 11th European Signal Processing Conference, September 3-6 2002, Toulouse, France.
68. Bernard Merialdo, B. Huet, I. Yahiaoui, Fabrice Souvannavong, ”Automatic Video Summarization”, International Thyrrenian Workshop on Digital Communications, Advanced Methods for Multimedia Signal Processing, September 8th - 11th, 2002, Palazzo dei Congressi, Capri, Italy.
69. Benoit Huet, G. Guarascio, N. Kern and B. Merialdo, ”Relational skeletons for retrieval in patent drawings”, IEEE International Conference Image Processing (ICIP2001), October 7-10 2001, Thessaloniki, Greece.
70. Ithery Yahiaoui, Bernard Merialdo et Benoit Huet, ”Automatic Summarization of Multi-episode Videos with the Simulated User Principle”, Workshop on MultiMedia Signal Processing (MMSP’01), October 3-5, 2001, Cannes, France.

71. Itheri Yahiaoui, Bernard Merialdo and Benoit Huet, "Optimal video summaries for simulated evaluation", European Workshop on Content-Based Multimedia Indexing, September 19-21, 2001 Brescia, Italy.
72. Itheri Yahiaoui, Bernard Merialdo and Benoit Huet, "AUTOMATIC VIDEO SUMMARIZATION", MMCBIR 2001 - Indexation et Recherche par le Contenu dans les Documents Multimedia, 24 et 25 septembre 2001, INRIA - Rocquencourt, France.
73. Ithery Yahiaoui, Bernard Merialdo et Benoit Huet, "Generating Summaries of Multi-Episodes Video", International Conference on Multimedia & Expo (ICME2001), August 22-25, 2001 Tokyo, Japan.
74. Itheri Yahiaoui, Bernard Merialdo and Benoit Huet, "Automatic construction of multi-video summaries", ISKO: Filtrage et résumé automatique de l'information sur les réseaux, July 5-6 2001, Nanterre, France.
75. Benoit Huet, Ithery Yahiaoui et Bernard Merialdo, "Multi-Episodes Video Summaries", International Conference on Media Futures 2001, 8-9 May 2001, Florence, Italy.
76. Arnd Kohrs, Benoit Huet, et Bernard Merialdo, "Multimedia Information Recommendation and Filtering on the Web", Networking 2000, May 14 - 19, 2000, Paris, France.
77. Merialdo B., S. Marchand-Maillet and B. Huet, "Approximate Viterbi decoding for 2D-Hidden Markov Models", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2000), Istanbul Turkey, June 5-9 2000.
78. Huet B. and E. R. Hancock, "Sensitivity Analysis for Object Recognition from Large Structural Libraries", IEEE International Conference on Computer Vision (ICCV99), Kerkyra, Greece, September 20-27, 1999.
79. Huet B. and E. R. Hancock, "Inexact Graph Retrieval", IEEE CVPR99 Workshop on Content-based Access of Image and Video Libraries (CBAIVL-99), Fort Collins, Colorado USA, June 22, 1999.
80. Huet B., A.D.J. Cross and E.R. Hancock, "Shape Retrieval by Inexact Graph Matching";, IEEE International Conference on Multimedia Computing and Systems (ICMCS'99), Florence, Italy, page 772-776, 7-11 June 1999.

81. Huet B. and E.R. Hancock, "Structural Sensitivity for Large-Scale Line-Pattern Recognition", Third International Conference on Visual Information Systems (VISUAL99), page 711-718, 2-4 June, 1999, Amsterdam, The Netherlands.
82. Huet B., A.D.J. Cross and E.R. Hancock, "Graph Matching for Shape Retrieval", Advances in Neural Information Processing Systems 11, Edited by M.J. Kearns, S.A. Solla and D.A. Cohn, MIT Press, June 1999.
83. Worthington P., B. Huet and E.R. Hancock, "Appearance-Based Object Recognition Using Shape-From-Shading", Proceeding of the 14th International Conference on Pattern Recognition (ICPR'98), Brisbane (Australia), page 412-416, 16-20 August 1998.
84. Huet B. and E.R. Hancock, "Relational Histograms for Shape Indexing", IEEE International Conference on Computer Vision (ICCV98), Mumbai India, page 563-569, Jan 1998.
85. Huet B. and E.R. Hancock, "Fuzzy Relational Distance for Large-scale Object Recognition", IEEE Conference on Computer Vision and Pattern Recognition (CVPR'98), Santa Barbara California USA, page 138-143, June 1998.
86. Huet B. and E.R. Hancock, "Pairwise Representation for Image Database Indexing", Sixth International Conference on Image Processing and its Applications (IPA97), Dublin (Ireland), 15-17 July 1997.
87. Huet B. and E.R. Hancock, "Cartographic Indexing into a Database of Remotely Sensed Images", Third IEEE Workshop on Applications of Computer Vision (WACV96), Sarasota Florida (USA), page 8-14, 2-4 Dec 1996.
88. Huet B. and E.R. Hancock, "Structural Indexing of infra-red images using Statistical Histogram Comparison", Third International Workshop on Image and Signal Processing (IWISP'96), Manchester (UK), 4-7 Nov 1996.
89. Charlton P. and Huet B., "Intelligent Agents for Image Retrieval", Research and Technology Advances in Digital Libraries, Virginia (USA), May 1995.
90. Charlton P. and Huet B., "Using Multiple Agents For Content-Based Image Retrieval", European Research Seminar on Advances in Distributed Systems, L'Alpe D'Huez (France), April 1995 .

- **National Conferences and Workshops**

  1. E. Galmar and B. Huet, "Méthode de segmentation par graphe pour le suivi de régions spatio-temporelles". CORESA 2005, 10èmes journées Compression et représentation des signaux audiovisuels, 7-8 Novembre 2005, Rennes, France.
  2. Fabrice Souvannavong, B. Merialdo and B. Huet, "Classification Sémantique des Macro-Blocs Mpeg dans le Domaine Compressé.", CORESA 2003,16 - 17 Janvier 2003, Lyon France.
  3. Itheri Yahiaoui, Bernard Merialdo, Benoit Huet, "User Evaluation of Multi-Episode Video Summaries", Indexation de documents et Recherche d'informations, GDR I3 et ISIS, July 9 2002, Grenoble, France.
  4. Itheri Yahiaoui, Bernard Merialdo, Benoit Huet, "Construction et Evaluation automatique de résumés multi-vidéos", Analyse et Indexation Multimédia, June 20 2002, Université Bordeaux 1, France.
  5. I. Yahiaoui, B. Mérialdo et B. Huet, "Construction automatique de résumés multi-vidéos", CORESA 2001, Nov 2001, Université de Dijon, France.
  6. I. Yahiaoui, B. Mérialdo et B. Huet, "Résumés automatiques de séquences vidéo", CORESA2000, 19-20 Octobre 2000, Université de Poitiers, Futuroscope, France.
  7. Worthington P., B. Huet and E.R. Hancock, "Increased Extend of Characteristic Views using Shape-from-Shading for Object Recognition", Proceeding of the British Machine Vision Conference (BMVC'98), Southampton (UK), page 710-719, 7-10 Sept 1998.
  8. Huet B. and E.R. Hancock, "Structurally Gated Pairwise Geometric Histograms for Shape Indexing", Proceeding of the British Machine Vision Conference (BMVC97), Colchester (UK), page 120-129, 8-11 Sept 1997.
  9. Huet B. and E.R. Hancock, "A Statistical Approach to Hierarchical Shape Indexing", Intelligent Image Databases (IEE and BMVA), London (UK), May 1996.

- **Technical Reports**

  1. Huet B., "Object Recognition from Large Libraries of Line-Patterns", PhD Thesis, University of York, Mai 1999.

2. Huet B., Parapadakis D., Konstantinou V. and Morse P., "The UIS-MS User Guide", University of Westminster AIRG/SCSISE Technical Report, 1994.
3. Huet B., "Recurrent Neural Networks for Temporal Sequences Recognition", MSc Thesis, University of Westminster, September 1993.
4. Huet B. and Houdry J.B., "MECABUS: Ensemble Logiciel et materiel d'aide a la conception et a la realisation d'automatismes industriels", Project Report, E.T.S. Electronic, April 1992.
5. Huet B., "Systeme d'aide a la decision pour CASI-RUSCO system 1800", Project Report, IBM (Corbeil Essonnes) and Ecole Superieure de Technologie Electrique, June 1990.
6. Huet B., "Systeme d'aide a la decision pour CASI-RUSCO system 1800, Guide du Developpeur", Technical Report, IBM (Corbeil Essonnes) , August 1990.

- **Invited Talks**
  - Université de Genève, Centre Universitaire d'Informatique, Computer Vision Group, July 1995. Presentation Title: "A Framework for Content-Based Retrieval of Images"
  - British Machine Vision Association Technical Meeting: Image and Video Databases, December 3rd 1997. Presentation Title: "Indexing in Line-Pattern Databases"
  - INRIA Sophia Antipolis, ARIANA project, March 27th 2000. Presentation Title: "Hierarchical Graph Based Techniques for Cartographic Content Based Indexing"
  - IMAG Grenoble (France), Working Group on Information Retrieval and Indexing (GDR I3 and GDR ISIS), July 9th 2002. Presentation Title: "User Evaluation of Multi-Episode Video Summaries".
  - Multimedia Workshop at Columbia University June 18th, 2004, Presentation title: "Relational LSA for Video Object Classification and Retrieval".
  - Working Group on Multimedia Information Retrieval and Indexing, Lyon - 08 juillet 2005, Presentation title: "Fusion de classifieur multimodaux pour la caracterisation de plan video"

  - SCATI - Journée évaluation des traitements dans un système de vision, 15 December 2005, Presentation title: "Evaluation d'algorithmes d'analyse vidéo: l'exemple TRECVid"
  - MultiMedia Workshop at Microsoft Research (Seattle), June 2006, Presentation Title: "Multimedia Reasearch @ Institut Eurecom"
  - Xerox Research Centre Europe, Thursday April 26 2007, Presentation title: "Multimodal Information Fusion for Semantic Concept Detection"
  - Sharp Laboratories (Tenri, Japan), January 2008, Presentation title: "Multimedia research for Interactive TV"
  - National University of Singapore (Singapore), July 2008, Presentation title: "Multimedia research at Eurecom"
  - Nayang Technological University (Singapore), September 2008, Presentation title: "Multimedia research at Eurecom"
  - UC Irvines (California, USA), Oct 2008, Presentation title: "Multimodal Emotion Recognition"
  - GdR-ISIS/IRIM: journee "Indexation scalable et cross-média", 26 novembre 2009, Paris. Presentation Title: "Réflexions de la communauté indexation multimédia sur le passage à l'échelle"
  - National University of Singapore (Singapore), July 2010, Presentation title: "Automatic Annotation of Online Social Videos"
  - VIGTA'12 Keynote Speaker (First International Workshop on Visual Interfaces for Ground Truth Collection in Computer Vision Applications) Capri, May 2012, Presentation title: "Multimedia Data Collection using Social Media Analysis".

- **Panels:**
  - CBMI'10: Large-Scale Multimedia Content Retrieval. (Panelist)
  - CIVR'10: Is there a future after CIVR'10? (Panelist)
  - VLSMCMR'10: (Panel Moderator) Very Large Scale Multimedia Corpus, Mining and Retrieval Panelists: Shin'Ichi Sato (NII, Japan), Apostol Natsev (IBM T. J. Watson Research Center, USA), Ed Chang (Google Inc., P.R. China), Remi Landais (Exalead, France).

- **Patent:**

- Joerg Deigmoeller, Gerhard Stoll, Helmut Neuschmied, Andreas Kriechbaum, José Bernardo Dos Santos Cardoso, Fausto José Oliveira de Carvalho, Roger Salgado de Alem, Benoit Huet, Bernard Mérialdo and Rémi Trichet, ”A method of adapting video images to small screen sizes”, European Patent No. WO2009115101 (A1) , pp 1-23.

2 Research Activities

There is a digital revolution happening right before our eyes, the way we communicate is rapidly changing dues to rapid technological advances. Pencil and paper communication is drastically reducing and being replaced with newer communication medium ranging from emails to sms/mms and other instant messaging services. Information/news used to be broadcasted only through official and dedicated channels such as television, radio or newspapers. The technology available today allows every single one of us to be individual information broadcasters whether through text, image or video using our personal connected mobile device. In effect, the current trend shows that video will soon become the most important media on the Internet. While the amount of multimedia content continuously increases there is still progress to be done for automatically understanding multimedia documents in order to provide means to index, search and browse them more effectively.

The objectives of this chapter are three-fold. First, we will motivate multimedia content modeling research in the current technological context. Secondly, a broad state of the art will provide the reader with a brief overview of the methodological trends of the field. Thirdly, a bird eye view of the various research themes I have supervised and/or conducted will be presented and will expose how contextual information has become an important additional source of information for multimedia content understanding.

2.1 Introduction/Motivation

During the last ten years, we have witnessed a digital data revolution. While the Internet and the world wide web have clearly enabled such an amazingly rapid growth, new electronic devices such as smart phones, tablet, etc.. have made it easier for people to capture, share and access multimedia information (text, images, audio, location, video...) continuously. However, searching and more specifically locating relevant multimedia information is becoming more of a challenge due to information/data overload. Just as an illustration, there were over 48 hours of video uploaded on YouTube [1] alone every minute in May 2011, and this keep growing at an impressive rate. Reuter [2] published in January 2012 that Google's video-share platform is now receiving approximately 60 hours of video per minute, a 25% increase over 8 months! Similarly impressive numbers are reported by photo/image online sharing platforms; 3 million new photos per day on Flickr [3] and a whooping 85 million photo uploaded every day on Facebook [4]. According to Cisco's Visual Networking Index [5]: Forecast and Methodology for 2010-2015; "It would take over 5 years to watch the amount of video that will cross global IP networks every second in 2015". Furthermore, "Internet video is now 40 percent of consumer Internet traffic, and will reach 62 percent by the end of 2015, not including the amount of video exchanged through P2P file sharing. The sum of all forms of video (TV, video on demand [VoD], Internet, and P2P) will continue to be approximately 90 percent of global consumer traffic by 2015". Given such figures the need to efficient and effective tools for finding online multimedia content is still very much on today's research agenda. It has clearly become impossible to manually annotate and check all online media content. Moreover, the sheer volume of data coupled with the number of users is creating new challenges for multimedia researchers in terms of algorithmic scalability, effectiveness and efficiency.

The scene is also changing rapidly in terms of the amount and the variety of contextual information available with the multimedia content uploaded on media sharing platforms. Media capturing devices (camcorder, camera, etc...) are getting increasingly ubiquitous and are rapidly converging towards a single highly connected portable device (i.e. the smartphone), every new device generation becoming more powerful and more mobile than the previous. The integration of numerous sensors, such as GPS, gyroscope, accelerometers, etc... on such devices provides important and unprecedented contextual information about the captured media. Indeed, nowadays when taking a photo or capturing a video using a mobile device, many extra information are automatically attached to the multimedia document. It is therefore possible to know where and when the photo was

[1] http://www.youtube.com

[2] http://www.reuters.com/article/2012/01/23/us-google-youtube-idUSTRE80M0TS20120123

[3] http://www.flickr.com

[4] http://www.facebook.com

[5] http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360_ns827_Networking_Solutions_White_Paper.html

taken (GPS/timer), under which conditions (focal length/aperture/flash), which direction the camera was pointing toward, etc... and the list grows with each new generation of devices. Images and videos content analysis approaches at large could benefit from such rich contextual information when processing the associated media.

Another aspect which is impacting multimedia research trends is the extraordinary success of social networks, which is contributing to massive growth of multimedia information exchange on the Internet. User of such services are happily and freely providing additional metadata to image and video through comments, tags, categories assignments, etc... Again, this extra information can prove to be particularly helpful for many multimedia processing tasks; such as media annotation, indexing and retrieval. However, such user generated metadata (comments in particular) should not be trusted. In many cases the comments are only relevant to the owner of the media or directed to his/her friends or relatives. Take the example of a photo of the Eiffel Tower which the owner annotated ”my wonderful summer holiday”, such description does not bring much meaningful information for the task of recognising the Eiffel tower in the photograph. A discussion with friends and relatives about how much he/she enjoyed his/her summer holiday may continue to enrich the comments associated to the image bringing little relevant information for content understanding/modeling. In other words, while some user provided metadata can contribute to better define the content of media documents others only bring additional noise. Therefore, it is mandatory to find way to curate user contributed metadata before relying on it for further processing.

The research field of multimedia content annotation, indexing and retrieval has been devoting much of its efforts to solving the well known Semantic Gap problem [Smeulders et al., 2000]. It refers to the difficulty for computer algorithms to detect high level semantic concepts (such as vehicle, animal, happiness, etc..) from the low level descriptors extracted automatically from the multimedia data. The wealth of contextual information surrounding media (i.e. photos and videos) nowadays enables researchers to propose novel and more effective algorithms for content analysis, thus reducing the Semantic Gap a little further.

Having introduced the current scene surrounding multimedia research, we now present the state of the art in the domain of multimedia content analysis. Then, in the following sections, we give an overview of a number of research directions which we have explored in the last 5 years or so. Finally, we provide a vision for the research themes we foresee as the most interesting and which we plan to study.

2.2 State of the Art

With almost 20 years of research since Multimedia Retrieval started to emerge as a scientific field [Niblack, 1993, Jain, 1993], a vast number of approaches have been proposed and studied leading to rather mature solutions [Hauptmann et al., 2008, Lee et al., 2006, Natsev et al., 2008, Snoek, 2010, Luan et al., 2011]. While at first multimedia retrieval was a simple extension of the databases search, offering retrieval on those images/videos featuring user provided keywords, the need for content based approaches never stopped to increase due both the extensive cost and subjectivity of media labeling by humans and the overwhelming rate at which new media are uploaded on the Internet. The first real multimedia retrieval systems required users to formulate their query based on low level properties (usually dominant colour) [Flickner et al., 1995, Smith and Chang, 1996, Pentland et al., 1996], on a sketch (hand drawn) [Eakins, 1989] or an image [Swain and Ballard, 1991]. Neither of those approaches really made it as the long awaited "killer app" due to their lack of practicality and the limited semantic coherence of the results with the query. When looking for an image or a video, is it not convenient to provide an initial query image to the search engine (unless searching for duplicates or near duplicates [Zhao et al., 2007, Naturel and Gros, 2008, Poullot and Satoh, 2010]). Hence, a novel paradigm entered the scene; automatic image annotation where image labels are produced by analysing the content (i.e. low level descriptors) of the media in order to be used by standard search engines (i.e. text based). Traditionally, the computational process involved in the labeling of an image or a video can be broadly decomposed in two parts. First, descriptors or low level features need to be extracted from the media. Then, a model is learned for each label, based on known occurrences of the label in the media. We shall now give a bird eye view of both the features and the models frequently employed in the field.

2.2.1 Low Level Features

Low level features extracted from the pixel map are the lowest form of representation in the visual content analysis chain. Such features can be computed globally over the entire image or at the region level. Whether to choose one over the other depends greatly on type of feature computed and the target application. In recent years, there has been a bias in favor for local/region descriptors in spite of the larger resulting descriptor size. Regions are obtained by either placing a grid over the image [Vailaya et al., 2001, Lim et al., 2003] or through a data-driven segmentation process. While the grid is the simplest form of image segmentation is it also the least effective for capturing the semantic of the image. The segmentation of images into homogeneous regions is an important research area of computer vision. Many approaches have been proposed based on clustering [Wang et al., 2001, Mezaris et al., 2003], region growing [Deng and Manjunath, 2001, Pratt and Jr., 2007], contour detection [Ciampini et al., 1998, Velasco and Marroquín, 2003], statistical model [Carson et al., 2002] or graph partitioning [Shi and Malik, 2000]. All have their advantages and drawbacks. Notably approaches based on clustering (i.e. K-means) require the number of desired regions to be known in advances. The accurate identification of the seeds is the main weakness for algorithms based on region growing. Contour based approaches suffer from the complexity of calibrating the edge detector parameter. Statistical models for segmenting regions require optimization (i.e. Expectation–Maximization algorithm) which is computationally demanding. Similarly, finding the optimal partition of the pixel graph is a demanding process. Recently, there has been a keen interest for representing images using features computed in the neighborhood of points of interest (such as corners) [Lowe, 1999, Bay et al., 2006]. The principal advantage of using local features over region descriptors is the improved robustness to imaging perturbations (change of view point, occlusion, etc...).

Now that the various image structuring elements have been listed we briefly review some of the main features one can compute for image and video annotation and indexing. For more comprehensive review the reader is referred to [Benmokhtar et al., 2011, Zhang et al., 2012]. Those features can be partitioned in three categories; Color, Texture and Shape.

Colors features are certainly the most commonly used for describing the content of an image or a video frame. The color of an image pixel is defined uniquely by 3 values within the chosen color space such as RGB, HSV, YCrCb and HMMD. While RGB (Red/Green/Blue) is very practical for technologies creating color (i.e. displays), some color spaces such as HSV better relate to human perception of color. There has been many colors descriptors proposed in the literature, they either be computed from the entire image, at the regions level or locally. The most compact and also simplest descriptor is the color moments [Flickner et al., 1995]. However, mean, variance and skew are rarely sufficient to describe effectively an image. Histograms [Swain and Ballard, 1990] provide a better representation of color distribution at the cost of higher vector dimensions, yet spatial information is missing. A reduction of the representation dimensions is proposed by the Scalable Color Descriptor [Manjunath et al., 2002] but spatial information is still not available. The Color Correlogram combines color with spatial distance between pixel pairs into a 3D histogram [Huang et al., 1997], providing spatial information at the cost of high dimensionality. Color Coherence Vector incorporates spatial information within the color histogram separating those isolated pixels from connected ones. However, both the dimension and the computation costs remain high. To reduce the color representation's dimension while capturing spatial properties, the Dominant Color Descriptor was introduced [Cieplinski, 2001]. It is particularly compact and has reasonably low complexity, giving it a fair advantage over other color descriptors for describing image regions.

The visual content of an image can also be described in terms of texture [Minka and Picard, 1996, Pentland et al., 1996]. While it is possible to compute the texture of an entire image, it does not make much sense and in most cases textures descriptors are extracted at the region level. Approaches can be classified in two categories; spatial or spectral according to the domain employed. Spatial texture extraction analyses the local structure at the pixel level. The most common spatial approaches are statistical such as the moments [Pratt and Jr., 2007] and the grey level cooccurrence matrix [Clausi and Yue, 2004] or model based such as Markov random fields [Cross and Jain, 1983]. The main drawback of these approaches is their sensitivity to noise and image distortions. Spectral texture features are extracted from the image's frequency domain transform. Well known approaches include Fourier transform [Zhou et al., 2001], Discrete Cosine transform [Smith and Chang, 1994] and Gabor filters [Jain and Farrokhnia, 1991]. Two of the three MPEG-7 texture descriptors are based on Gabor filter responses; Homogeneous Texture Descriptor [Manjunath et al., 2001] and Texture Browsing Descriptor [Manjunath and Ma, 1996]. Those approaches require square regions to be computed; the texture of an free form region is obtained from averaging the sub-square regions features. Robustness to noise is a strong positive property of Spectral approaches which can also be scale and orientation invariant, as in Gabor filters.

The description of shape is highly relevant for recognizing and retrieving objects in images. However, while humans can easily compare shapes, researchers are struggling to achieve similar performance. The literature reports two main categories of shape descriptors, those describing the shape as a contour and those extracting features from a region. The simplest form of region descriptors consist in computing geometric properties of the region, such as area, moments, circularity, etc... Moment invariants features were employed by the first image retrieval systems such as QBIC [Flickner et al., 1995]. Unfortunately, those descriptors only loosely describe shapes and handle shape transformations rather poorly. The Region-based Shape Descriptor (R-SD) provides a better representation of both the interior and boundary of the shape based on Angular Radial Transformation [Manjunath et al., 2002]. It is both compact and robust to segmentation artifacts but does not provide high recall rate in practical situation. Other shape descriptors analyse and express the variations of the shape contours. A number of structural approaches have been proposed based on chain code [Jr., 1961], polygon approximation [Huet and Hancock, 1999, Wolfson and Rigoutsos, 1997] and curve fitting [Rosin and West, 1989]. The difficulty facing structural approaches concerns the decomposition of the contour in basic elements, leading to multiple representation for similar shapes. Global approaches are more inclined to capture the overall shape appearance yet may lack robustness with respect to occlusion. Beside the simple global descriptors such as area, circularity or eccentricity, the Curvature Scale Space (CSS) descriptor [Mokhtarian and Mackworth, 1986] represents the evolution of the contour in terms of curvature at varying scales in a single vector. Spectral transforms (such as Fourier [Chellappa and Bagdazian, 1984] or Wavelet [Tieng and Boles, 1997] descriptors) have also been used to represent object shape contours successfully. However, neither the region based or contour based descriptors provide representation capabilities in line with with human perception of generic shapes. They are however capable of performing object recognition for domain specific tasks.

While the signal processing research community has produced a significant number of low level descriptors, it is still unclear which is/are the most appropriate for understanding multimedia content. In an effort to boost research and multimedia application development, the Moving Picture Expert Group (MPEG) proposed a standard multimedia content description. Many of the low level features listed above are part of MPEG-7 [Manjunath et al., 2002]. However, MPEG-7 produces a substantial data overhead which is rather incompatible with the increasing consumption of multimedia content over low bandwidth network and devices (i.e. mobile phone and tablets). In such application, it is important to have compact yet descriptive description of image content. The trend is to extract and store visual characteristics surrounding specific image locations (i.e. corners) as in Scale-Invariant Feature Transform (SIFT) [Lowe, 1999], Gradient Location and Orientation Histogram (GLOH) [Mikolajczyk and Schmid, 2005], and Speeded-Up Robust Features (SURF) [Bay et al., 2006]. The later favoring descriptor size and low computation time comparing with SIFT.

2.2.2 Models

In the early years of content-based image and video retrieval, the visual representation was also the model [Swain and Ballard, 1991, Flickner et al., 1995, Picard, 1995]. A similarity, or alternatively a distance, measure is employed to compare the low level features of the two images. The L1-norm (Manhattan distance), L2-norm (Euclidean distance), Mahanalobis distance [Mahalanobis, 1936] and the Earth-mover's distance [Rubner et al., 1998] were among the most frequently used. Such approaches suffered drastically from the lack of semantic in the results returned. As an attempt to solve this issue, researchers devised approaches to include users in the loop through relevance feedback [Smeaton and Crimmins, 1997, Rui et al., 1997, Chua et al., 1998, Benitez et al., 1998, Zhou and Huang, 2003]. In spite of improved performance, optimizing features based on human feedback was not sufficient to bridge the semantic gap. Nowadays, the common approach consist in automatically labeling visual content with semantic tags which can then be used for search and indexing using off the shelf text based approaches, such as PageRank [Brin and Page, 1998] or Okapi BM25 [Robertson et al., 1994].

There are roughly two categories of annotation methods; One in which the occurrence of concepts in the multimedia content is estimated using content analysis and classification techniques and the other which combines the metadata associated with the document and its low level features to perform annotation.

The basic workflow of an automatic visual annotation system starts with feature extraction (as seen in the previous section), continues eventually with feature encoding and pooling, before machine learning is performed to learn the models. The most popular feature encoding and pooling method is the widely known Bag-of-Word/Codebook approach which has been ported from natural language processing to computer vision [Csurka et al., 2004]. Once represented using the Bag-of-Word (BoW) model, the concepts can be learned using either a generative or a discriminative model. The prominent drawback of the BoW model originates from its inability to capture the structure and geometric distribution of the image features. For this reason a number of approaches aimed at modeling the spatial relationships have been proposed; spatial feature co-occurences [Savarese et al., 2006], relative position of codewords [Sudderth et al., 2005] or spatial pyramid [Lazebnik et al., 2006].

Support Vector Machines (SVM) [Vapnik and Chapelle, 2000] have become the most widely employed classifier due to their ability to identify optimal class boundaries on both linearly-separable and non-linearly separable problems through the use of kernel-based data transforms [Shawe-Taylor and Cristianini, 200
As a discriminative model, SVMs are generally employed in one-against-all situations [Chapelle et al., 1999, Tong and Chang, 2001, Yan et al., 2003, Shi et al., 2004]. In other words, a SVM learns and models one concept as

a single-class problem, using the positive samples from that concept and the rest of the dataset (all remaining concepts) as negative samples. Although the SVM is capable of learning from few positive samples, it is highly affected by unbalanced datasets [Natsev et al., 2005]. This is an important factor since typically there are far more negative visual examples of a concept than positive. In practice, as many SVM as there are concepts to be modeled are needed. The final concept detection result is based on the SVM's responses (or output probabilities). If a single label is desired (winner take all) the concept with the highest score will be selected, but when multiple concepts are expected the final solution is obtained through simple thresholding or classifier fusion approaches [Essid et al., 2011]. Additionally, the many ways in which the multiple low level features available for detecting concepts are handled is often referred to as feature or low level fusion [Benmokhtar et al., 2011] and leads to diverse system architectures with varying complexity.

Artificial Neural Networks (ANNs) [Haykin, 2007] are an alternative to SVMs and offer the possibility to handle multiple concepts simultaneously. There are not as widely employed as SVM but interesting results have been obtained for automatic image and region annotation [Park et al., 2004, Zhao et al., 2008, Benmokhtar and Huet, 2011]. The main advantage of ANNs is their ease of use and implementation. However, the choice of the architecture (number of layers, number of neurons per layer and neuron activation function) usually results from empirical study. In addition, the ANN learning phase generally requires a larger training dataset than for SVM and is not guaranteed to reach a global optimum. Nonetheless, it remains an attractive alternative in situations such as learning from social media where extensive training samples are available but contain non negligible noise.

Decision Trees (DT) have not been extensively used for modeling visual concepts but there seems to be growing interest in this classification method due to its ability to learn concepts from a limited number of samples and to express them in human understandable terms [Shearer et al., 2001, Wong and Leung, 2008, Liu et al., 2008]. Decision Tree algorithms differ in the type of attribute they handle (discrete for ID3 [Quinlan, 1986] and continuous for C4.5 [Quinlan, 1996] and CART [Breiman et al., 1984]), the way the decision at each node of the tree is chosen (Information gain for ID3, Gain ratio for C4.3 and Gini coefficient for CART) and the structure of the resulting tree (binary for CART versus n-ary for ID3 and C4.3). The work on Random Forest [Moosmann et al., 2008, Uijlings et al., 2011] is a good example of how decision tree can provide efficient and effective image classification. Associative classification is another data mining approach which constructs rules describing the statistically significant patterns in a dataset [Yin and Han, 2003, Mangalampalli et al., 2010]. While the standard approach shares a common drawback with decision trees, better performance on discrete than continuous attributes, performance on the par or better with SVMs have been reported [Quack et al., 2007].

While previous approaches (SVM and DT) provide binary classification and therefore a single label per item, there are many instances in visual content analysis where multiple label are preferred. Probabilistic Bayesian approaches allow such multiple instance multiple label modeling of the content [Dietterich et al., 1997]. The annotation, which corresponds to the posterior probability, is computed in the Bayesian framework from the priors and the conditional probabilities which can either be modeled using non-parametric (clustering) [Vailaya et al., 2001, Shi et al., 2005] or parametric (Gaussian distribution) [Carneiro et al., 2007, Li and Wang, 2008] approaches. Non-parametric approaches are usually more efficient than parametric methods which require expensive optimization process to learn the parameters. However, in both approaches the annotation process is generally too slow to be employed on large collections.

A common issue to all trainable classification approaches, is the availability of a sufficiently large dataset and it's associated ground truth. The most significant effort of the past decade is certainly the high-level feature extraction task of the TRECVid evaluation campaign [Smeaton and Over, 2003], which has provided hours of video content and for which participants jointly participated in the annotation of a few concepts [Ayache and Quénot, 2007]. Indeed, only a selection of a few concepts from the LSCOM onthologie [Naphade et al., 2006] were available and investigated. Recently, larger corpus have been made available to and by the multimedia community thanks to the extensive quantities of medias available through online sharing platform (via specific APIs). Some are associated with recurring evaluation campaigns (such as PASCAL [Everingham et al., 2010] or MediaEval [Larson et al., 2011]) while others are made available by individual research groups (NUS-WIDE [Chua et al., 2009], ImageNet [Deng et al., 2009] or MCG-WEBV [Cao et al., 2009]).

Due to the availability of text surrounding images on Internet Web pages, researchers have considered the possibility of combining the textual information with the visual information in order to improve both annotation and retrieval accuracy; An idea which dates back to 1994 [Srihari and Burhans, 1994] and the work of Srihari *et* al. on face labeling from newspaper using captions [Srihari, 1991]. Forsyth *et* al. [Barnard and Forsyth, 2001] brought the idea to the computer vision and multimedia retrieval community. Their statistical approach models occurrences and co-occurrences of words and region features. Another approach [Cai et al., 2004] consists in grouping images using text first and then using visual features to re-organise images. The biggest issue with using associated text, is to ensure its correctness. Efforts addressing this issue have led to approaches using WordNet [Miller, 1995] to evaluate how words are correlated [Jin et al., 2005, Li and Sun, 2006], or by studying the co-occurrence of words associated with images [Wang et al., 2006, Joshi et al., 2007, Benmokhtar and Huet, 2009, Li et al., 2009]. Another, more recent idea, is to employ data which have either been curated or which originate from known/trusted sources; such as online shopping catalogs for products [Li et al., 2012].

In this section, we have given a bird eye view of the field of multimedia content analysis and understanding. For more detailed readings about the state of the art of this extremely broad research domain the following publications are suggested [Smeulders et al., 2000, Snoek and Worring, 2009, Hanjalic and Larson, 2010, Wang et al., 2010, Bhatt and Kankanhalli, 2011, Lu et al., 2011, Zhang et al., 2012].

The remaining of this chapter concerns some of the research activities undertaken within my research group and under my guidance. The topics range from the low-level representation of visual content to event mining from social media via multimodal fusion and human emotion recognition... In other words, a wide range of topic is covered. Furthermore, the reader should notice a gradually emerging trend; The use of contextual information is becoming increasingly important in the approaches employed to achieve multimedia content analysis and understanding.

2.3 Graph-Based Moving Objects Extraction

The first step of most content based approach consist in identifying the location at which low level features should be extracted. In the large majority of cases, video content analysis is in fact performed at the image level (either though key-frame extraction or sub-sampling). When extracting regions from an image, grouping those regions into coherent, semantic objects, is a non trivial task. Here, we are interested in leveraging from the video motion to automatically segment objects using spatio-temporal segmentation. The challenge is to extract coherent objects from the raw pixel data, considering spatial and temporal aspects simultaneously. Such approaches generally suffer from high complexity and is therefore not practical for processing large video volumes [DeMenthon and Doermann, 2003, Greenspan et al., 2004]. To reduce the computational cost, we consider a new approach that decomposes the problem at different grouping levels including pixel, frame regions, and spatio-temporal regions [Galmar and Huet, 2006]. For each level of the framework, depicted in figure 2.1, we have devised low-complexity graph algorithms and rules for merging regions incrementally and testing the coherence of the groups formed. In this way, we aim to establish strong relations between regions at both local and global levels. The observation of results shows that the process tends to balance between lifetime and coherence of between extracted objects.

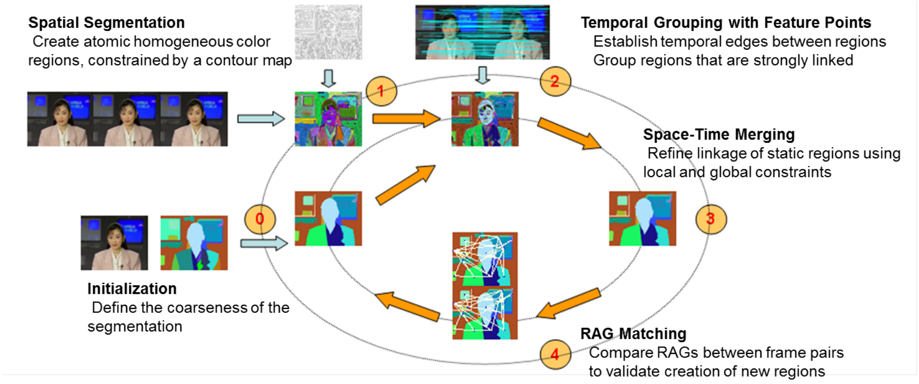

Spatial Segmentation
Create atomic homogeneous color regions, constrained by a contour map
Temporal Grouping with Feature Points
Establish temporal edges between regions
Group regions that are strongly linked
Space-Time Merging
Refine linkage of static regions using local and global constraints
Initialization
Define the coarseness of the segmentation
RAG Matching
Compare RAGs between frame pairs to validate creation of new regions
0
1
2
3
4

Figure 2.1: Our Spatio-Temporal Volume Extraction Framework

The resulting spatio-temporal volumes represented as volume adjacency graphs can then be employed as the basis for extracting the visual features of a semantic video segmentation system [Galmar et al., 2008] or a video object retrieval system [Galmar and Huet, 2008]. It is worth mentioning the work of Wei *et* al. [Jiang et al., 2010] which extends the idea of spatio-temporal video elements to audio-visual atoms for generic video concept classification.

2.4 Fusion of MultiMedia descriptors

With the plethora of low-level feature available to describe multimedia content, each performing better under some specific circumstances than others, it is interesting to study the combined use of multiple descriptors. While most approaches perform feature fusion by averaging or concatenating the low level features [Benmokhtar et al., 2011] (also referred to as static fusion), our work on fusion of multimedia descriptors aims toward machine learned fusion techniques [Benmokhtar and Huet, 2006, Benmokhtar et al., 2008] (or dynamic fusion). In practice, machine learned fusion at the feature level is rather complex, due to the lack of extensive labeled datasets which are necessary for obtaining reliable fusion parameters. In spite of the recent progress provided by support vector machines on the topic of classification based on complex input vectors, the issue of content based classification using information fusion remains among the challenging research topics.

The objective of feature fusion is to reduce the redundancy, the uncertainty and the ambiguity of signatures. Under this conditions, the fused feature vector should yield to better classification performance. Our contribution to feature fusion is characterised by a novel dynamic approach which uses a Neural Network Coder (NNC). The idea is for the NNC to learn the most effective way to compress the feature vector in such a way that it is able to reconstruct it with limited error. In our framework, the compressed feature vector, the hidden layer of the NNC, becomes the input of an SVM classifier which is trained for detecting high level semantic concepts (as in the TRECVid high level concept detection task [Smeaton and Over, 2003]). The performance of the NNC feature fusion is compared with two other fusion approaches: one static (concatenation) and one dynamic (Principal Component Analysis).

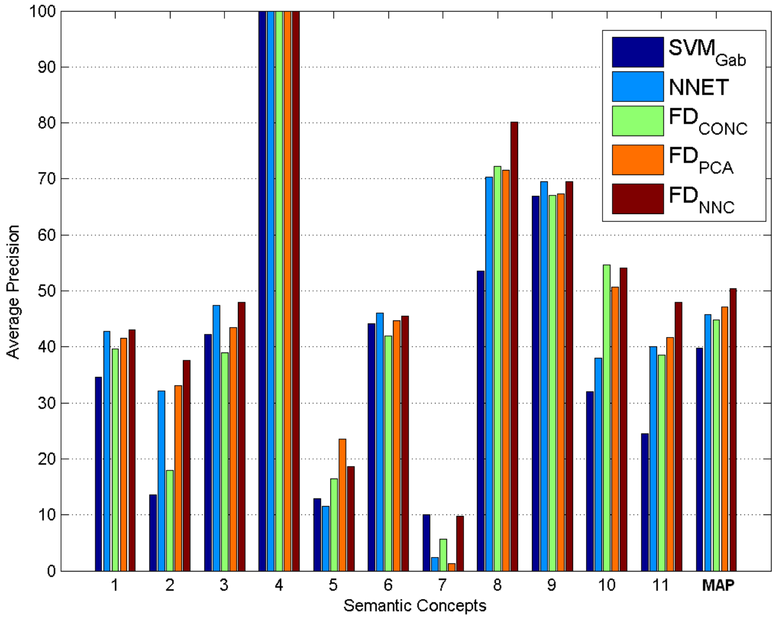


Figure 2.2: Classification performance comparison for 3 Feature Fusion (FD_{xxx}) approaches and 2 without feature fusion (SVM and NNET)

Figure 2.2 shows the results of the 3 feature fusion approaches (FD_{NNC}

is our proposed approach, $\mathrm{FD}_{\mathrm{PCA}}$ is a variant where feature compression is performed using PCA and $\mathrm{FD}_{\mathrm{CONC}}$ is the concatenation of all features) along with the results originating from 2 approaches with no feature fusion. $\mathrm{SVM}_{\mathrm{Gab}}$ corresponds to a system using only 1 low level feature to achieve semantic classification, the Gabor feature in this case which produced the highest performance on this dataset among independent features. The NNET approach is the result of our research on classifier fusion. It is a Neural Network based on Evidence Theory which takes as input the output of the SVM classifiers operating on individual low level features to perform high level fusion. The results for individual concepts (1-11) show the superiority of the NNC feature fusion approach over all other alternatives. One of our findings, is that in some cases feature fusion may lead to better performance than classification fusion [Benmokhtar and Huet, 2006, Benmokhtar et al., 2008, Benmokhtar, 2009].

2.5 Structural Representations for Video Objects

One of the major drawback of local or regions based visual descriptors is that they do not capture the image spatial properties, and when they do, it is at a significant computational cost. We have proposed novel approach for indexing and retrieving video objects based on both geometric and structural information [Souvannavong et al., 2004, Souvannavong et al., 2005]. The objective here is to improve the quality of video indexing, retrieval and summarization by looking at objects (or image regions) within the video instead of the entire frame.

Our effort to achieve this task follows two separate tracks. The first concentrates on the issue of adapting graph based methods in order to efficiently perform the matching of the complex data structures representing video objects on the very large data volumes required for video analysis, we explore the construction of efficient index structures. The other aims at extending a video object classification system using Latent Semantic Analysis (LSA) on image regions with the addition of structural constraints. The LSA technique offers promising results [Sivic and Zisserman, 2003, Souvannavong et al., 2004, Souvannavong et al., 2004, Hörster et al., 2008] in spite of the fact that the automatically segmented regions are characterized by some visual attributes (color, texture and possibly shape) but does not make use of the relationship (connectivity and relative position) between regions (object sub-parts). We have devised a number of techniques aimed at incorporating relational information within the representation and classification. Additionally, we have thoroughly studied and evaluated the alternative structural approaches with our basic implementation [Souvannavong et al., 2005]. The results presented in figure 2.3, show the retrieved video shots which most closely match the query objects (on the left hand side) using either the LSA (bottom row) or our proposed relational-LSA (top row). From this example the benefits of employing structural information are exposed clearly. Further experiments have shown improvement of performance can be achieved by using the relational-LSA. However, the performance of the relational-LSA are somewhat undermined by the frequent instability of the region extraction process [Souvannavong et al., 2005]. Theses observations have led to study of algorithms using both image regions and points of interest, as well as research on the topic of spatio-temporal segmentation algorithms (section 2.3).

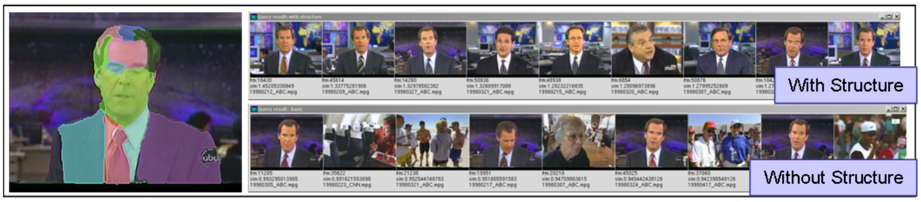

Query result: with structure
frm:16430 sim:1.45205330849 19980212_ABC.mpg
frm:45614 sim:1.33775281906 19980209_ABC.mpg
frm:14260 sim:1.32978582382 19980327_ABC.mpg
frm:50836 sim:1.32689917088 19980321_ABC.mpg
frm:48938 sim:1.28232216835 19980215_ABC.mpg
frm:6854 sim:1.28096973896 19980320_ABC.mpg
frm:50876 sim:1.27995252609 19980307_ABC.mpg
With Structure
Query result: basic
frm:11205 sim:0.993295013905 19980305_ABC.mpg
frm:35622 sim:0.991621553898 19980223_CNN.mpg
frm:21238 sim:0.952544748783 19980321_ABC.mpg
frm:19951 sim:0.951805591583 19980217_ABC.mpg
frm:20218 sim:0.94759863615 19980307_ABC.mpg
frm:45025 sim:0.945442438126 19980324_ABC.mpg
frm:37060 sim:0.942398548126 19980417_ABC.mpg
Without Structure

Figure 2.3: TRECVid retrieval example comparing LSA with Relational-LSA

2.6 Spatio-Temporal Semantic Segmentation

As a natural extension to our work on spatio-temporal segmentation of video objects, we propose a framework where spatio-temporal segmentation and object/region labeling are coupled to achieve semantic annotation of video shots [Galmar et al., 2008]. On one hand, spatio-temporal segmentation utilizes region merging and matching techniques to group visual content. On the other hand, semantic labeling of image regions is obtained by computing and matching a set of visual descriptors to a database. The integration of semantic information within the spatio-temporal grouping process sets two major challenges. Firstly, the computation of visual descriptors and their matching to the database are two complex tasks. Secondly, the relevance of the semantic description depends also on the accuracy of visual descriptors, which means that the volumes should have sufficient size. To this aim, we introduce a method to group semantically spatio-temporal regions within video shots. We extract spatio-temporal volumes from small temporal segments. Then, the segments are sampled temporally to produce frame regions. These regions are semantically labeled and the result is propagated within the spatio-temporal volumes. After this initial labeling stage, we perform joint propagation and re-estimation of the semantic labels between video segments. The idea is to start by matching volumes with relevant concepts, re-evaluate and propagate the semantic labels within each segment and repeat the process until no more matches are found.

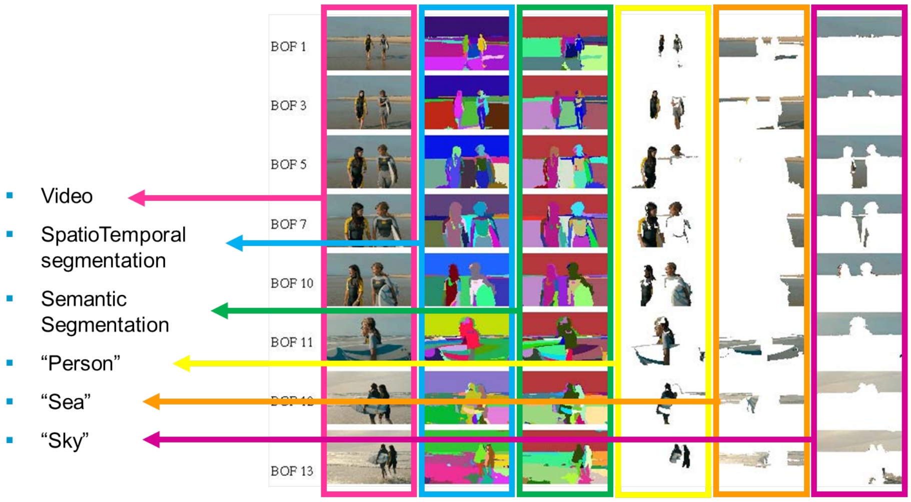


Figure 2.4: Spatio-Temporal Semantic Segmentation of a Video Sequence

The result of this approach on a "beach" video sequence is presented in figure 2.4. The first column shows some frames from the video, the second the result of the spatio-temporal segmentation at the level of block of frames (BOF) and the third the harmonized segmentation over the entire sequence (note the consistency in coloring). The remaining three columns show the regions labeled with the semantic concepts "person", "sea" and "sky" respectively. While the approach still shows some limitation when smaller regions with similar visual content (but different semantic content) are present (see the people's shadows on BOF 13 wrongly labeled as "sky"), overall, the dominant regions/objects of the scene have been assigned the correct semantic annotations.

2.7 Fusion of MultiMedia Classifiers

Fusion of multiple classification algorithms is currently employed by the most advanced multimedia indexing and annotation systems yet it is still an active research topic. Some approaches are based on the selection of the, or some of the, best classifiers in as a mean to perform fusion. Others, use the scores obtained by many classifiers via basic mathematical operations such as sum, product, min-max, etc. Fusion may also be seen as a classification problem using Bayesian methods or support vector machines. More advanced systems attempt to perform both classification and fusion at once. This is the case of Boosting algorithms [Freund and Schapire, 1996] which combine the results of many simple classifiers in order to improve overall classification performance.

One approach we have proposed, to address the fusion of classifier problem, consists in determining the fusion formula with a genetic algorithm [Souvannavong and Huet, 2005]. The hierarchical structure which represents the fusion function is learned during the genetic optimization process along with the selection of the operators (i.e. min, max, sum, product, etc...) and weighting parameters. The resulting fusion chain corresponds to a binary tree of fusion operators.

Another approach has also been proposed [Benmokhtar and Huet, 2007], based on Neural Network Evidence Theory (NNET). The technique applied consists in applying Demspter-Shafer theory [Shafer, 1976] to the problem of classifier fusion, in order to benefit from both belief (or support) and plausibility information. We have devised a Neural Network which learns and performs such a combination of classifier outputs. Theses approaches along with well known techniques reported in the literature (such as GMM, Bayesian Naïve, Multilayer Perceptron and SVM) have been thoroughly evaluated and compared on the TRECVid'05 and '06 dataset [Benmokhtar, 2009]. The results show the benefits of combining the output of multiple classifiers.

Individual semantic concepts are best represented or described by their own set of descriptors. Intuitively, color descriptors could be better to detect certain concepts such as "sky, snow, waterscape, and vegetation", than "car, studio, meeting". With this observation in mind, we propose to weight each low-level feature at the fusion level according to its entropy and perplexity measure. The novel "perplexity-based weighted descriptors" are fed along with the classifier's outputs to our evidential combiner NNET, to obtain an adaptive classifier fusion called PENN (Perplexity-based Evidential Neural Network) [Benmokhtar and Huet, 2008].

In figure 2.5, we show the performance of a simple system "No-weight" where all descriptors are taken as equal in terms of relevance to all semantic concepts (this corresponds to the NNET case), and compare it with four evolution models of weights (Softmax, Sigmoid, Gaussian, and Verhulst). Our proposed model based on Verhulst has the best average precision for all TRECVid'07 semantic concepts. As an example, to detect the "face",

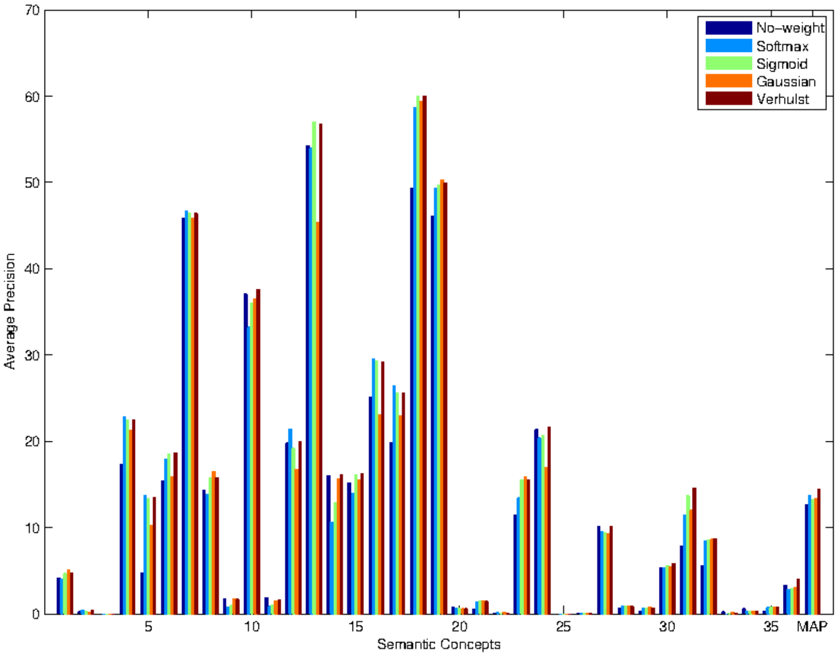


Figure 2.5: Performance comparison of 5 high level fusion approaches across the 36 TRECVid'07 concepts

"person" or "meeting" concepts, PENN gives more importance to "FaceDetector", "ContourShape", "ColorLayout", "ScalableColor" and "EdgeHistogram" than others descriptors. For the "person" concept, the improvement is as high as 11%, making it the best performing run. Overall the addition of the perceptual weights leads to improved fusion performances.

2.8 Human Emotion Recognition

Recognising human emotions from video data provides valuable high level semantic cues about the content of the video and is certainly relevant for multimedia content understanding and indexing. Our contribution is SAMMI: Semantic Affect-enhanced MultiMedia Indexing, a framework explicitly designed for extracting reliable real-time emotional information through multimodal fusion of affective cues [Paleari and Huet, 2008]. While our aim in this work is to annotate video content with emotional information for content indexing, there are many other application for such high level concepts.

For example, domains in which human emotion recognition would have a significant impact are e-learning and remote meetings where the affective state of the participants is known to be important to the success of the activity [Paleari et al., 2009].

Our approach to emotion recognition from video aims are identifying the 6 basic "universal" human emotions (anger, disgust, fear, happiness, sadness, and surprise) as defined by [Ekman et al., 1987]. To achieve this we train classifiers on two modalities; Visual (Facial expressions) and Audio (Vocal prosodic information). Multimodal fusion of affective cues extracted through these two modalities is used to increase precision and reliability of the recognition [Paleari et al., 2009].

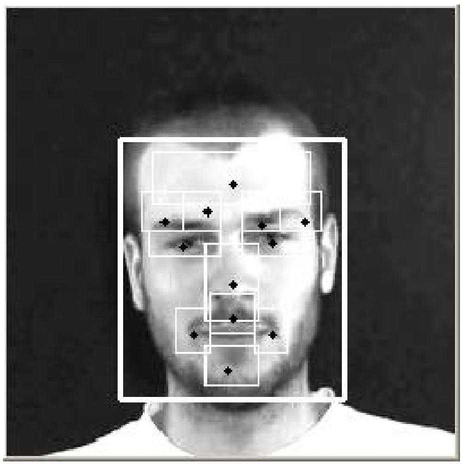


Figure 2.6: Facial Feature Points.

We extract information about the facial expressions by tracking the position of eleven, automatically extracted points of interests (see figure 2.6). Audio is analyzed and features such as pitch, pitch contours and MFCC (Mel Frequency Cepstral Coefficient) are extracted. A thorough study of the discriminative power of the various low level features for emotion recognition has been performed and is available in [Paleari et al., 2010b]. Classification of video and audio features is achieved using individual classifiers (SVM and NN). Classifier fusion is employed to improve recognition reliability by taking into account the complementarity between classifiers.

In Table 2.1, we compare the results obtained from the NNET classifier fusion (section 2.7) with the unimodal NN and SVM classifier outputs. These experiments have been completed using the eNTERFACE'05 database [Martin et al., 2006]. The results show a significant gain in accuracy when multimodal cues are combined to decide on the emotion ex-

	Anger	Disgust	Fear	Happiness	Sadness	Surprise	**CR$^+$**	**MAP**
Video NN	0.420	0.366	0.131	0.549	0.482	0.204	0.321	0.205
Audio NN	0.547	0.320	0.151	0.496	0.576	0.169	0.354	0.234
Video SVM	0.342	0.342	0.193	0.592	0.426	0.244	0.320	0.211
Audio SVM	**0.627**	0.220	0.131	0.576	0.522	0.162	0.361	0.253
NNET	0.542	**0.388**	**0.224**	**0.633**	**0.619**	**0.340**	**0.428**	**0.337**

Table 2.1: Emotion recognition accuracy.

pressed by the person being monitored. On this dataset, some emotions are difficult to identify accurately (Disgust, Fear, Surprise) while for others (Anger, Happiness and Sadness) results are much more encouraging. Low performance for some emotional states could be inherent to the dataset employed. In [Martin et al., 2006] it is mentioned that some videos do not represent the desired emotional expression to a satisfactory level. This is essentially due to the fact that subjects' portrayed in the dataset are not trained actors and mostly non-native English speakers.

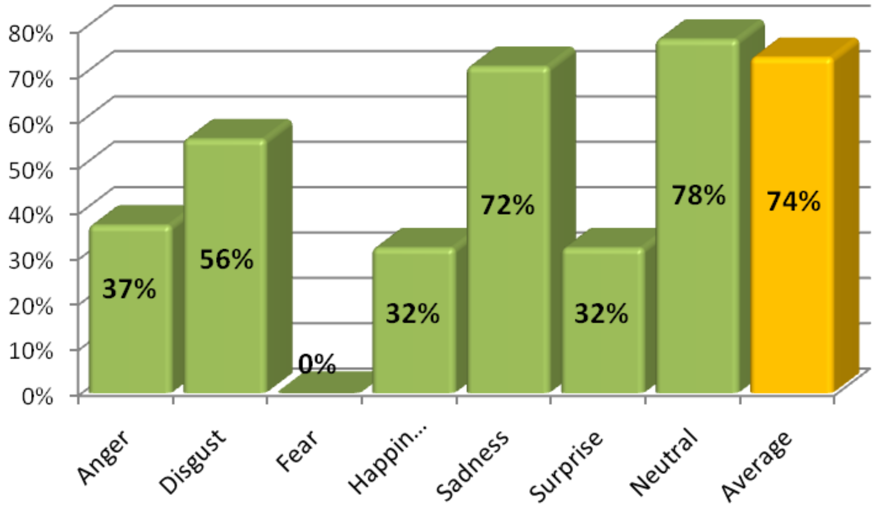


Figure 2.7: Emotion recognition accuracy on real video sequences (TV shows and Movies)

We have also experimented with the possibility of recognising emotion from movies and other TV material using the models learned from the eNTERFACE'05 dataset [Paleari et al., 2010a]. In order to evaluate our approach, a collection of 107 short YouTube videos showing at least one character in frontal view was created and labeled by a group of people (530 tags in total). Figure 2.7 shows the accuracy at which our system is able to recognise real world multimodal human emotions. On average the system identifies accurately 74% of emotions, thanks to the addition of the "neutral" emotional state (which composes over one third of the dataset). Again, our approach performed badly for "fear", but other emotion were detected with either moderate (Anger, Happiness, Surprise) or good accuracy (Disgust, Sadness, Neutral). The results on this dataset confirm that our emotion recognition approach is sufficiently generic for person independent detection of human emotions in real video sequences.

2.9 Large Scale Multimedia Annotation

With the extraordinary success of multimedia sharing website (YouTube, Flickr and Facebook to mention only a few), the colossal amount of media documents (and more particularly video) available on the Internet is reenforcing the need for semantic analysis. We have addressed the problem of Web video annotation using both content-based information originating from visual characteristics and textual information associated with the multimedia documents [Liu and Huet, 2010, Liu and Huet, 2010]. At first, we have created a video dataset for our research. To address the online videos analysis problem, the size of the collected dataset should be as large as we can possibly handle. Using the meta-data associated with the most popular videos on YouTube, we selected 1875 popular and meaningful tags and used them as query seeds to retrieve 42000 videos. For each of the video in our dataset we perform low level feature extraction. Since we have participated in several edition of TRECVID [Smeaton and Over, 2003], we have a number of high level feature detectors along with their corresponding representative ground truth (training set). However, there is an important difference of content between the TRECVID dataset and the large scale corpus we have collected. In addition, it is unconceivable to manually annotate all videos uploaded on the web. Our initial study focuses on an approach where the training set is extented using selected web content for refining and improving high level concept models without human interaction. For the purpose of evaluation, we have manually labeled 8000 video shots out from our dataset with the 39 semantic concepts used in TRECVID 2007. A third of the labeled data is used to train an SVM and is evaluated on remaining two thirds of the data. The trained detectors are then employed to detect semantic concepts on the large scale corpus for which ground truth is not available. The video shots with high detection scores are selected as new training examples which are then learned by the SVM concept detectors.

We also study how to utilize the text metadata to assist visual fea-

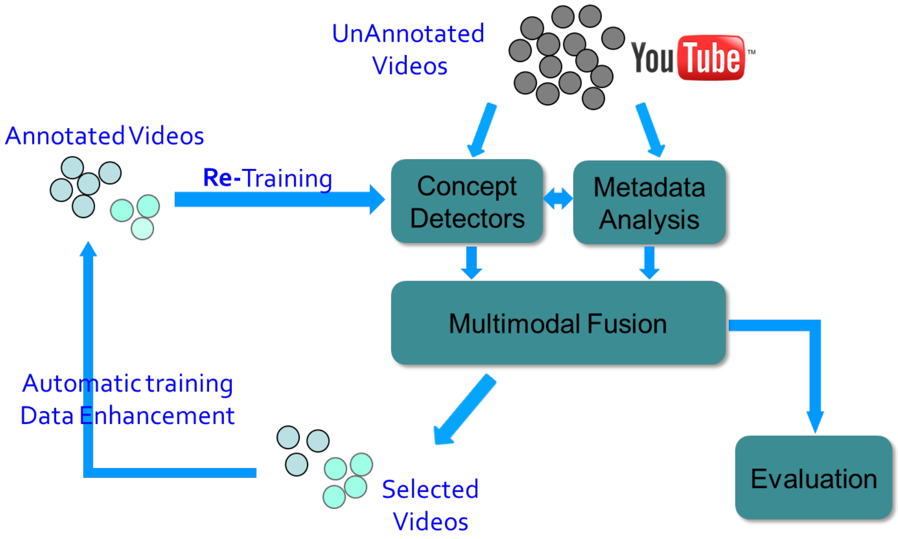


Figure 2.8: The Large Scale Automatic Concept Refinement Framework

ture and enhance the performance of concept classifier on unlabeled videos. Compared with traditional videos, social videos are commonly accompanied with metadata such as tags, description, script, etc, which are provided by the users themselves. Although those textual information are usually erroneous, sparse, and not accurate enough to provide the required knowledge for effective content-based retrieval, the analysis of the auxiliary text shows possibility of improving the performance of traditional multimedia information analysis approaches. Additionally, when correctly tagged by their authors or other contributors, such information could really benefit concept modeling.

This problem is addressed by the framework shown in figure 2.8. We query with keywords for each concept from our dataset, and initialize the annotation of all the shots with such concept for each returned video entity. Then, trained visual concept detectors are run on those shots, and used to sort the result by visual similarity. Those video shots whose probability exceeds a given threshold are reserved to augment the training set.

Figure 2.9 shows the accuracy of 4 approaches for detecting high-level concepts in video shots; the original model trained on TRECVid data (Before Refining), the model refined using new training samples based on visual information only (Visual Refining), the model refines using new training samples obtained from both textual and visual information pruning (Tag Refining) and finaly an approach based on text only (Tag Query). Our experiments show that the automatic training data enhancement significantly

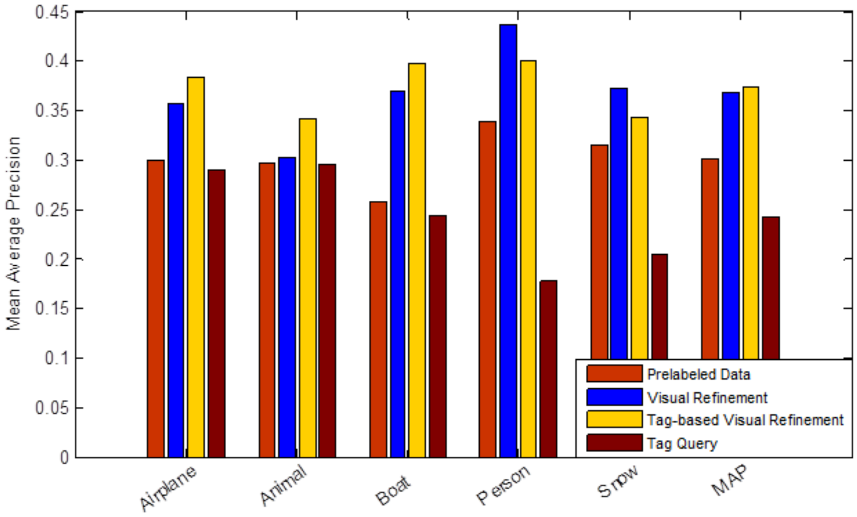


Figure 2.9: Semantic concept detector refining accuracy (Average Precision)

improves the accuracy of the semantic concept detectors without requiring additional human efforts. Looking at figure 2.9 a little closer, we see that using the model refinement approaches provide improved results with respect to the initial models. The results using "Tag Query" only clearly show that the quality of the tags associated with online media fluctuates greatly; It always gives the worse retrieval accuracy. On average over all concepts, the initial pruning of the dataset using tags followed by visual inspection ("Tag Refining") gives improved results over the approach using visual inspection alone. The improvements are due to the fact that, thanks to the double constraints (text and visual) on the candidate video shots, less erroneous exemplars are added to the training set for concept detector refinement.

2.10 Mining Social Events from the Web

Events are a natural way for humans to organize information, media are no exception. In the work presented here, an event relates to a scheduled activity grouping a number of persons in a given location during a given time-span. Examples of such events are happenings like Concerts, Shows, Conferences, Sport Games etc... as featured in the MediaEval Social Event

Detection Challenge [Papadopoulos et al., 2011]. This should not be confused withanother type of event found in the litterature which are concerned with detecting human actions [Gravier et al., 2012, zhong Lan et al., 2012, Fiscus, 2012].

It is observed that many photos and videos are captured and shared on the Web during and after social events take place. People frequenly upload media captured with their mobile phones on media sharing platforms or social networks while or shortly after attending a show or any other event. Here, we aims at detecting these events from social media data leveraging on burst detection techniques [Liu et al., 2011]. Our approach to detect and identify events consists of 3 steps:

- Location Monitoring: finding the bounding-box of venues.
- Temporal Analysis: detecting events by analyzing the uploading behavior along time.
- Event Topic Identification: identifying detected events' topics through tag analysis.

Venue bounding-boxes are learned from the photos taken during past events. We extract the GPS information from photos tagged with the eventID, remove potential outliers and keep the min rectangle from the remaining photos as the representative venue bounding-box. An example of such bounding-box is depicted in figure 2.10(right) for the venue "Hammersmith Apollo" in London, UK.

We analyze the uploading behavior (the number of daily media uploads) on a given bounding box during a period. Events are detected by burst detection techniques. We use the combined number of photos and different owners (the number of people who shared them) to express the activity at a given location. The detection is based on whether the number of uploads exceeds a threshold or not. We have experimented various thresholding approaches and identified that using the median over a temporal window of one month produces on average the most accurate results. The figure 2.10(right) illustrates the activity over a one month period at the "Hammersmith Apollo" in London, UK.

The approach as been evaluated on a dataset of 242 social events which took place in 9 different venues all over the world. Although, not all of these 242 events have been captured and shared online by users, our approach is able to accurately detect 67 of them. Interestingly, out of the 67 events we automatically detected 17 were not referenced in the LastFM [6]

[6] http://www.last.fm

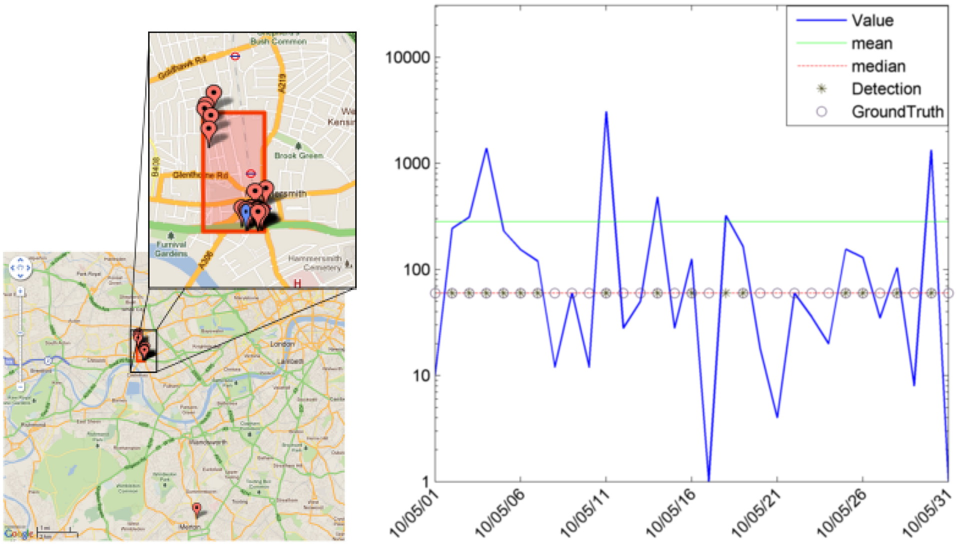


Figure 2.10: An example of Venue Bounding Box (left) The uploading pattern for a venue over a month (right)

event repository. Users are increasingly capturing and sharing medias on social portals when attending events. This enables to accurately detect and identify social events such as concerts, festivals, shows, etc... in an implicit manner [Liu et al., 2011a].

2.11 Event Media Mining

Events are often documented by people through different media such as videos and photos. Many of those documents will be uploaded on online social sharing platforms such as Facebook, Flickr and YouTube (to mention only the largest). However, only few of those photos and videos available on the web are explicitly associated to an event using a machine tag from a major online event directory such at Last.fm, Eventful or Upcoming. Our goal is to mine the web for as much as possible media resources that have NOT been tagged with a lastfm:event=xxx machine tag but that should still be associated to an event description.

We are investigating several approaches to find those photos and videos to which we can then propagate the rich semantic description of the event improving the recall accuracy of multimedia query for events.

Starting from an event description, the three dimensions from the LODE [7]

[7] Linking Open Descriptions of Events

model [Shaw et al., 2009] can easily be mapped to metadata available in Flickr and be used as search query in these two sharing platforms: the "what"-dimension that represents the title, the "where"-dimension that gives the geo-coordinates attached to a media, and the "when"-dimension that is matched with either the taken date or the upload date of a media. Querying Flickr or YouTube with just one of these dimensions bring far too many results: many events took place on the same date or at nearby locations and the titles are often ambiguous. Consequently, we will query the media sharing sites using at least two dimensions. We also find that there are recurrent annual events with the same title and held in the same location, which makes the combination of "title" and "geo tag" inaccurate. We consider the two combinations "title"+ "time" and "geotag"+ "time" for performing search query and extending media that could be relevant for a given event.

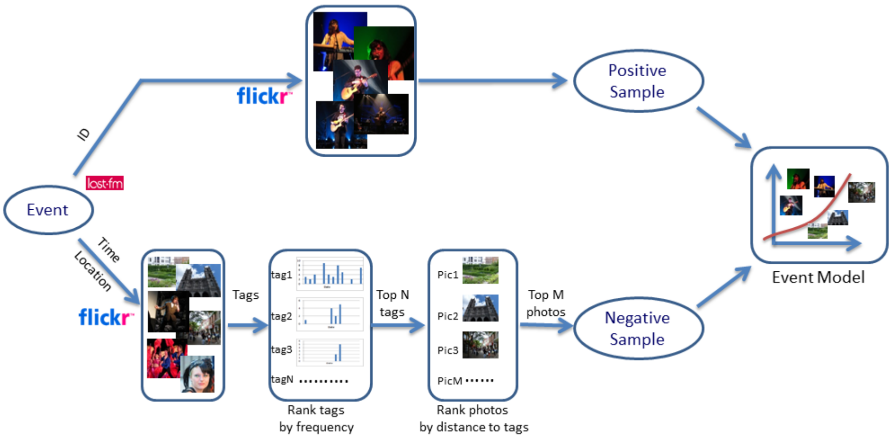


Figure 2.11: Our proposed framework for Mining and Modeling Social Events

Searching online sharing platforms (such as Flickr or YouTube) as described above retrieves many photos with a clear description and association to events, but also returns many photos which do not originate from the event itself. In particular, photos without any textual description (only "geotag" + "time") may not be related to the event under consideration. Multimedia content analysis can is used to address this issue and discard the photos which do not have sufficient visual similarity with media known to arise from the event. In our framework [Liu and Huet, 2012], shown in figure 2.11, we assume that the photos that are corresponding to the same event should be similar visually. Visual pruning is employed to remove the noisy photos from the results of the event identification model. Candidate photo are compared with those which have been explicitly associated to the event using machine tag. Photos are then sorted according to their shortest distance to one of the representative. The bigger the distance and the less similar the photo is with the photo cluster, so we prune the photos with such a large distance. Experimentally, we remove those photos whose distance from the nearest model photo is larger than the mean distance between model photos.

Owner refinement is another approach we employ to improve the detection results [Liu et al., 2011b]. Based on the assumption that a person cannot attend more than one event at a given time, all the photos that have been taken by the same person during the event duration can reasonable be assigned to the same event. Using this heuristic, it is possible to retrieve photos which do not have any textual description.

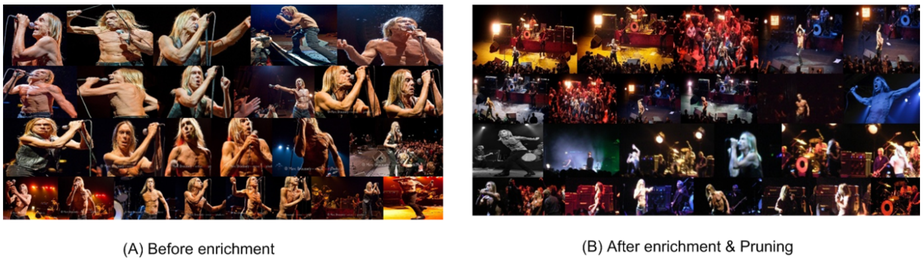


Figure 2.12: Illustrating an event using mined web media documents

From the resulting photo set a visual mosaic is created to show a vivid interface for users. Figure 2.12 shows the visual media for a known event (Date: May 2010, Location: Hammersmith Apollo, Title: iggy stooges) before and after our Event Media Mining approach is performed. During the media mining and enrichment phase, we expect to bring more diverse photos into the collection. The visual on the left (A) is generated from the relevant photos (featuring the event machine tag) while visual (B) shows the collection of images resulting from our enriching and visual pruning method. We can clearly see the increased visual diversity of the scenes between the two sets. The final set of images illustrating the "Iggy stooges" event will be composed of both sets.

3 Conclusions and Future Directions

We have entered a ubiquitous world where technologies are enabling us to capture, share and receive multimedia information regardless of our location and activity. As a result our communication behavior is changing. It is changing in both the way we are sending information as well as how we are receiving and consuming it. Technology is giving us the tools to broadcast our personal "news" to whom we see fit, while at the same time allowing us to experience information in an increasingly personalised fashion. As the amount of multimedia data circulating on the Internet increases at an overwhelming pace, the need for methods automatically extracting the semantic embedded within digital documents is gaining importance. The benefits of understanding the content of multimedia documents are many folds, ranging from providing cues for indexing and retrieval systems to feeding hints for intelligence/knowledge creation systems. Without multimedia content understanding the Internet could soon become a large scale multimedia data graveyard...

In Chapter 2, we have presented a number of approaches contributing to the state of the art for multimedia content analysis and understanding. There is a common theme to most, if not all, the research we have undertaken in those past few years; Bringing Context to Content. The way in which context is dealt with varies from one approach to the next, although we can distinguish between two types of context; Contextual information available within the document itself on one side (internal) and external information on the other (contextual information that arises from data associated with the document such as geo-coordinates or EXIF data). The idea of using context in addition to content has been strongly promoted and supported by Ramesh Jain [Sinha and Jain, 2008] whom invented a new word to express it: Contenxt. Furthermore, according to the latest Accenture Technology Vision [1] context-based technologies will be at the core of the next generation of digital services.

[1] http://www.accenture.com/us-en/technology/technology-labs/Pages/insight-accenture-technology-vision-2012.aspx

In our spatio-temporal video segmentation work (Section 2.3) contextual information is present at almost pixel level, when deciding whether a group of pixel should be merged or not with its neighbor. Another example making use of internal contextual information was presented in section 2.5. In our work on structural representations for video objects, a Latent Semantic Analysis framework is employed to learn the latent association between high level concepts concepts and the context in which individual image regions occur (co-occurrence of neighboring regions). The work on high-level fusion (Section 2.7) also benefits from internal contextual information. Indeed, our Evidence Theory based Neural Network, is making extensive use of co-detected concepts in order to compute the final confidence score in a semantic concept's presence. However, the final step of the fusion process [Benmokhtar and Huet, 2011] consist in an ontology-based re-ranking which provide external knowledge and improved performance to the approach. When semantically labeling those segmented regions (Section 2.6) obtained thanks to our spatio-temporal video segmentation approach (Section 2.3), external knowledge was also brought in thanks to an ontology. Our current and most recent work largely builds on the availability of context data. The concept model refinement approach (Section 2.9) combines visual feature comparison and user contributed annotations to search and identify new positive training sample directly from the web. The external context provided by user annotation and comment helps creating visual semantic models with higher performance and enhanced generalisation properties. Both the works on social event detection (Section 2.10) and event media mining (Section 2.11) make extensive use of contextual information available in conjunction with media documents shared from the Web. Our proposed approaches, employ image and video geo-coordinates, capture time, machine tags, annotations to detect and identify events automatically. A visual model corresponding to specific events [Liu and Huet, 2012] can learned from the data gathered during detection and identification in order to mine the media sharing platforms (such as Flickr) for more related multimedia material. The media mining makes use of yet another contextual information, the media owner (the uploader of the document), in order to collect media for which no additional information is available and enhance the diversity of the illustrations related to a given social event. Last but not least, our emotion recognition work (Section 2.8), can been seen as a context creation process for enriching media featuring humans with affective cues. Such information is valuable in many situation, such as for example, when classifying movies into genre or when providing feedback about user experience in a serious game. Overall, the addition of contextual information, whether internal or external to the media, contributes significantly to understanding the content of multimedia documents.

The field of multimedia content analysis and understanding has now passed its infancy [Smeulders et al., 2000]. Significant progress have been made and numerous solution proposed [Snoek and Worring, 2009]. During a panel session at the prestigious ACM Multimedia Conference of 2006, Tat-Seng Chua controversially suggested that since only incremental progress is being made on video search, the problem is mostly solved. The statement certainly created the desired stir within the community and it was Alex Hauptmann whom a few month later provided a reply based on experimental results showing that video search approaches were still offering poor generalisation properties [Hauptmann et al., 2007]. Are today's search engines able to search multimedia documents (image and videos) with the same level of details, the same granularity, the same efficacy as they do for text? There is clearly scope for improvement, and our objectives in term of future research directions aim in that direction.

In Accenture's Technology Vision [2] the ability to share data with others enhances its value when and if properly managed. Moreover, Social-driven IT will have a true business-wide impact. As most of Internet traffic is soon to be video content, understanding multimedia content is becoming an even more challenging and important enabling step for many applications, among which the long awaited multimedia semantic search. What is happening in a video?, who are the people being viewed? what are they talking about? where and when is the action taking place? These are some of the typical questions which one would like to have answers to about multimedia content, regardless of whether it is shared on dedicated platforms or privately owned. These semantic cues about the content have importance to summarization [Huet and Mérialdo, 2006], recommendation [Kohrs et al., 2000], business intelligence, etc... Here, we will highlight three research topics aiming at answering semantic questions regarding multimedia content; Are these multimedia content related? What event is this depicted in this multimedia document? What is this video about?

The area of digital television is right in the middle of the convergences of technologies, devices and media. While current interactive television attracts only very few users, many admit using a second screen (laptop, tablet or smartphone) while watching TV to search and browse the web for additional information about programs [3]. This implies that interactive television is of interest only just not in the way it is currently proposed.

[2] http://www.accenture.com/us-en/technology/technology-labs/Pages/insight-accenture-technology-vision-2012.aspx

[3] Research by Nielsen in the US found that 70% of tablet users and 68% of smartphone owners use their device while watching television

Viewers are interested by additional content supporting the program they are seeing. There is a need for mining the web for additional content and for automatically presenting the most relevant ones to the viewer. Additional content may vary from live discussions on Twitter or Facebook, to articles on Wikipedia or news channels (BBC, AFP, etc...) or other somewhat similar or at least related audiovisual programs originating from various sources (YouTube, DailyMotion, etc...). Identifying related audiovisual contents is a challenging task which goes well beyond near duplicate detection. Is it the task we are involved in within the newly started LinkedTV [4] european project. Our role is to mine the web of media for content related to broadcasted material. Our initial approach will extend the visual concept refinement framework (presented in section 2.9), to extract, annotate and index a wide range of semantic concepts ranging from people, objects and events based on both audio-visual content and contextual information.

Events are a natural way for humans to structure and organise their activities. Whether it is a vacation, a meeting, a birthday or any other activity, events take place at a given time and place, and feature one or more persons. Events can be categorised as public if open to all (presidential election ballot, a football game, a concert, etc...) or private if attended by invited people only (wedding, birthday, holiday, etc....). Our current work on social event detection [Liu et al., 2011b] and media mining [Liu and Huet, 2012] is essentially concentrating to very specific events (i.e. public events taking place in specific venues). It is well suited for identifying an event based on a collection of media as well as searching for additional documents illustrating the event. The answer to the question ”What event is this depicted in this multimedia document?” requires a more generic and more complex approach combining multimodal content analysis and web-scale data analysis/mining in order to associate the media segment with the corresponding event. This research direction is supported by the Alias [5] (AAL) and the EventMap (EIT ICT Labs) projects. This years premier Multimedia conference (ACM Multimedia) proposes nothing less than 2 grand challenges related to event mining: NTT Docomo Challenge: Event Understanding through Social Media and its Text-Visual Summarization and Technicolor Challenge: Audiovisual Recognition of Specific Events. This show how timely our research in this direction is.

The increase of video content on the Internet, reflects a shift in the way people are communicating. As an example, while it was necessary to read specialized press to obtain technical reviews about consumer products, it

[4] http://www.linkedtv.eu/

[5] www.aal-alias.eu

is becoming common for consumer themselves to post video detailing their experience of a product on media sharing platforms such as YouTube. The information embedded in such videos, coupled with more traditional product reviews constitute extremely valuable knowledge for consumers as well as products manufacturers and retailers. Due to the wealth and diversity of online information on brands and products, it is still a challenge to systematically process, analyse and easily comprehend such valuable information. Harvesting the social web in order to extract, understand and index multimedia documents related to products at large is directly in line with some of the research we have undertaken, and which we intend to study further. In particular, the visual concept refinement framework (presented in section 2.9) could be extended to model products and the emotion recognition system (section 2.8) specialised to sense the reviewer's mood (happy, pleased, disapointed etc...). This will contribute significant semantic information for business intelligence applications, one of the key technologies of the next three to five years in Accenture's vision.

Bibliography

[Ayache and Quénot, 2007] Ayache, S. and Quénot, G. (2007). Evaluation of active learning strategies for video indexing. *Sig. Proc.: Image Comm.*, 22(7-8):692–704.

[Barnard and Forsyth, 2001] Barnard, K. and Forsyth, D. A. (2001). Learning the semantics of words and pictures. In *ICCV*, pages 408–415.

[Bay et al., 2006] Bay, H., Tuytelaars, T., and Gool, L. J. V. (2006). Surf: Speeded up robust features. In Leonardis, A., Bischof, H., and Pinz, A., editors, *ECCV (1)*, volume 3951 of *Lecture Notes in Computer Science*, pages 404–417. Springer.

[Benitez et al., 1998] Benitez, A. B., Beigi, M., and Chang, S.-F. (1998). Using relevance feedback in content-based image metasearch. *IEEE Internet Computing*, 2(4):59–69.

[Benmokhtar, 2009] Benmokhtar, R. (2009). *Multi-level fusion for content-based semantic multimedia indexing and retrieval.* PhD thesis, Thesis Eurecom / Telecom ParisTech.

[Benmokhtar and Huet, 2006] Benmokhtar, R. and Huet, B. (2006). Classifier fusion: Combination methods for semantic indexing in video content. In Kollias, S. D., Stafylopatis, A., Duch, W., and Oja, E., editors, *ICANN (2)*, volume 4132 of *Lecture Notes in Computer Science*, pages 65–74. Springer.

[Benmokhtar and Huet, 2007] Benmokhtar, R. and Huet, B. (2007). Neural network combining classifier based on dempster-shafer theory for semantic indexing in video content. In Cham, T.-J., Cai, J., Dorai, C., Rajan, D., Chua, T.-S., and Chia, L.-T., editors, *MMM (2)*, volume 4352 of *Lecture Notes in Computer Science*, pages 196–205. Springer.

[Benmokhtar and Huet, 2008] Benmokhtar, R. and Huet, B. (2008). Perplexity-based evidential neural network classifier fusion using mpeg-7 low-level visual features. In Lew, M. S., Bimbo, A. D., and Bakker, E. M., editors, *Multimedia Information Retrieval*, pages 336–341. ACM.

[Benmokhtar and Huet, 2009] Benmokhtar, R. and Huet, B. (2009). Ontological reranking approach for hybrid concept similarity-based video shots indexation. In *WIAMIS*, pages 226–229. IEEE Computer Society.

[Benmokhtar and Huet, 2011] Benmokhtar, R. and Huet, B. (2011). An ontology-based evidential framework for video indexing using high-level multimodal fusion. *Multimedia Tools and Applications*, pages 1–27. 10.1007/s11042-011-0936-5.

[Benmokhtar et al., 2008] Benmokhtar, R., Huet, B., and Berrani, S.-A. (2008). Low-level feature fusion models for soccer scene classification. In *ICME*, pages 1329–1332. IEEE.

[Benmokhtar et al., 2011] Benmokhtar, R., Huet, B., Richard, G., and Essid, S. (2011). *Feature extraction for multimedia analysis.* Book chapter NÂ4 in &quot;Multimedia Semantics: Metadata, Analysis and Interaction&quot;, Wiley, July 2011, ISBN: 978-0-470-74700-1.

[Bhatt and Kankanhalli, 2011] Bhatt, C. A. and Kankanhalli, M. S. (2011). Multimedia data mining: state of the art and challenges. *Multimedia Tools Appl.*, 51(1):35–76.

[Bimbo et al., 2010] Bimbo, A. D., Chang, S.-F., and Smeulders, A. W. M., editors (2010). *Proceedings of the 18th International Conference on Multimedea 2010, Firenze, Italy, October 25-29, 2010.* ACM.

[Breiman et al., 1984] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees.* Wadsworth.

[Brin and Page, 1998] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7):107–117.

[Cai et al., 2004] Cai, D., He, X., Li, Z., Ma, W.-Y., and Wen, J.-R. (2004). Hierarchical clustering of www image search results using visual, textual and link information. In Schulzrinne, H., Dimitrova, N., Sasse, M. A., Moon, S. B., and Lienhart, R., editors, *ACM Multimedia*, pages 952–959. ACM.

[Cao et al., 2009] Cao, J., Zhang, Y., Song, Y., Chen, Z., Zhang, X., and Li., J. (2009). Mcg-webv: A benchmark dataset for web video analysis. Technical report, Institute of Computing Technology, China.

[Carneiro et al., 2007] Carneiro, G., Chan, A. B., Moreno, P. J., and Vasconcelos, N. (2007). Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(3):394–410.

[Carson et al., 2002] Carson, C., Belongie, S., Greenspan, H., and Malik, J. (2002). Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(8):1026–1038.

[Chapelle et al., 1999] Chapelle, O., Haffner, P., and Vapnik, V. (1999). Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10(5):1055–1064.

[Chellappa and Bagdazian, 1984] Chellappa, R. and Bagdazian, R. (1984). Fourier coding of image boundaries. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(1):102–105.

[Chua et al., 1998] Chua, T.-S., Low, W.-C., and Chu, C.-X. (1998). Relevance feedback techniques for color-based image retrieval. In *MMM*, pages 24–31. IEEE Computer Society.

[Chua et al., 2009] Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., and Zheng, Y.-T. (July 8-10, 2009). Nus-wide: A real-world web image database from national university of singapore. In *Proc. of ACM Conf. on Image and Video Retrieval (CIVR'09)*, Santorini, Greece.

[Ciampini et al., 1998] Ciampini, R., Blanc-Féraud, L., Barlaud, M., and Salerno, E. (1998). Motion-based segmentation by means of active contours. In *ICIP (2)*, pages 667–670.

[Cieplinski, 2001] Cieplinski, L. (2001). Mpeg-7 color descriptors and their applications. In Skarbek, W., editor, *CAIP*, volume 2124 of *Lecture Notes in Computer Science*, pages 11–20. Springer.

[Clausi and Yue, 2004] Clausi, D. A. and Yue, B. (2004). Texture segmentation comparison using grey level co-occurrence probabilities and markov random fields. In *ICPR (1)*, pages 584–587.

[Cross and Jain, 1983] Cross, G. R. and Jain, A. K. (1983). Markov random field texture models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 5(1):25–39.

[Csurka et al., 2004] Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22.

[DeMenthon and Doermann, 2003] DeMenthon, D. and Doermann, D. S. (2003). Video retrieval using spatio-temporal descriptors. In Rowe, L. A., Vin, H. M., Plagemann, T., Shenoy, P. J., and Smith, J. R., editors, *ACM Multimedia*, pages 508–517. ACM.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Li, F.-F. (2009). Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE.

[Deng and Manjunath, 2001] Deng, Y. and Manjunath, B. S. (2001). Unsupervised segmentation of color-texture regions in images and video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(8):800–810.

[Dietterich et al., 1997] Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.*, 89(1-2):31–71.

[Eakins, 1989] Eakins, J. P. (1989). SAFARI - a shape retrieval system for engineering drawings. *Proceedings of 11th BCS Information Retrieval Specialist Group Research Group Colloquium on Information Retrieval*, pages 50–71.

[Ekman et al., 1987] Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W. A., Pitcairn, T., and Ricci-Bitti, P. E. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, 53(4):712–717.

[Enser et al., 2004] Enser, P., Kompatsiaris, Y., O'Connor, N. E., Smeaton, A. F., and Smeulders, A. W., editors (2004). *Image and Video Retrieval: Third International Conference, CIVR 2004, Dublin, Ireland, July 21-23, 2004. Proceedings*, volume 3115 of *Lecture Notes in Computer Science*. Springer.

[Essid et al., 2011] Essid, S., Campedel, M., Richard, G., Piatrik, T., Benmokhtar, R., and Huet, B. (2011). *Machine learning techniques for multimedia analysis*. Book chapter NÂo5 in &quot;Multimedia Semantics: Metadata, Analysis and Interaction&quot;, Wiley, July 2011, ISBN: 978-0-470-74700-1.

[Everingham et al., 2010] Everingham, M., Gool, L. J. V., Williams, C. K. I., Winn, J. M., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338.

[Fiscus, 2012] Fiscus, J. (2012). 2012 TRECVID multimedia event detection track. http://www.nist.gov/itl/iad/mig/med12.cfm.

[Flickner et al., 1995] Flickner, M., Sawhney, H. S., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., and Yanker, P. (1995). Query by image and video content: The qbic system. *IEEE Computer*, 28(9):23–32.

[Forsyth, 2006] Forsyth, D., editor (2006). *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*. IEEE Computer Society.

[Freund and Schapire, 1996] Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. In Saitta, L., editor, *ICML*, pages 148–156. Morgan Kaufmann.

[Galmar et al., 2008] Galmar, E., Athanasiadis, T., Huet, B., and Avrithis, Y. S. (2008). Spatiotemporal semantic video segmentation. In [Sikora et al., 2008], pages 574–579.

[Galmar and Huet, 2006] Galmar, E. and Huet, B. (2006). Graph-based spatio-temporal region extraction. In Campilho, A. C. and Kamel, M. S., editors, *ICIAR (1)*, volume 4141 of *Lecture Notes in Computer Science*, pages 236–247. Springer.

[Galmar and Huet, 2008] Galmar, E. and Huet, B. (2008). Spatiotemporal modeling and matching of video shots. In *ICIP*, pages 5–8. IEEE.

[Gravier et al., 2012] Gravier, G., Demarty, C.-H., Baghdadi, S., and Gros, P. (2012). Classification-oriented structure learning in bayesian networks for multimodal event detection in videos. *Multimedia Tools and Applications.*

[Greenspan et al., 2004] Greenspan, H., Goldberger, J., and Mayer, A. (2004). Probabilistic space-time video modeling via piecewise gmm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:384–396.

[Hanjalic and Larson, 2010] Hanjalic, A. and Larson, M. (2010). Advances in multimedia retrieval, part i: frontiers in multimedia search. In [Bimbo et al., 2010], pages 1773–1774.

[Hauptmann et al., 2008] Hauptmann, A. G., Wang, J. J., Lin, W.-H., Yang, J., and Christel, M. G. (2008). Efficient search: the informedia video retrieval system. In [Luo et al., 2008], pages 543–544.

[Hauptmann et al., 2007] Hauptmann, A. G., Yan, R., and Lin, W.-H. (2007). How many high-level concepts will fill the semantic gap in news video retrieval? In Sebe, N. and Worring, M., editors, *CIVR*, pages 627–634. ACM.

[Haykin, 2007] Haykin, S. (2007). *Neural Networks: A Comprehensive Foundation (3rd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

[Hörster et al., 2008] Hörster, E., Lienhart, R., and Slaney, M. (2008). Continuous visual vocabulary models for plsa-based scene recognition. In [Luo et al., 2008], pages 319–328.

[Huang et al., 1997] Huang, J., Kumar, R., Mitra, M., Zhu, W.-J., and Zabih, R. (1997). Image indexing using color correlograms. In *CVPR*, pages 762–768. IEEE Computer Society.

[Huet and Hancock, 1999] Huet, B. and Hancock, E. R. (1999). Line pattern retrieval using relational histograms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(12):1363–1370.

[Huet and Mérialdo, 2006] Huet, B. and Mérialdo, B. (2006). *Automatic video summarization.* Book chapter in &quot;Interactive Video, Algorithms and Technologies&quot; by Hammoud, Riad (Ed.), 2006, XVI, 250 p, ISBN: 3-540-33214-6.

[Ikeuchi et al., 2003] Ikeuchi, K., Faugeras, O., Malik, J., Triggs, B., and Zisserman, A., editors (2003). *9th IEEE International Conference on Computer Vision (ICCV 2003), 14-17 October 2003, Nice, France.* IEEE Computer Society.

[Jain and Farrokhnia, 1991] Jain, A. K. and Farrokhnia, F. (1991). Unsupervised texture segmentation using gabor filters. *Pattern Recognition*, 24(12):1167–1186.

[Jain, 1993] Jain, R. (1993). Nsf workshop on visual information management systems. *SIGMOD Rec.*, 22:57–75.

[Jiang et al., 2010] Jiang, W., Cotton, C. V., Chang, S.-F., Ellis, D., and Loui, A. C. (2010). Audio-visual atoms for generic video concept classification. *TOMCCAP*, 6(3).

[Jin et al., 2005] Jin, Y., Khan, L., Wang, L., and Awad, M. (2005). Image annotations by combining multiple evidence & wordnet. In [Zhang et al., 2005], pages 706–715.

[Joshi et al., 2007] Joshi, D., Naphade, M., and Natsev, A. (2007). Semantics reinforcement and fusion learning for multimedia streams. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, CIVR '07, pages 309–316, New York, NY, USA. ACM.

[Jr., 1961] Jr., R. D. F. (1961). Algorithm 32: Multint. *Commun. ACM*, 4(2):106.

[Kohrs et al., 2000] Kohrs, A., Huet, B., and Mérialdo, B. (2000). Multimedia information recommendation and filtering on the Web. In *Networking 2000, Broadband Communications, High Performance Networking, and Performance of Communication Networks, May 14-19, 2000, Paris, France*, Paris, FRANCE.

[Larson et al., 2011] Larson, M., Rae, A., Demarty, C.-H., Kofler, C., Metze, F., Troncy, R., Mezaris, V., and Jones, G. J. F., editors (2011). *Working Notes Proceedings of the MediaEval 2011 Workshop, Santa Croce in Fossabanda, Pisa, Italy, September 1-2, 2011*, volume 807 of *CEUR Workshop Proceedings*. CEUR-WS.org.

[Lazebnik et al., 2006] Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In [Forsyth, 2006], pages 2169–2178.

[Lee et al., 2006] Lee, H., Smeaton, A. F., O'Connor, N. E., and Smyth, B. (2006). User evaluation of físchlár-news: An automatic broadcast news delivery system. *ACM Trans. Inf. Syst.*, 24(2):145–189.

[Li et al., 2012] Li, G., Wang, M., Lu, Z., Hong, R., and Chua., T.-S. (2012). In-video product annotation with web information mining. *ACM trans. on Multimedia Computing, Communications, and Applications TOMCCAP*. In Press.

[Li and Wang, 2008] Li, J. and Wang, J. Z. (2008). Real-time computerized annotation of pictures. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(6):985–1002.

[Li and Sun, 2006] Li, W. and Sun, M. (2006). Automatic image annotation based on wordnet and hierarchical ensembles. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 3878 of *Lecture Notes in Computer Science*, pages 417–428. Springer Berlin / Heidelberg. 10.1007/11671299_44.

[Li et al., 2009] Li, X., Snoek, C. G. M., and Worring, M. (2009). Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia*, 11(7):1310–1322.

[Lim et al., 2003] Lim, J.-H., Tian, Q., and Mulhem, P. (2003). Home photo content modeling for personalized event-based retrieval. *IEEE MultiMedia*, 10.(4):28–37.

[Liu and Huet, 2010] Liu, X. and Huet, B. (2010). Automatic concept detector refinement for large-scale video semantic annotation. In *ICSC*, pages 97–100. IEEE.

[Liu and Huet, 2010] Liu, X. and Huet, B. (2010). Concept detector refinement using social videos. In *VLS-MCMR'10, International workshop on Very-large-scale multimedia corpus, mining and retrieval, October 29, 2010, Firenze, Italy*, Firenze, ITALY.

[Liu and Huet, 2012] Liu, X. and Huet, B. (2012). Social event modeling from web media data. In *Submitted to ICMR'12, the ACM International Conference on Multimedia Retrieval, Hong Kong.*

[Liu et al., 2011a] Liu, X., Huet, B., and Troncy, R. (2011a). Eurecom @ mediaeval 2011 social event detection task. In Larson, M., Rae, A., Demarty, C.-H., Kofler, C., Metze, F., Troncy, R., Mezaris, V., and Jones, G. J. F., editors, *MediaEval*, volume 807 of *CEUR Workshop Proceedings*. CEUR-WS.org.

[Liu et al., 2011b] Liu, X., Troncy, R., and Huet, B. (2011b). Finding media illustrating events. In [Natale et al., 2011], page 58.

[Liu et al., 2011] Liu, X., Troncy, R., and Huet, B. (2011). Using social media to identify events. In *WSM'11, 3rd ACM Multimedia Workshop on Social Media, November 18-December 1st, 2011, Scottsdale, Arizona, USA*, Scottsdale, UNITED STATES.

[Liu et al., 2008] Liu, Y., Zhang, D., and Lu, G. (2008). Region-based image retrieval with high-level semantics using decision tree learning. *Pattern Recognition*, 41(8):2554–2570.

[Lowe, 1999] Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157.

[Lu et al., 2011] Lu, Y., Sebe, N., Hytnen, R., and Tian, Q. (2011). Personalization in multimedia retrieval: A survey. *Multimedia Tools Appl.*, 51(1):247–277.

[Luan et al., 2011] Luan, H.-B., Zheng, Y.-T., Wang, M., and Chua, T.-S. (2011). Visiongo: Towards video retrieval with joint exploration of human and computer. *Inf. Sci.*, 181(19):4197–4213.

[Luo et al., 2008] Luo, J., Guan, L., Hanjalic, A., Kankanhalli, M. S., and Lee, I., editors (2008). *Proceedings of the 7th ACM International Conference on Image and Video Retrieval, CIVR 2008, Niagara Falls, Canada, July 7-9, 2008.* ACM.

[Mahalanobis, 1936] Mahalanobis, P. C. (1936). On the generalised distance in statistics. *National Institute of Sciences of India*, 2(1):49–55.

[Mangalampalli et al., 2010] Mangalampalli, A., Chaoji, V., and Sanyal, S. (2010). I-fac: Efficient fuzzy associative classifier for object classes in images. In *ICPR*, pages 4388–4391. IEEE.

[Manjunath and Ma, 1996] Manjunath, B. S. and Ma, W.-Y. (1996). Texture features for browsing and retrieval of image data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(8):837–842.

[Manjunath et al., 2001] Manjunath, B. S., Ohm, J.-R., Vasudevan, V. V., and Yamada, A. (2001). Color and texture descriptors. *IEEE Trans. Circuits Syst. Video Techn.*, 11(6):703–715.

[Manjunath et al., 2002] Manjunath, B.-S., Salembier, P., and Sikora, T. (2002). *Introduction to MPEG-7: Multimedia content description interface.* Wiley-Interscience.

[Martin et al., 2006] Martin, O., Kotsia, I., Macq, B. M., and Pitas, I. (2006). The enterface'05 audio-visual emotion database. In Barga, R. S. and Zhou, X., editors, *ICDE Workshops*, page 8. IEEE Computer Society.

[Mezaris et al., 2003] Mezaris, V., Kompatsiaris, I., and Strintzis, M. G. (2003). An ontology approach to object-based image retrieval. In *ICIP (2)*, pages 511–514.

[Mikolajczyk and Schmid, 2005] Mikolajczyk, K. and Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(10):1615–1630.

[Miller, 1995] Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

[Minka and Picard, 1996] Minka, T. P. and Picard, R. W. (1996). Interactive learning with a "society of models". In *CVPR*, pages 447–452. IEEE Computer Society.

[Mokhtarian and Mackworth, 1986] Mokhtarian, F. and Mackworth, A. K. (1986). Scale-based description and recognition of planar curves and two-dimensional shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(1):34–43.

[Moosmann et al., 2008] Moosmann, F., Nowak, E., and Jurie, F. (2008). Randomized clustering forests for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:1632–1646.

[Naphade et al., 2006] Naphade, M. R., Smith, J. R., Tesic, J., Chang, S.-F., Hsu, W. H., Kennedy, L. S., Hauptmann, A. G., and Curtis, J. (2006). Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91.

[Natale et al., 2011] Natale, F. G. B. D., Bimbo, A. D., Hanjalic, A., Manjunath, B. S., and Satoh, S., editors (2011). *Proceedings of the 1st International Conference on Multimedia Retrieval, ICMR 2011, Trento, Italy, April 18 - 20, 2011*. ACM.

[Natsev et al., 2005] Natsev, A., Naphade, M. R., and Tesic, J. (2005). Learning the semantics of multimedia queries and concepts from a small number of examples. In [Zhang et al., 2005], pages 598–607.

[Natsev et al., 2008] Natsev, A., Smith, J. R., Tesic, J., Xie, L., and Yan, R. (2008). Ibm multimedia analysis and retrieval system. In [Luo et al., 2008], pages 553–554.

[Naturel and Gros, 2008] Naturel, X. and Gros, P. (2008). Detecting repeats for video structuring. *Multimedia Tools Applications*, 38(2):233–252.

[Niblack, 1993] Niblack, W., editor (1993). *Storage and Retrieval for Image and Video Databases, 31 January - 5 February 1993, San Jose, CA, USA*, volume 1908 of *SPIE Proceedings*. SPIE.

[Paleari et al., 2009] Paleari, M., Benmokhtar, R., and Huet, B. (2009). Evidence theory-based multimodal emotion recognition. In Huet, B., Smeaton, A. F., Mayer-Patel, K., and Avrithis, Y. S., editors, *MMM*, volume 5371 of *Lecture Notes in Computer Science*, pages 435–446. Springer.

[Paleari et al., 2010a] Paleari, M., Chellali, R., and Huet, B. (2010a). Bimodal emotion recognition. In Ge, S. S., Li, H., Cabibihan, J.-J., and Tan, Y. K., editors, *ICSR*, volume 6414 of *Lecture Notes in Computer Science*, pages 305–314. Springer.

[Paleari et al., 2010b] Paleari, M., Chellali, R., and Huet, B. (2010b). Features for multimodal emotion recognition: An extensive study. In *Cybernetics and Intelligent Systems (CIS), 2010 IEEE Conference on*, pages 90 –95.

[Paleari and Huet, 2008] Paleari, M. and Huet, B. (2008). Toward emotion indexing of multimedia excerpts. In *Content-Based Multimedia Indexing, 2008. CBMI 2008. International Workshop on*, pages 425 –432.

[Paleari et al., 2009] Paleari, M., Singh, V., Huet, B., and Jain, R. (2009). Toward environment-to-environment (E2E) affective sensitive communication systems. In *MTDL'09, Proceedings of the 1st ACM International Workshop on Multimedia Technologies for Distance Learning at ACM Multimedia, October 23rd, 2009, Beijing, China*, Beijing, CHINA.

[Papadopoulos et al., 2011] Papadopoulos, S., Troncy, R., Mezaris, V., Huet, B., and Kompatsiaris, I. (2011). Social event detection at MediaEval 2011: Challenges, dataset and evaluation. In *MEDIAEVAL 2011, MediaEval Benchmarking Initiative for Multimedia Evaluation, September 1-2, 2011, Pisa, Italy*, Pisa, ITALY.

[Park et al., 2004] Park, S. B., Lee, J. W., and Kim, S.-K. (2004). Content-based image classification using a neural network. *Pattern Recognition Letters*, 25(3):287–300.

[Pentland et al., 1996] Pentland, A. P., Picard, R. W., and Sclaroff, S. (1996). Photobook: Content-based manipulation of image databases. *International Journal of Computer Vision*, 18(3):233–254.

[Picard, 1995] Picard, R. W. (1995). Light-years from lena: video and image libraries of the future. In *ICIP*, pages 310–313.

[Poullot and Satoh, 2010] Poullot, S. and Satoh, S. (2010). Detecting screen shot images within large-scale video archive. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3203 –3207.

[Pratt and Jr., 2007] Pratt, W. K. and Jr., J. E. A. (2007). Digital image processing, 4th edition. *J. Electronic Imaging*, 16(2):029901.

[Quack et al., 2007] Quack, T., Ferrari, V., Leibe, B., and Gool, L. J. V. (2007). Efficient mining of frequent and distinctive feature configurations. In *ICCV*, pages 1–8. IEEE.

[Quinlan, 1986] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106.

[Quinlan, 1996] Quinlan, J. R. (1996). Bagging, boosting, and c4.5. In Clancey, W. J. and Weld, D. S., editors, *AAAI/IAAI, Vol. 1*, pages 725–730. AAAI Press / The MIT Press.

[Robertson et al., 1994] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M., and Gatford, M. (1994). Okapi at trec-3. In *TREC*, pages 0–.

[Rosin and West, 1989] Rosin, P. L. and West, G. A. W. (1989). Segmentation of edges into lines and arcs. *Image and Vision Computing*, 7(2):109–114.

[Rubner et al., 1998] Rubner, Y., Tomasi, C., and Guibas, L. J. (1998). A metric for distributions with applications to image databases. In *ICCV*, pages 59–66.

[Rui et al., 1997] Rui, Y., Huang, T. S., and Mehrotra, S. (1997). Content-based image retrieval with relevance feedback in mars. In *ICIP (2)*, pages 815–818.

[Savarese et al., 2006] Savarese, S., Winn, J. M., and Criminisi, A. (2006). Discriminative object class models of appearance and shape by correlatons. In [Forsyth, 2006], pages 2033–2040.

[Shafer, 1976] Shafer, G. (1976). *A Mathematical Theory of Evidence.* Princeton University Press, Princeton.

[Shaw et al., 2009] Shaw, R., Troncy, R., and Hardman, L. (2009). Lode: Linking open descriptions of events. In Gómez-Pérez, A., Yu, Y., and Ding, Y., editors, *ASWC*, volume 5926 of *Lecture Notes in Computer Science*, pages 153–167. Springer.

[Shawe-Taylor and Cristianini, 2004] Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis.* Cambridge University Press.

[Shearer et al., 2001] Shearer, K., Bunke, H., and Venkatesh, S. (2001). Video indexing and similarity retrieval by largest common subgraph detection using decision trees. *Pattern Recognition*, 34(5):1075–1091.

[Shi and Malik, 2000] Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905.

[Shi et al., 2004] Shi, R., Feng, H., Chua, T.-S., and Lee, C.-H. (2004). An adaptive image content representation and segmentation approach to automatic image annotation. In [Enser et al., 2004], pages 545–554.

[Shi et al., 2005] Shi, R., Jin, W., and Chua, T.-S. (2005). A novel approach to auto image annotation based on pairwise constrained clustering and semi-naïve bayesian model. In Chen, Y.-P. P., editor, *MMM*, pages 322–327. IEEE Computer Society.

[Sikora et al., 2008] Sikora, T., Siu, W. C., Zhang, J., Guan, L., Dugelay, J.-L., Wu, Q., and Li, W., editors (2008). *International Workshop on Multimedia Signal Processing, MMSP 2008, October 8-10, 2008, Shangri-la Hotel, Cairns, Queensland, Australia.* IEEE Signal Processing Society.

[Sinha and Jain, 2008] Sinha, P. and Jain, R. (2008). Semantics in digital photos: a *c*ontenxtual analysis. *Int. J. Semantic Computing*, 2(3):311–325.

[Sivic and Zisserman, 2003] Sivic, J. and Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In [Ikeuchi et al., 2003], pages 1470–1477.

[Smeaton and Over, 2003] Smeaton, A. and Over, P. (2003). Trecvid: Benchmarking the effectiveness of information retrieval tasks on digital video. In Bakker, E., Lew, M., Huang, T., Sebe, N., and Zhou, X., editors, *Image and Video Retrieval*, volume 2728 of *Lecture Notes in Computer Science*, pages 451–456. Springer Berlin / Heidelberg.

[Smeaton and Crimmins, 1997] Smeaton, A. F. and Crimmins, F. (1997). Relevance feedback and query expansion for searching the web: A model for searching a digital library. In Peters, C. and Thanos, C., editors, *ECDL*, volume 1324 of *Lecture Notes in Computer Science*, pages 99–112. Springer.

[Smeulders et al., 2000] Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380.

[Smith and Chang, 1994] Smith, J. R. and Chang, S.-F. (1994). Transform features for texture classification and discrimination in large image databases. In *ICIP (3)*, pages 407–411.

[Smith and Chang, 1996] Smith, J. R. and Chang, S.-F. (1996). Visualseek: A fully automated content-based image query system. In Aigrain, P., Hall, W., Little, T. D. C., and Jr., V. M. B., editors, *ACM Multimedia*, pages 87–98. ACM Press.

[Snoek, 2010] Snoek, C. G. M. (2010). The mediamill search engine video. In [Bimbo et al., 2010], pages 1323–1324.

[Snoek and Worring, 2009] Snoek, C. G. M. and Worring, M. (2009). Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 2(4):215–322.

[Souvannavong et al., 2004] Souvannavong, F., Hohl, L., Méerialdo, B., and Huet, B. (2004). Using structure for video object retrieval. In *CIVR'04, International Conference on Image and Video Retrieval, July 21-23, 2004, Dublin City University, Ireland / Also published in LNCS Volume 3115/2004*, Dublin City University, IRELAND.

[Souvannavong et al., 2005] Souvannavong, F., Hohl, L., Mérialdo, B., and Huet, B. (2005). Structurally enhanced latent semantic analysis for video object retrieval. *IEE Proceedings on Vision, Image and Signal Processing*, 152(6).

[Souvannavong and Huet, 2005] Souvannavong, F. and Huet, B. (2005). Hierarchical genetic fusion of possibilities. In *EWIMT 2005, 2nd European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology, November 30-December 1st, 2005, London, UK*, London, UNITED KINGDOM.

[Souvannavong et al., 2004] Souvannavong, F., Mérialdo, B., and Huet, B. (2004). Improved video content indexing by multiple latent semantic analysis. In [Enser et al., 2004], pages 483–490.

[Souvannavong et al., 2004] Souvannavong, F., Mérialdo, B., and Huet, B. (2004). Latent semantic analysis for an effective region-based video shot retrieval system. In *6th ACM SIGMM International Workshop on Multimedia Information Retrieval, ACM Multimedia 2004, October 15-16, 2004, New York, USA*, New York, UNITED STATES.

[Srihari, 1991] Srihari, R. K. (1991). Piction: A system that uses captions to label human faces in newspaper photographs. In Dean, T. L. and McKeown, K., editors, *AAAI*, pages 80–85. AAAI Press / The MIT Press.

[Srihari and Burhans, 1994] Srihari, R. K. and Burhans, D. T. (1994). Visual semantics: Extracting visual information from text accompanying pictures. In Hayes-Roth, B. and Korf, R. E., editors, *AAAI*, pages 793–798. AAAI Press / The MIT Press.

[Sudderth et al., 2005] Sudderth, E. B., Torralba, A., Freeman, W. T., and Willsky, A. S. (2005). Learning hierarchical models of scenes, objects, and parts. In *ICCV*, pages 1331–1338. IEEE Computer Society.

[Swain and Ballard, 1990] Swain, M. J. and Ballard, D. H. (1990). Indexing via color histograms. In *ICCV*, pages 390–393. IEEE.

[Swain and Ballard, 1991] Swain, M. J. and Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision*, 7(1):11–32.

[Tieng and Boles, 1997] Tieng, Q. M. and Boles, W. (1997). Recognition of 2d object contours using the wavelet transform zero-crossing representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(8):910 –916.

[Tong and Chang, 2001] Tong, S. and Chang, E. Y. (2001). Support vector machine active learning for image retrieval. In Georganas, N. D. and Popescu-Zeletin, R., editors, *ACM Multimedia*, pages 107–118. ACM.

[Uijlings et al., 2011] Uijlings, J. R. R., de Rooij, O., Odijk, D., Smeulders, A. W. M., and Worring, M. (2011). Instant bag-of-words served on a laptop. In [Natale et al., 2011], page 69.

[Vailaya et al., 2001] Vailaya, A., Figueiredo, M. A. T., Jain, A. K., and Zhang, H. (2001). Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10(1):117–130.

[Vapnik and Chapelle, 2000] Vapnik, V. and Chapelle, O. (2000). Bounds on error expectation for support vector machines. *Neural Computation*, 12(9):2013–2036.

[Velasco and Marroquín, 2003] Velasco, F. A. and Marroquín, J. L. (2003). Growing snakes: active contours for complex topologies. *Pattern Recognition*, 36(2):475–482.

[Wang et al., 2006] Wang, C., Jing, F., Zhang, L., and Zhang, H. (2006). Image annotation refinement using random walk with restarts. In Nahrstedt, K., Turk, M., Rui, Y., Klas, W., and Mayer-Patel, K., editors, *ACM Multimedia*, pages 647–650. ACM.

[Wang et al., 2001] Wang, J. Z., Li, J., and Wiederhold, G. (2001). Simplicity: Semantics-sensitive integrated matching for picture libraries. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(9):947–963.

[Wang et al., 2010] Wang, M., Sebe, N., Mei, T., Li, J., and Aizawa, K. (2010). Large-scale image and video search: Challenges, technologies, and trends. *J. Visual Communication and Image Representation*, 21(8):771–772.

[Wolfson and Rigoutsos, 1997] Wolfson, H. and Rigoutsos, I. (1997). Geometric hashing: an overview. *Computational Science Engineering, IEEE*, 4(4):10 –21.

[Wong and Leung, 2008] Wong, R. C. F. and Leung, C. H. C. (2008). Automatic semantic annotation of real-world web images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1933–1944.

[Yan et al., 2003] Yan, R., Yang, J., and Hauptmann, A. G. (2003). Automatically labeling video data using multi-class active learning. In [Ikeuchi et al., 2003], pages 516–523.

[Yin and Han, 2003] Yin, X. and Han, J. (2003). Cpar: Classification based on predictive association rules. In Barbará, D. and Kamath, C., editors, *SDM*. SIAM.

[Zhang et al., 2012] Zhang, D., Islam, M. M., and Lu, G. (2012). A review on automatic image annotation techniques. *Pattern Recognition*, 45(1):346–362.

[Zhang et al., 2005] Zhang, H., Chua, T.-S., Steinmetz, R., Kankanhalli, M. S., and Wilcox, L., editors (2005). *Proceedings of the 13th ACM International Conference on Multimedia, Singapore, November 6-11, 2005.* ACM.

[Zhao et al., 2007] Zhao, W.-L., Ngo, C.-W., Tan, H.-K., and Wu, X. (2007). Near-duplicate keyframe identification with interest point matching and pattern learning. *Multimedia, IEEE Transactions on*, 9(5):1037 –1048.

[Zhao et al., 2008] Zhao, Y., Zhao, Y., Zhu, Z., and Pan, J.-S. (2008). A novel image annotation scheme based on neural network. In Pan, J.-S., Abraham, A., and Chang, C.-C., editors, *ISDA (3)*, pages 644–647. IEEE Computer Society.

[zhong Lan et al., 2012] zhong Lan, Z., Bao, L., Yu, S.-I., Liu, W., and Hauptmann, A. G. (2012). Double fusion for multimedia event detection. In Schoeffmann, K., Mérialdo, B., Hauptmann, A. G., Ngo, C.-W., Andreopoulos, Y., and Breiteneder, C., editors, *MMM*, volume 7131 of *Lecture Notes in Computer Science*, pages 173–185. Springer.

[Zhou et al., 2001] Zhou, F., Feng, J. F., and Shi, Q. Y. (2001). Texture feature based on local fourier transform. In *ICIP (2)*, pages 610–613.

[Zhou and Huang, 2003] Zhou, X. S. and Huang, T. S. (2003). Relevance feedback in image retrieval: A comprehensive review. *Multimedia Systems*, 8:536–544. 10.1007/s00530-002-0070-3.

4 Selected Publications

In order to provide the reader a more in-depth view of some of the research topics presented earlier, here is a selection of relevant publications:

- Eric Galmar, Thanos Athanasiadis, Benoit Huet, Yannis Avrithis, "Spatiotemporal semantic video segmentation", MMSP 2008, 10th IEEE International Workshop on MultiMedia Signal Processing, October 8-10, 2008, Cairns, Queensland, Australia

- Marco Paleari and Benoit Huet, "Toward emotion indexing of multimedia excerpts", CBMI 2008, 6th International Workshop on Content Based Multimedia Indexing, June, 18-20th 2008, London, UK
  Best student paper award

- Rachid Benmokhtar, Benoit Huet, "An ontology-based evidential framework for video indexing using high-level multimodal fusion", Multimedia Tools and Application, Springer, December 2011

- Xueliang Liu, Benoit Huet, "Concept detector refinement using social videos", VLS-MCMR'10, ACM Multimedia International workshop on Very-large-scale multimedia corpus, mining and retrieval, October 29, 2010, Firenze, Italy , pp 19-24

- Xueliang Liu, Raphaël Troncy, Benoit Huet, "Finding media illustrating events", ICMR'11, 1st ACM International Conference on Multimedia Retrieval, April 17-20, 2011, Trento, Italy

Spatiotemporal Semantic Video Segmentation

E. Galmar [*1], Th. Athanasiadis [†3], B.Huet [*2], Y. Avrithis [†4]

[*] *Département Multimédia, Eurécom, Sophia-Antipolis, France*
[1] galmar@eurecom.fr [3] huet@eurecom.fr

[†] *Image, Video & Multimedia Systems Laboratory, NTUA, Greece*
[2] thanos@image.ntua.gr [4] iavr@image.ntua.gr

Abstract—**In this paper, we propose a framework to extend semantic labeling of images to video shot sequences and achieve efficient and semantic-aware spatiotemporal video segmentation. This task faces two major challenges, namely the temporal variations within a video sequence which affect image segmentation and labeling, and the computational cost of region labeling. Guided by these limitations, we design a method where spatiotemporal segmentation and object labeling are coupled to achieve semantic annotation of video shots. An internal graph structure that describes both visual and semantic properties of image and video regions is adopted. The process of spatiotemporal semantic segmentation is subdivided in two stages: Firstly, the video shot is split into small block of frames. Spatiotemporal regions (volumes) are extracted and labeled individually within each block. Then, we iteratively merge consecutive blocks by a matching procedure which considers both semantic and visual properties. Results on real video sequences show the potential of our approach.**

I. Introduction

The development of video databases has impelled research for structuring multimedia content. Traditionally, low-level descriptions are provided by image and video segmentation techniques. The best segmentation is achieved by the human eye, performing simultaneously segmentation and recognition of the object thanks to a strong prior knowledge about the objects' structures. To generate similar high-level descriptions, a knowledge representation should be used in computer-based systems. One of the challenges is to map efficiently the low-level descriptions with the knowledge representation to improve both segmentation and interpretation of the scene.

We propose to associate spatiotemporal segmentation and semantic labeling techniques for joint segmentation and annotation of video shots. From one hand, semantic labeling brings information from a domain of knowledge and enables recognition of materials and concepts related to the objects. From the other hand, spatiotemporal segmentation decomposes a video shot into continuous volumes that are homogeneous with respect to a set of features. These extracted volumes represent an efficient medium to propagate semantic labels inside the shot.

Various approaches have been proposed for segmenting video shot into volumes. 3D approaches take as input the whole set of frames and give coherent volumes optimizing a global criterion [1], at the expense of an important computational cost. A few methods provide mid-level description of the volumes. In [2], volumes are modeled by a gaussian mixture model including color and position. Another example is given in [3], where volumes are considered as small moving linear patches. We have previously demonstrated that with a 2D+T (time) method [4] we can obtain a good trade-off between efficiency and accuracy of the extracted volumes. Recent progress has been also observed for scene interpretation and the labeling of image regions. In [5], an experimental platform is described for semantic region annotation. Integration of bottom-up and top-down approaches in [6] provides superior results in image segmentation and object detection. Region growing techniques have been adapted to group low-level regions using their semantic description instead of their visual features [7].

The integration of semantic information within the spatiotemporal grouping process sets two major challenges. Firstly, region labeling is obtained by computing visual features and match them to the database, which induces an important computational cost. Secondly, the relevance of the semantic description depends also on the accuracy of visual descriptors, whose extraction requires sufficient area of the volumes. These considerations suggest that use of semantic information during the early stages of the segmentation algorithm would be highly inefficient and ineffective if not misleading. Therefore, we add semantic information when the segmentation has produced a relatively small number of volumes. To this aim, we introduce a method to group semantically spatiotemporal regions within video shots.

The paper is organized as follows: In section II we give an overview of the strategy. Section III introduces the graph representation used for video shots. Section IV and V details the building steps of our approach: the labeling of temporal volumes and its propagation to the whole shot, respectively. Finally, results are illustrated in section VI and conclusions are drawn in section VII.

II. Overview of the strategy

The overall framework for the application is shown in fig.1. The considered video sequences are restricted to single shots, i.e. video data has been captured continuously from the camera and there are no cuts. Because of occlusion, shadowing, viewpoint change or camera motion, object material is prone to important spatial and temporal variations that makes maintaining an object as a unique volume difficult. To overcome the limits of the spatiotemporal stage, a video shot is decomposed into a sequence of smaller Block of Frames (BOF).

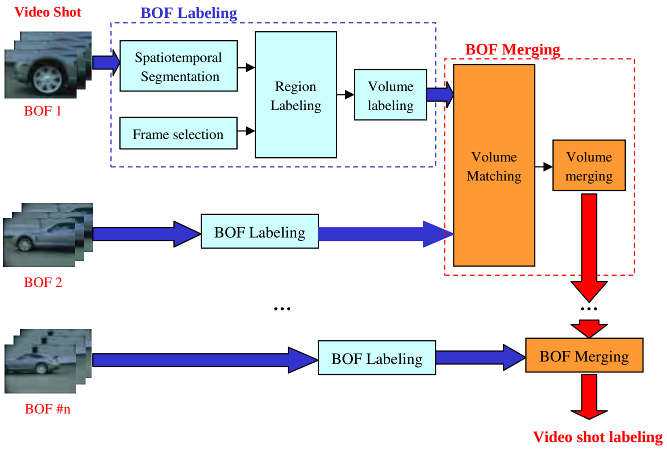


Fig. 1. The proposed framework for semantic video segmentation.

Semantic video shot segmentation is then achieved by an iterative procedure on the BOFs and operates in two steps, labeling of volumes within the BOF and merging with the previous BOF, which we will refer to as *intra-BOF* and *inter-BOF* processing respectively. During intra-BOF processing, spatiotemporal segmentation decomposes each BOF into a set of volumes. The resulting 2D+T segmentation map is sampled temporally to obtain several frame segmentation maps, each one consisting of a number of non overlapping regions. These regions are semantically labeled and the result is propagated within the volumes. A semantic region growing algorithm is further applied to group adjacent volumes with strong semantic similarity. During inter-BOF processing, we perform joint propagation and re-estimation of the semantic labels between consecutive video segments. The volumes within each BOF are matched by means of their semantic labels and visual features. This allows to extend the volumes through the whole sequence and not just within a short BOF. The semantic labels of the matched volumes are re-evaluated and changes are propagated within each segment. Finally both BOFs are merged and the process is repeated on the next BOF.

III. Graph Representation of Video Shots

Following MPEG-7 descriptions, one video shot is structured hierarchically in video segments. Firstly a shot is divided into M Blocks of Frames (BOF) B_i ($i \in [1, M]$), each one composed of successive frames F_t, $t \in [1, |B_i|]$. Spatiotemporal segmentation decomposes each B_i into a set of video regions (or volumes) S_{B_i}. Each volume $a \in S_{B_i}$ is subdivided temporally into frame regions $R_a(t)$, $F_t \in B_i$. Finally, frame segmentation at time t is defined as the union of frame regions of all volumes intersecting frame F_t: $S_t = \bigcup_{a \cap F_t \neq \emptyset} R_a(t)$. The elements composing the BOF are represented in fig.2.

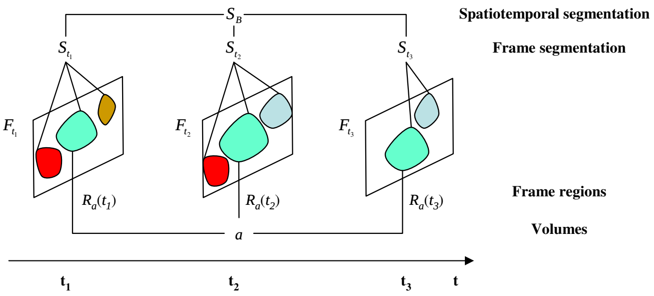


Fig. 2. Spatial and temporal decomposition of a BOF B_i.

A video segment (image or video shot) can represent a structured set of objects and is naturally described by an Attributed Relational Graph (ARG) [8]. Formally, an ARG is defined by spatiotemporal entities represented as a set of vertices V and binary spatiotemporal relationships represented as a set of edges E: $ARG \equiv \langle V, E \rangle$. Letting S_{B_i} be a segmentation of a BOF B_i, a volume $a \in S_{B_i}$ is represented in the graph by vertex $\mathrm{v}_a \in V$, where $\mathrm{v}_a \equiv \langle a, \mathcal{D}_a, \mathcal{L}_a \rangle$. $\mathcal{D}_a$ is a set of low-level MPEG-7 visual descriptors for volume a, while $\mathcal{L}_a$ is the fuzzy set of labels for that volume (defined over the crisp set of concepts C) with membership function μ_a:

$$\mathcal{L}_a = \sum_{i=1}^{|C|} c_i / \mu_a(c_i), \quad c_i \in C \tag{1}$$

Two neighbor volumes $a, b \in S_{B_i}$ are related by a graph edge $e_{ab} \equiv \langle (\mathrm{v}_a, \mathrm{v}_b), s^{\mathcal{D}}_{ab}, s^{\mathcal{L}}_{ab} \rangle$. $s^{\mathcal{D}}_{ab}$ is the visual similarity of volumes a and b, calculated from their set of MPEG-7 descriptors $\mathcal{D}_a$ and $\mathcal{D}_b$.. Several distance functions are used for each descriptor, so we normalize those distances linearly to the unit range and compute their visual similarity $s^{\mathcal{D}}_{ab}$ by their linear combination. $s^{\mathcal{L}}_{ab}$ is a semantic similarity value based on the fuzzy set of labels of the two volumes $\mathcal{L}_a$ and $\mathcal{L}_b$:

$$s^{\mathcal{L}}_{ab} = \sup_{c_i \in C} \left(t(\mathcal{L}_a, \mathcal{L}_b) \right), \quad a \in S, b \in N_a \tag{2}$$

where N_a is the set of neighbor volumes of a and t is a t-norm of two fuzzy sets. Intuitively, eq.2 states that the semantic similarity $s^{\mathcal{L}}_{ab}$ is the highest degree, implied by our knowledge, that volumes a and b share the same concept.

IV. Intra-BOF Labeling

To label a new BOF, we exploit the spatiotemporal segmentation to build visual and semantic description efficiently, using only a few frames. The following subsections present the criterion used for selecting these frames, the extraction of visual and semantic attributes of video regions and how those attributes are used for merging operations of volumes within the BOF.

A. Frame Selection

Once the segmentation masks are obtained for the whole BOF, region descriptor extraction and labeling tasks are substantially reduced by selecting a set of frames within the video segment. Choosing an important number of frames will lead to a complete description of the BOF but will require more time to process. On the contrary, using a single frame is more efficient but important volumes may not receive labels.

We consider a set of frames T and its corresponding frame segmentations $S_T = \{S_t\}$, $t \in T$ and measure the total span of the intersected volumes. Given a fixed size for T we choose the set T_{sel} that maximizes the span of the labeled volumes:

$$T_{sel} = \underset{T}{\operatorname{argmax}} \sum_{a \cap S_T \neq \emptyset} |a| \tag{3}$$

where $|a|$ is the size of volume a. Compared with fixed sampling, the criterion offers scalability for the extracted descriptors in function of the desired total volume span for the shot. Indeed the span increases with the number of frames selected.

B. Video Region Description

In previous work [5] we have shown how extracted visual descriptors can be matched to visual models of concepts. This region labeling process is applied to the selected frames (according to criteria discussed in section IV-A), resulting to an initial fuzzy labeling of regions with a set of concepts. The fuzzy set of labels $\mathcal{L}_a$ of a volume a is obtained by gathering the contributions from each frame region using a fuzzy aggregation operator :

$$\mu_a(c) = \frac{\sum_{t \in T_{sel}} \mathcal{A}(R_a(t)) \mu_{R_a(t)}(c)}{\sum_{t \in T_{sel}} \mathcal{A}(R_a(t))} \tag{4}$$

This operator weights the confidence degrees with the importance given to the frame regions. These weights $\mathcal{A}(R_a(t))$, are obtained by a measure of temporal consistency of frame regions.

Besides the semantic labeling, volumes are also described by low-level visual descriptors. Most MPEG-7 descriptors are originally intended for frame regions, but can be extended to volumes with the use of aggregation operators. For histogram-based descriptors, common operators are *mean*, *median* and *intersection* of bins. We select the *mean* operator since we consider homogeneous short-length volumes. In addition to descriptors, we also store the sizes and center of the volumes and its spatiotemporal bounding box for fast localization.

C. Semantic Volume Growing

Spatiotemporal segmentation usually creates more volumes than the actual number of objects present in the BOF. We examine how a variation of a traditional segmentation technique, the Recursive Shortest Spanning Tree (RSST) can be used to create more coherent volumes within a BOF. The idea is that neighbor volumes, sharing the same concepts, as expressed by the labels assigned to them, should be merged, since they define a single object.

To this aim, we modify the RSST algorithm to operate on the fuzzy sets of labels $\mathcal{L}$ of the volumes in a similar way as if it worked on low-level features (such as color, texture) [7]. The modification of the traditional algorithm to its semantic equivalent lies on the re-definition of the two criteria: (i) The similarity between two neighbor volumes a and b (vertices v_a and v_b in the graph), based on which graph's edges are sorted and (ii) the termination criterion. For the calculation of the semantic similarity between two vertices, we use $s_{ab}^{\mathcal{L}}$ defined in eq.2.

For one iteration of the semantic RSST, the process of volume merging decomposes in the following steps: Firstly, the edge e_{ab} that has the maximum semantic similarity $s_{ab}^{\mathcal{L}}$ is selected; vertices v_a and v_b are merged. Vertex v_b is removed completely from the ARG, whereas v_a is updated appropriately. This update procedure consists of two actions:

- Re-evaluation of the degrees of membership of the labels in a weighted average fashion from the union of the two volumes:

$$\mu_a(c) \leftarrow \frac{|a|\mu_a(c) + |b|\mu_b(c)}{|a| + |b|} \tag{5}$$

- Re-adjustment of the ARG edges by removing edge e_{ab} and re-evaluating the weights of the affected edges incident to a or b.

This procedure terminates when the edge e^* with maximum semantic similarity in the ARG is lower than a threshold, which is calculated in the beginning of the algorithm, based on the histogram of all semantic similarity values of the set of all edges E.

V. Inter-BOF Processing

In the previous section we dealt with segmentation and labeling of volumes within each single BOF. Here we examine how to extend volumes over consecutive BOF and for this purpose we develop techniques of visual and semantic volume matching. Semantic grouping is first performed on volumes with dominant concepts (i.e. concepts with high degree of confidence), then concepts are propagated temporally and spatially with the use of both semantic and visual similarity.

A. BOF Matching

We consider the merging of two successive BOF represented by their ARGs G_1 and G_2. It is not worth computing all volume matches between the two ARGs. As we consider continuous sequences, semantic objects are coherent spatially and temporally. In consequence, numerous matches can be pruned by exploiting spatiotemporal location of the volumes.

We establish temporal connections between G_1 and G_2 by selecting candidate matches from G_1 to G_2 and G_2 to G_1. Let G be the merged graph of G_1 and G_2. At the beginning, $G = G_1 \cup G_2$. Given vertices $\mathrm{v}_a \in G_1$ and $\mathrm{v}_b \in G_2$, v_a is connected to v_b in G if the bounding box of b intersects a truncated pyramid that represents the possible locations for a in the new BOF. The pyramid top base is defined by the bounding box of a. The bottom base is enlarged by a factor $D_s = v_{max}T_{max}$ where v_{max} is the maximum displacement between two frames and T_{max} is the height of the pyramid along the temporal axis. The connections are established in both forward and backward temporal directions. As a result, v_a owns an edge list of candidate matches $E_a = \{e_{ab} | \mathrm{v}_b \in G_2\}$. A list E_b is created similarly for v_b.

After creating the list of candidate matches, we match volumes with reliable or *dominant* concepts. A concept $c^* \in C$

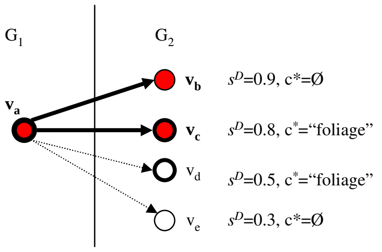


Fig. 3. Matching of dominant volumes. Dominant volumes are represented with thick circles.

is considered *dominant* for a volume $a \in G$ if the following condition is satisfied:

$$\left\{ \begin{array}{lll} \mu_a(c_1) & > & T_{dom} \\ \mu_a(c_1) & > & T_{sec}\mu_a(c_2) \end{array} \right. \tag{6}$$

c_1 and c_2 are respectively the concepts with highest and second highest degrees of membership. A dominant concept has degree of memberships above T_{dom} and is more important than all other concepts, with minimum ratio of T_{sec}.

The best match for one dominant volume may not be dominant because its visual appearance changes during the sequence. For this reason, we match either dominant volumes that have sufficient visual similarity or one dominant volume to any volume in case they have perfect visual match. The criterion to match a dominant volume a to a volume b, $e_{ab} \in E_a$, is based on both semantic and visual attributes. Let c_a^* and c_b^* be the dominant concepts of $\mathcal{L}_a$ and $\mathcal{L}_b$. If b is dominant but $c_a^* \neq c_b^*$, then no matching is done. In case c_b^* is empty, then e_{ab} has to be the best visual match from a, otherwise we compute the normalized rank of the visual similarity $s^{\mathcal{D}}$ in decreasing order, whose values do not depend of the descriptors used. Formally the criterion is validated if:

$$\left\{ \begin{array}{ll} rank\left(s_{ab}^{\mathcal{D}}\right) = 1 & \text{if} \quad c_b^* = \emptyset \\ \left\{ \begin{array}{l} c_a^* = c_b^* \\ \frac{|E_a| - rank\left(s_{ab}^{\mathcal{D}}\right)}{|E_a| - 1} > T_s \end{array} \right. & \text{otherwise} \end{array} \right. \tag{7}$$

T_s indicates the tolerance allowed on visual attributes. When T_s is close to 1, only the best visual match is considered. If T_s is set to 0.5, half of the matches are kept.

The aforementioned procedure is illustrated fig.3. In the example, v_a is linked to v_c as it shares the same concept "foliage" and the visual similarity is the second best ($s_{ac}^{\mathcal{D}} = 0.8$). v_a is also linked to v_b since the similarity between a and b is the best one ($s_{ab}^{\mathcal{D}} = 0.9$). v_d is not matched even if it shares the same dominant concept as they are visually different from v_a. Indeed only dominant matches with good similarity are kept.

Since region and volume labeling are processes with a certain degree of uncertainty, reliable semantic concepts do not emerge from every volume, either due to the limited domain of the knowledge base, the imperfections of the segmentation, or the material itself. Therefore, we introduce volume matching using low-level visual attributes, expecting the semantics of these volumes to be recognized with more certainty in a subsequent part of the sequence. To avoid propagating matching errors and hamper the accuracy of the volumes, we only consider the matches with the strongest similarities and we are most confident in. Let e_a^* and e_b^* be the edges in lists E_a and E_b which have maximum visual similarity. a and b are matched and e_{ab} is a first best match, i.e. $e_{ab} \equiv e_a^* \equiv e_b^*$.

B. Update and Propagation of Labels

After the matching process, volumes are merged and their semantic and visual properties are computed using the aggregation operators, defined eq. 4. For this reason, new evidence for semantic similarity can be found in the merged graph as new dominant volumes are likely to be found. We do not merge further these volumes at this stage, so as to keep the accuracy of the visual description as they may correspond to different materials belonging to the same concept. Instead of this, the concepts of dominant volumes are propagated in the merged graph G. Let a be a non-dominant volume, $\mathrm{v}_a \in G$; we define a set of candidate dominant concepts $C_a = \{c \in C | \mu_a(c) > T_c\}$. For a concept $c \in C_a$, we compute the degrees of membership $\mu_a'(c)$ resulting from the aggregation of v_a and its neighbor vertices in G with dominant concept c:

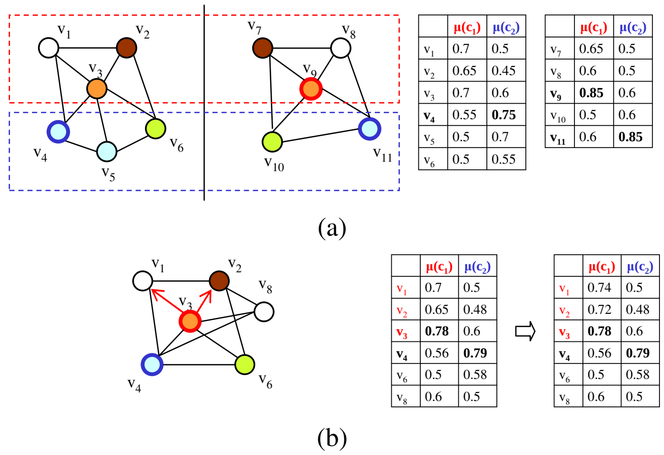
$$\mu_a'(c) = \frac{\sum_{b \in N_a^c} |b| \mu_b(c)}{\sum_{b \in N_a^c} |b|} \tag{8}$$

where $N_a^c = a \cup \{b \in N_a | c_b^* = c\}$ is the aforementioned neighborhood and $|b|$ is the current size of volume b. The concept $c^* \in C_a$, maximizing $\mu_a'(c)$, is selected and all degrees of membership of $\mathcal{L}_a$ and the size $|a|$ are updated by the aggregation of volumes in $N_a^{c^*}$. This propagation is performed in the whole graph G recursively. Let G^D be the subgraph of G containing only the dominant volumes of G and their incident edges. Once non-dominant volumes in G are processed, new dominant volumes may emerge in the subgraph $G' = G - G^D$. The update procedure is repeated considering G' as the whole graph until no more dominant volumes are found: $G^D = \emptyset$. Consequently, degrees of membership of non-dominant volumes tend to increase using the neighborhood context, correcting the values from the initial labeling.

Fig.4 gives an example of the inter-BOF merging and propagation of labels after that. The ideal semantic segmentation would be composed of two objects with dominant concepts c_1 and c_2. Before merging, a few dominant volumes are detected (v_4, v_9, v_{11}) in the two BOFs. After merging (fig.4(b)) the degrees of membership are re-evaluated according to eq. 5 and semantic weights are computed on the new edges. New evidence for semantic similarity is found between volumes (v_3, v_1) and (v_3, v_2), since v_3 has been matched with dominant volume v_9. Thus, due to propagation of concept c_1, v_1 and v_2 are linked to the dominant volume v_3 and their degrees of membership are increased according to eq. 8.

VI. Experimental Results

We illustrate the potential of the method on a set of examples. The knowledge domain encompasses various elements encountered in a natural scene, such as "sea", "sky", "foliage"



	μ(c1)	μ(c2)
v_1	0.7	0.5
v_2	0.65	0.45
v_3	0.7	0.6
v_4	0.55	**0.75**
v_5	0.5	0.7
v_6	0.5	0.55

	μ(c1)	μ(c2)
v_7	0.65	0.5
v_8	0.6	0.5
v_9	**0.85**	0.6
v_{10}	0.5	0.6
v_{11}	0.6	**0.85**

	μ(c1)	μ(c2)
v_1	0.7	0.5
v_2	0.65	0.48
v_3	**0.78**	0.6
v_4	0.56	**0.79**
v_6	0.5	0.58
v_8	0.6	0.5

⇨

	μ(c1)	μ(c2)
v_1	0.74	0.5
v_2	0.72	0.48
v_3	**0.78**	0.6
v_4	0.56	**0.79**
v_6	0.5	0.58
v_8	0.6	0.5

Fig. 4. Merging of two BOFs. (a) Matching between two BOF. (b) Merging of a BOF and update of semantic labels. Ideal semantic segmentation is represented by the dashed boxes. Matched volumes are marked with similar colors, and dominant volumes are indicated with thick circles. Here, $T_{dom} = 0.75$ and $T_{sec} = 1.25$.

or "person". The proposed example sequences are composed of 650 and 100 frames, respectively. The BOF duration in the second sequence is $|B| = 10$ frames while for the first sequence we increase the duration to $|B| = 50$ frames, to show the behavior of the method at a larger scale while maintaining reduced computational costs.

The first example shows two girls walking on the beach (fig.5). Firstly, the girls are approaching the camera (BOF 1-5). Then they are observed in a close-up view (BOF 6-10). Finally the camera rotates quickly by 180 degrees to shoot them backside. Relevant concepts "person", "sky" and "sea" are detected within the shot. First we can see that the sky area is recognized all along the sequence. Although its aspect slightly changes at the end, it is still detected as dominant in the labeling stage and thus merged as a single volume. We can notice that isolated areas are also labeled "sky", as their material is visually close to this concept (BOF 5, 13). For the same reason, only part of the sea is identified at the right. In contrast, the left part is not dominant, but is correctly grouped by visual matching from BOF 3 to 10. After that, the sea areas are detected easily being shot in front view. The detection of "person" is more challenging since the related object includes different materials. In BOF 1 each silhouette is identified correctly standing as a single volume. The left girl's area is propagated from BOF 3 to 10. After that point it is completely occluded in BOF 11 and the concept is re-detected within a new volume in BOF 13. For the girl on the right the labeling is more uncertain as part of her suit and head have been confused with the background area (BOF 5, 7, 11). However, the upper part is still detected and propagated from BOF 5 to 9 and from 10 to 12 while the view is changing.

The second example shows a woman talking in front of her car (fig.6). The detected concepts include "person" and "foliage". The head and the coat both belong to the "person" concept and can be viewed as a single object, but are still separated in the semantic segmentation (fig.6(c)), which is an advantage as they are visually different. In BOF 4 only the coat is recognized (fig.6(d)). The reason is that the head has been partly confused with the background in the spatiotemporal segmentation. In such case, the volume is not matched, as its visual properties are different from the other volumes in the previous and subsequent BOF. In the right part of the sequence, the upper branches are well identified as "foliage" and are merged in a single volume from BOF 1 to 4 (fig.6(c)). From BOF 6 to 8, the branches are occluded by the woman. As a consequence the volumes are more fragmented and less homogeneous, so they are not linked to the previous part of the sequence. In BOF 10, the volume material in this area is homogeneous and the branches are correctly identified again.

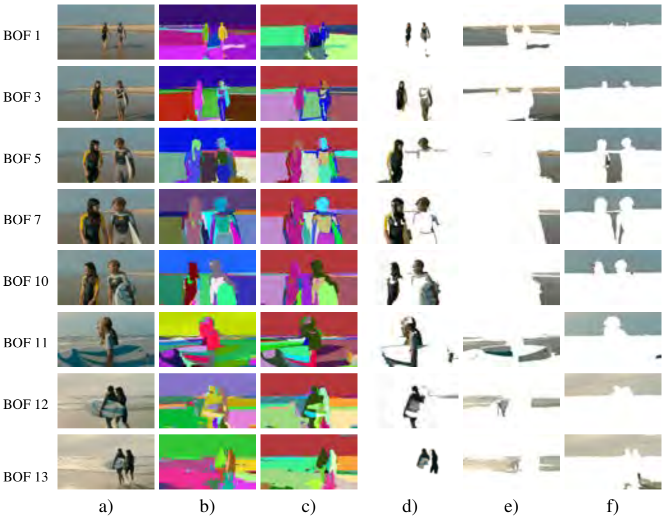


Fig. 5. Video semantic segmentation. (a) Frames in various BOF. (b) Spatiotemporal segmentation. (c) Semantic segmentation and inter-BOF matching. Volumes are extended throughout the shot (note the consistency in coloring). (d) Concept "person". (e) "sea". (f) "sky".

		Foliage	**Person**	**Sea**	**Sky**	**Overall**
Ex.1	*Acc*	x	0.74	0.87	0.96	0.89
	Score	x	0.62	0.65	0.78	0.71
Ex.2	*Acc*	0.81	0.86	x	x	0.84
	Score	0.55	0.64	x	x	0.61

TABLE I
EVALUATION OF THE SEGMENTATION RESULTS.

Evaluation of the results for the above sequences is presented in table I. Each concept is associated to a semantic object (ground truth). The accuracy measure (Acc) [9] relates to the quality of the segmented volumes (fig.5-6(c)), unifying precision and recall. The evaluation score [7] gives a further measure of belief for the object labeling in every image. Unsurprisingly concept "sky" obtains the best result for all measures. For "foliage" sparse texture of the material and fragmentation of the volumes result in a lower score of 0.55. Concept "sea" has a higher detection score of 0.65, color and texture being relatively stable. Concept "person" is detected although some background can be included in the object

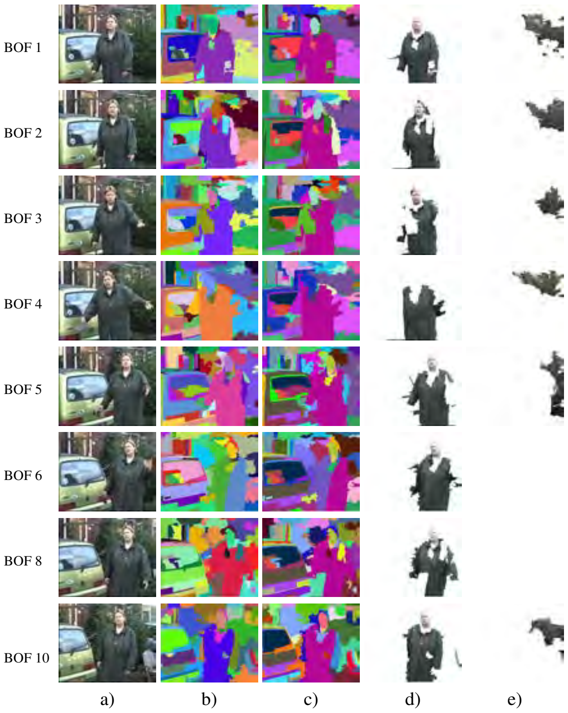


Fig. 6. Video semantic segmentation. (a) Frames in various BOF. (b) Spatiotemporal segmentation. (c) Semantic segmentation and inter-BOF matching. (d) Concept "person". (e) "foliage".

($Acc = 0.74$ for Ex.1). Finally, the overall detection scores are 0.71 for Ex.1 and 0.61 for Ex.2.

We further analyze the effects of the BOF decomposition on the efficiency of the approach. Fig.7 shows the repartition of the overall running time for the sequence of the first example (650 frames). The procedure is composed of four steps: (i) spatiotemporal segmentation, (ii) visual descriptor extraction and region labeling with the knowledg-assisted analysis system (KAA [5]), (iii) the construction of the ARGs (including the semantic RSST) and (iv) the inter-processing stage that merges the BOFs. Processing frames independently ($|B| = 1$) generates an important computational cost because of the labeling of every image of the sequence. The impact on the overall complexity is reduced with the spatiotemporal scheme ($|B| > 1$) that allows temporal sampling of the frames. For the evaluation, a single frame has been selected for each block, so that running time decreases inversely with the BOF length. Regarding the other components, we can notice that large BOF sizes lead to increase the time required for producing the spatiotemporal segmentation of the BOF. However, the additional cost is largely compensated with the gain in the region labeling stage. For the final merging stage, the running time for different BOF sizes is comparable. Indeed, the step is dominated by loading and updating the frame segmentation maps of which number does not depend of the BOF size, while the merging of the ARGs has lower complexity.

Overall, the gain with the proposed approach reaches a factor up to 12 ($|B| = 50$). Thus, the analysis shows the benefit of the framework in terms of complexity, extending single image annotation to continuous sequences efficiently.

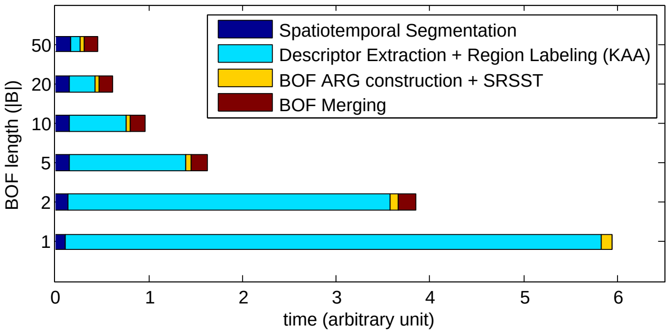


Fig. 7. Repartition of the overall running time of the first example, in function of the BOF length. Complexity is reduced when the BOF length increases.

VII. Conclusions

This paper presents a new approach for simultaneous segmentation and labeling of video sequences. Spatiotemporal segmentation is presented as an efficient solution to alleviate the cost of region labeling, compensating semantic with visual information when the former is missing. Our approach groups volumes with relevant concepts together while maintaining a spatiotemporal segmentation for the entire sequence. This enables the segmented volumes to be annotated at a subsequent point in the sequence. First experiments on real sequences show that the application is promising, though enhancements can still be achieved in the early spatiotemporal segmentation and labeling stage. Further challenge will be to consider structured objects instead of materials, leading towards scene interpretation and detection of complex events.

Acknowledgment

This research was supported by the European Commission under Contract FP6027026-K-Space.

References

[1] H.-Y. S. Y. Li, J. Sun, "Video object cut and paste," in *SIGGRAPH*, 2005.

[2] H. Greenspan, J. Goldberger, and A. Mayer, "Probabilistic space-time video modeling via piecewise gmm," *IEEE Trans. PAMI*, vol. 26, no. 3, pp. 384–396, Mar. 2004.

[3] D. DeMenthon and D. Doermann, "Video retrieval using spatio-temporal descriptors," in *ACM MM*, 2003, pp. 508–517.

[4] E. Galmar and B. Huet, "Graph-based spatio-temporal region extraction," in *ICIAR*, 2006, pp. 236–247.

[5] T. Athanasiadis, V. Tzouvaras, V. Petridis, F. Precioso, Y. Avrithis, and Y. Kompatsiaris, "Using a multimedia ontology infrastructure for semantic annotation of multimedia content," in *5th Int'l Workshop on Knowledge Markup and Semantic Annotation*, 2005.

[6] S. U. Eran Borenstein, Eitan Sharon, "Combining top-down and bottom-up segmentation," in *8th Conference on Computer Vision and Pattern Recognition Workshop*, 2004.

[7] T. Athanasiadis, P. Mylonas, Y. Avrithis, and S. Kollias, "Semantic image segmentation and object labeling," *IEEE Trans. Circuits ans Systems for Video Technology*, vol. 17, March 2007.

[8] S. Berretti, A. D. Bimbo, and E. Vicario, "Efficient matching and indexing of graph models in content-based retrieval," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 11, no. 12, pp. 1089–1105, Dec. 2001.

[9] F. Ge, S. Wang, and T. Liu, "Image-segmentation evaluation from the perspective of salient object extraction," in *CVPR*, 2006, pp. 1146–1153.

TOWARD EMOTION INDEXING OF MULTIMEDIA EXCERPTS

Marco Paleari, Benoit Huet

Eurecom Institute
Multimedia Department
2229, route des cretes, Sophia Antipolis, France

ABSTRACT

Multimedia indexing is about developing techniques allowing people to effectively find media. Content-based methods become necessary when dealing with large databases. Current technology allows exploring the emotional space which is known to carry very interesting semantic information. In this paper we state the need for an integrated method which extracts reliable affective information and attaches this semantic information to the medium itself. We describe SAMMI, a framework explicitly designed to fulfill this need and we present a list of possible applications pointing out the advantages that the emotional information can bring about. Finally, different scenarios are considered for the recognition of the emotions which involve different modalities, feature sets, fusion algorithms, and result optimization methods such as temporal averaging or thresholding.

1. INTRODUCTION

The increasing number of media publicly available on the web encourage the search for effective indexing and retrieval systems.

Current technologies use metadata which are extracted from the text surrounding the media itself supposing a tight link between these two elements. Unfortunately, this is not always the case and the text surrounding a piece of media is often other than its description, furthermore, it is rarely accurate or as complete as we would like.

In recent years, academia has been developing automatic content based methods to extract information about media excerpts. Audio and video are being analyzed to extract both low level features, such as tempo, texture, or color, and abstracted attributes, e.g. person (in an image), genre, and others.

It is well known that in most medias, in most human communications forms, and notably in art expressions, emotions represent a non-negligible source of information.

Even though studies from this community [1] acknowledge that emotions are an important characteristic of media and that they might be used in many interesting ways as semantic tags, only few efforts [2, 3, 4, 5, 6] have been done to link emotions to content-based indexing and retrieval of multimedia.

Salway and Miyamory [2, 3] analyze the text associated to a film searching for occurrences of emotionally meaningful terms; Chan et al. [4] analyze pitch and energy of the speech signal of a film; Kuo [5] canalizes features such as tempo, melody, mode, and rhythm to classify music and [6] uses information about textures and colors to extrapolate the emotional meaning of an image.

The evaluation of these systems lack of completeness but when the algorithms are evaluated they allow to positively index as much as 85% of media showing the feasibility of this kind of approach.

Few more researches have been studying algorithms for emotion recognition from humans. Pantic and Rothkrantz [7] provides a thorough state of the art in this field of study. The main techniques involve facial expressions, vocal prosody or physiological signals such as heart rate or skin conductivity and attain as much as 90% recognition rate through unimodal classification algoritms.

Generally, the described algorithms work only under a number of lab condition and major constraints. Few system have analyzed the possibility of using bimodality to improve reliability or to reduce the number of constraints. Busso et al. [8] for example use bimodal audio and visual algorithms to improve the recognition rate reaching as much as 92% on 4 emotions (anger, happiness, sadness, and neutral).

In this paper, we present a general architecture which extracts affective information and attaches this semantic information to the medium itself. Different modalities, feature sets, and fusion schemes are compared in a set of studies.

This paper is organized as follows. Section 2 discusses the audio-video database we use for this study. Section 3 introduces some possible scenarios in which emotions can actively be used to improve media searches. Section 4 discusses the architecture that we designed to extrapolate emotions and link them to the medium itself. Section 5 describes the various experiments we have conducted and their results. Finally Section 6 presents our conclusions and cues for future work.

2. THE ENTERFACE DATABASE

The eNTERFACE database [9] is a publicly available audio-visual emotion database. The base contains videos of 44 subjects coming from 14 different nationalities. 81% of the subjects were men, 31% wore glasses and 17% had a beard, finally one of the subjects was bald (2%).

Subjects were told to listen to six different short stories, each of them eliciting a particular emotion (anger, disgust, fear, happiness, sadness, and surprise), and to react to each of the situation uttering 5 different predefined sentences. Subjects were not given further constraints or guidelines regarding how to express the emotion and head movements were allowed.

The base finally contains 44 (subjects) by 6 (emotions) by 5 (sentences) shots. The average video length is about 3 seconds summing up to 1320 shots and more than one hour of videos. Videos are recorded in a lab environment: subjects are recorded frontal view with studio lightening condition and gray uniform background. Audio is recorded with an high quality microphone placed at around 30 cm from the subject mouth.

The eNTERFACE audio-visual database is a good emotion database and is the only publicly available that we have found for bimodal audio and video; it does, nevertheless present some limitations:

1. The quality of the encoding is mediocre: the 720x576 pixels videos are interlaced and encoded using DivX 5.05 codec at around 230 Kbps using 24 bits color information resulting, sometimes, in some blocking effect.

2. Subjects are not trained actors possibly resulting in a mediocre emotional expression quality.

3. Subjects were asked to utter sentences in English even though this was not, in most cases, their natural language; this may result in a low quality of the prosodic emotional modulation.

4. Not all of the subject learned their sentences by heart resulting in a non negligible percentages of videos starting with the subjects looking down to read their sentences.

5. The reference paper acknowledge some videos (around 7.5%) does not represent in a satisfactory way the desired emotional expressions. These videos were, in theory, rejected but this is apparently not the case in the actual database.

This kind of drawbacks introduce some difficulties but it allows us to develop algorithms which should be robust in realistic scenarios. A user study will, nevertheless, be conducted in the future to evaluate the human ability to recognize the emotions presented in the database and generally to evaluate the database quality.

3. SCENARIOS

In many cases, it is very interesting to use emotions for indexing and retrieval tasks. For example one could argue it is simpler to define music as “romantic” or “melancholic” than to define its genre, tempo or melody. Similarly, film and book genres are strongly linked to emotions as can clearly be seen in the case of comedies or horrors. For these reasons we argue that content based semantic tags need to be coupled to emotions to build complete and flexible systems.

One example showing the importance of a multidisciplinary approach, is automatic movie summarisation. Indeed when constructing a summary, both specific objects/events (such as explosions, kissing, and others depending on the movie genre), and scenes regarding specific emotions (both elicited in the public or shown by the character) should be selected. For example, we could say that the summary of an action movie should elicit “suspense” or that an horror movie should show fearful people. Fusion between the two modalities (events and emotions) will enable selecting scenes according to both their content and affective meaning (e.g. thrilling gunfights, romantic speech, etc.)

The same principles can be applied to an indexing scenario: an action movie could be, for example, characterized by the fact of having an ongoing rotation of relevant emotions and for having explosion or shooting scenes. A documentary about demolitions through controlled explosions, though, will contain the very same explosions as an horror movie will contain relevant emotions. The presence of emotions will discriminate between the documentary and the action movie and the presence of explosion will discriminate between the action and the horror genres. Using a combined approach should, therefore, correctly index the videos.

In other domains the input about user emotions could help improving the quality of the feedback to the user himself; this is the case of gaming, telemedicine, e-learning, communications, and all human-computer interactions when the affective state plays an important role in the interaction.

We have seen, so far, how emotions can join other media content descriptors in order to improve upon the performance of content-based retrieval and semantic indexing systems. We have briefly listed some other scenarios in which emotions can play an important role. In the next section we describe “Semantic Affect-enhanced MultiMedia Indexing” (SAMMI), a framework we are developing which allows creating such a kind of systems.

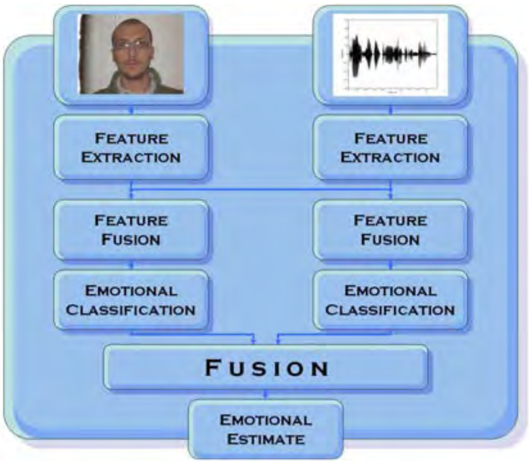


Fig. 1. Bimodal emotion recognition

4. THE GENERAL FRAMEWORK

This section describes “Semantic Affect-enhanced MultiMedia Indexing” (SAMMI), a framework explicitly designed for extracting reliable real-time emotional information through multimodal fusion of affective cues and to use it for emotion-enhanced indexing and retrieval of videos.

Two main limitations of existing work on emotion-based indexing and retrieval have been shown: 1) emotion estimation algorithms are very simple and not very reliable and 2) emotions are generally used without being coupled with any other content information.

In SAMMI, emotions are estimated exploiting the intrinsic multimodality of the affective phenomena. Indeed, emotions have a visual component (facial expression, gestures, etc.), an auditory component (vocal prosody, words uttered, etc.), and generally modulates both volitional and un-volitional behaviors (autonomous nervous system, action choices, etc.).

SAMMI (see Fig. 1) takes in account two modalities: the visual and the auditory ones. In particular it analyzes facial expressions and vocal prosody.

Video analysis. The Intel OpenCV library [10] is used to analyze the visual part of the signal. At first, the video is analyzed and a face is searched and eventually found using the Adaboost technique. When the face is found and its position on a video frame is defined 12 regions are defined which corresponds to emotionally meaningful regions on the face (see Fig. 3 (forehead, 2 regions on the left & right brows, eyes, nose, upper & lower central mouth, and 2 mouth corners).

For each region a certain number of points is found which will be easy to follow with the Lukas Kanade algorithm [11]. The position of the points inside one region is averaged to increase the stability of the algorithms in case one point is lost or it is moved outside its true position because of video imperfections (in Fig. 3 the small dots on the left represent the followed points while the bigger dots on the right represent the center of mass of the points belonging to one region). These points are followed along the video and the coordinates of the 12 centers of mass are saved.

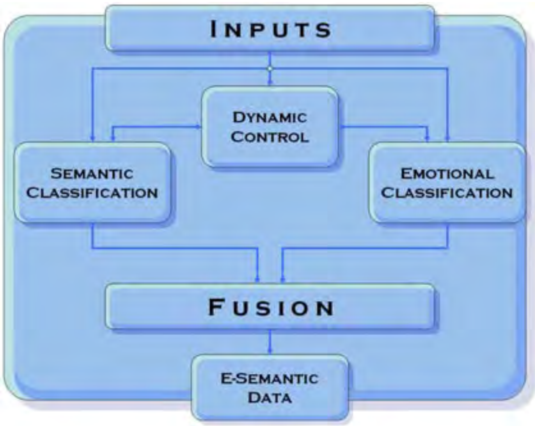


Fig. 2. SAMMI’s architecture

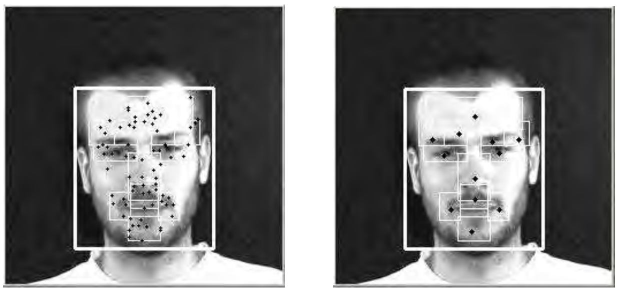


Fig. 3. Followed feature points (FP)

This process generates 12[1] couples (x and y components) of signals which are windowed and analyzed to extract meaningful feature vectors. Different window size have been preliminarily analyzed and a length of 25 frames (1 sec) has been selected as optimal. Two possible feature vectors, a statistical and a polynomial representations, will be presented and compared in the next section. The next step consists in training a classifier with the computed feature vectors. Two classifiers, i.e. the Neural Networks (NN) and the Support Vector Machine, will be compared in the next section.

Audio Analysis. The audio is analyzed offline with the help of PRAAT [12], a powerful open source toolkit for audio, and particularly speech, analysis. Thanks to this software pitch, formants, linear predictive coefficients (LPC), mel-frequency cepstral coefficients (MFCC), harmonicity and intensity of the speech are computed. Again these signals are windowed (1 sec windows), 2 different feature vectors are computed (statistical and polynomial), and two different classifiers are trained (NN and SVM). This approach is substantially equivalent to the one described in [13].

[1] Only 11 of the 12 feature points are actually followed because the upper central mouth point was judged not stable enough

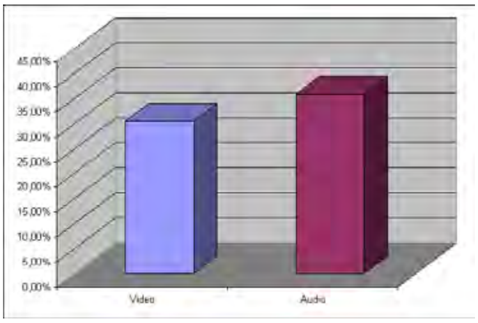


Fig. 4. Video vs. Audio average emotion estimation

As described before, SAMMI (see Fig. 1) takes advantage of multimodality to increase estimation reliability. Emotional information are, thus, fused together and two different levels of fusions are available. These will be described in the next section: 1) the feature level fusion (i.e. the feature vectors are fused and classified together) and 2) the decision level fusion (i.e. the emotional estimates are "averaged" together).

The emotional estimate is part of the general framework described in Fig. 2. Indeed, to develop the applications listed in Section 3, it is necessary to couple the emotional information with other content based semantic information.

Other techniques need to be developed to extract other content-based semantic information and to fuse together these two kinds of information [14, 15].

Dynamic control (Fig. 2) is used to adapt the multimodal fusion according to the qualities of the various modalities at hand. Indeed, if lighting is inadequate the use of color information should be limited and the emotion estimate should privilege the auditory modality.

We have overviewed SAMMI, a framework for semantic tagging of medias enhanced by affective information. In the next section we presents and compare different the techniques for emotion recognition that we have tested.

5. RESULTS

This section relates the results of several studies we have conducted which give an idea about how different modalities, features sets, fusion techniques, and or other algorithms can influence the results of the emotion estimation.

5.1. Modalities: Video vs Audio

SAMMI can analyze 2 different modalities: 1) video and 2) audio. Figure 4 shows how the two modalities perform on average. It is possible to notice that on average our processing of the audio signal is 17% more reliable than the processing of the video signal.

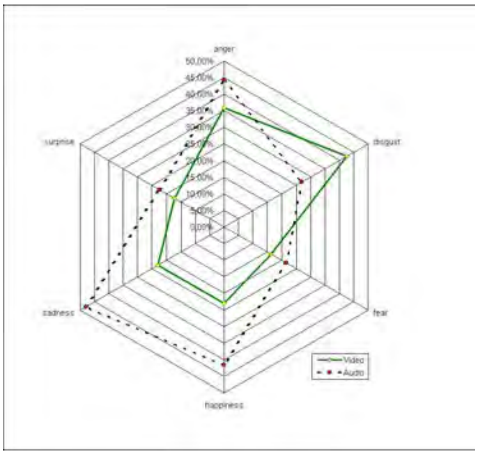


Fig. 5. Video vs. Audio emotion estimation

Observing Figure 5[2] we can observe the behavior of the algorithms for the 6 analyzed emotions. In general, we estimate that, for equivalent average score, a figure resembling more to an hexagon centered in the origin is the optimal (which is also equivalent to say that the best figure is the one maximizing the minimum recognition score or that we want to maximize the inscribed area).

In this case the audio processing figure is the best but we can observe an interesting behavior: both audio and video work better on some particular emotions but these emotions are not the same for the two modalities. In particular, even if the video processing seems less reliable than the audio one, the emotion "disgust" is better recognized from the facial expression than from the vocal prosody. Combining the information coming from the two modalities should therefore improve the overall results.

5.2. Feature Sets

For both modalities we have been testing two feature sets as well as a third resulting from the concatenation of the previous feature vectors. The signals resulting from processing both video and audio were described with 1) a statistical model (based on mean, variance, standard deviation, 5 quantiles, max (and its position), and min (and its position)) and 2) a fifth order polynomial approximation.

In Figures 6 and 7 we can see the effect of the different feature sets on the results for the audio modality. One can observe the polynomial analysis works the worse but still carries some information which can improve the results of the statistical analysis. This is probably due to the choice of using a fifth order polynomial regression which is probably too sensitive to small changes in the original data.

We can see that the particular data we had trained make the system perform very badly on the surprise emotion. A

[2] Please notice that the perfect emotion detector will be represented by an hexagon having for vertexes the 100% probability; the random generator will be represented by an hexagon crossing at around 17%.

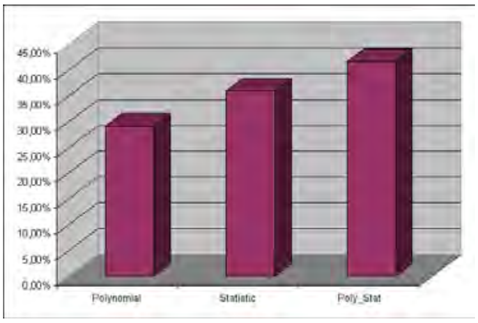


Fig. 6. Average recognition score: different feature sets

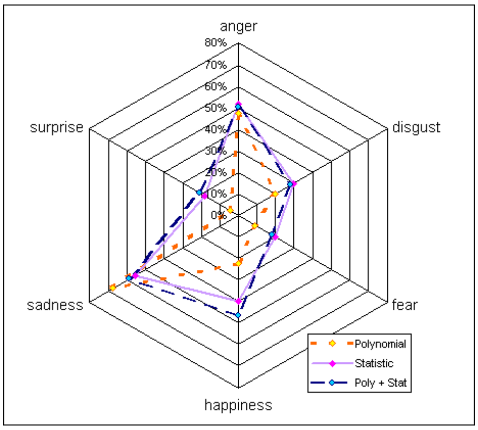


Fig. 7. Emotion recognition using different feature sets

statistical representation of the signals, other than improving the average score, also improves the distribution of the results decreasing the score of the two preferred emotions (sadness and anger) and improving every other score (Fig. 7).

5.2.1. *Low Level Features sets: 11 vs 64 points*

For the video analysis we have also tested a system other than the one based on feature points. The process is basically the same: the face is found and some regions are defined for which the movement is estimated; this time the regions are result of a regular grid of 8 by 8 cells (See Figures 8 & 9).

This approach is equivalent to consider a dense motion flow instead of a feature point (FP) tracking. We have trained the database with these new data and tested it with the three feature sets. The results (Fig. 10, 11) show a similar average recognition rate but a “nicer shape” resulting from the feature point tracking algorithm. This process could, nevertheless, be used to improve results through fusion; in fact the recognition score for half of the emotions is better recognized from the dense flow algorithms than from the FP tracking.

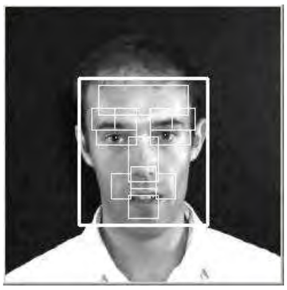


Fig. 8. 12 regions

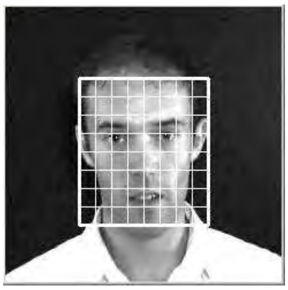


Fig. 9. 64 regions

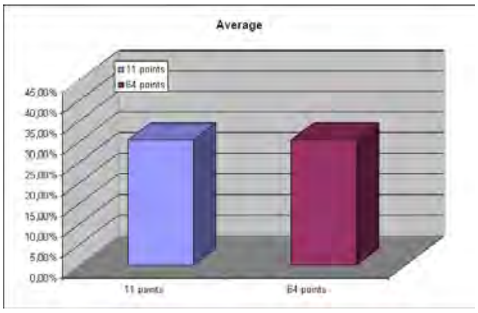


Fig. 10. Average recognition rate: 11 vs. 64 points

5.3. Classifier: SVM vs NN

Another factor which influences the results is the choice of the classifiers. We have been testing neural networks (NN) and support vector machines (SVM). We have employed Matlab to create feed-forward backpropagation neural networks. We did vary the number of neurons between 15 and 100.

We have adopted the libSVM [16] for training and testing SVM. A radial basis function (RBF) has been used as kernel as suggested in [16]. All other parameters have been left to default values.

We can observe in Figure 12 that the average result is quite similar. Nevertheless, we observe, once again, that the distribution of the results (Fig. 13) is quite different suggesting the possibility to exploit fusion of the two results to improve the final recognition score.

5.4. Detectors and Classifiers

Until now the performance of the system have been tested by training one single classifier for the six emotions. However every single emotion has a different temporal behavior. It is therefore possible that the classifier designed for one emotion would work worse on the others and viceversa. This conclusion leads to one solution: substitute 6 detectors to the single classifier. Every detector will be trained to react to one single emotion maximizing the recognition rate.

The results (shown in Figures 14 & 15) show how the adoption of 6 detectors in place of one classifiers for emotion

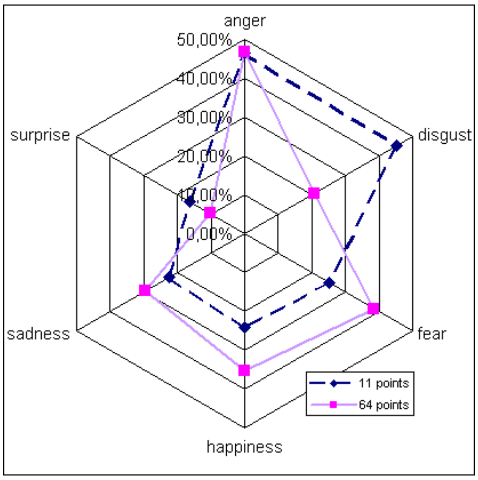


Fig. 11. Emotion estimation: 11 vs. 64 points

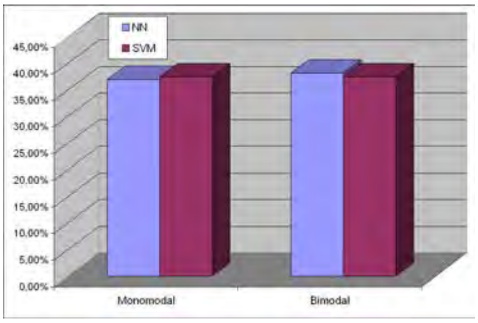


Fig. 12. Average recognition rate: SVM vs. NN

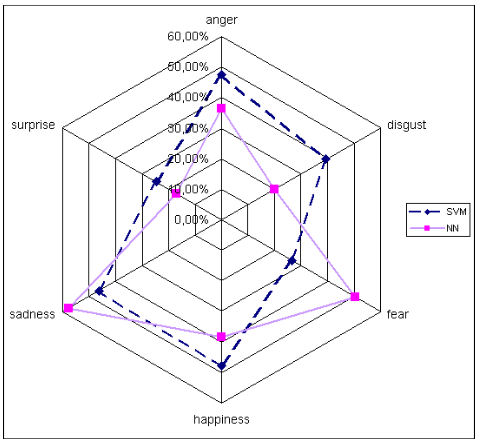


Fig. 13. Emotion estimation: SVM vs. NN

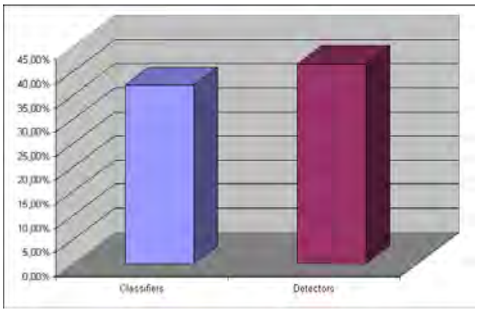


Fig. 14. Average recognition rate: classifier vs. detectors

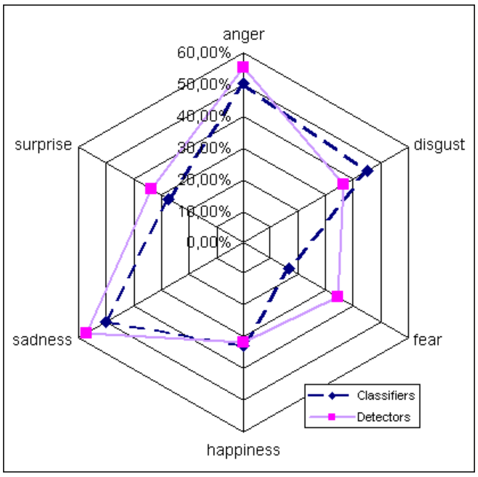


Fig. 15. Emotion estimation: classifier vs. detectors

recognition (from the facial expression) not only improves the average score but also improves the distribution of the results.

5.5. Multimodal Fusion: Features vs Decisions

In this section we want to show how fusion could improve results. In particular we are comparing monomodal systems (Audio and Video) with three different kind of fusion.

The first, namely feature fusion, consist in training the classifier (a NN in Figure 16) with feature vectors resulting from the concatenation of the monomodal video and audio feature vectors. With decision fusion we mean a simple averaging of the outputs of the two monomodal classifiers. With optimized decision fusion we finally mean the weighted average of the two output. The recognition scores for the different emotion of the two monomodal system are used as weights.

Different fusion paradigms leads to different results. We were expecting feature fusion to preserve more information and therefore to work better. Unexpectedly decision fusion (and optimized decision fusion), while employing very simple algorithms works 5% (15%) better than feature fusion. This may suggest that the original data are noisy with respect to emotions and that the emotion estimation is finally noisy

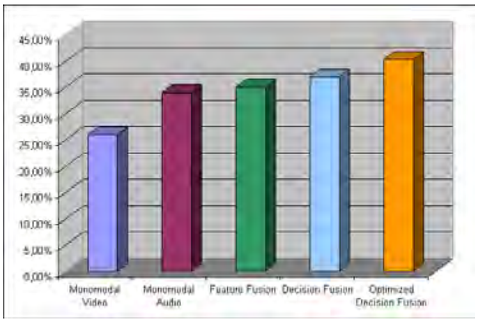


Fig. 16. Average Recognition rate: multimodal fusion

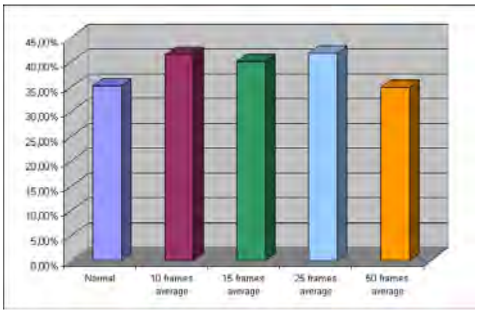


Fig. 17. Average recognition rate: temporal average

too. An average between two modalities does therefore increase the recognition score by reducing the noise (which is statistically independent on the two modalities).

5.6. Temporal Averaging

Starting from the previous conclusions we have tried to perform a temporal averaging of the classification output with the results shown in Figures 17 & 18. The result of such an operation shows a relative improvement of about 18% on the average recognition rate. Furthermore this simple operation does, once again, improve the distribution of the scores improving the recognition rate of the emotions which were less recognized leaving the more recognized emotions untouched. We can notice that averaging windows with lenght between 10 and 25 frames result in similar average score. Bigger averaging windows decreases the score particularly reducing the recognition score of the two emotions fear and happiness.

5.7. Thresholding

Another technique to improve the results is to apply a threshold to the output data. In some scenarios one does not need a frame-by-frame emotion estimation but only to detect when a strong emotion is expressed. In this cases applying a threshold may result in a good technique to improve results in the detriment of the number of emotional estimation. We have tried two different paradigm for thresholding.

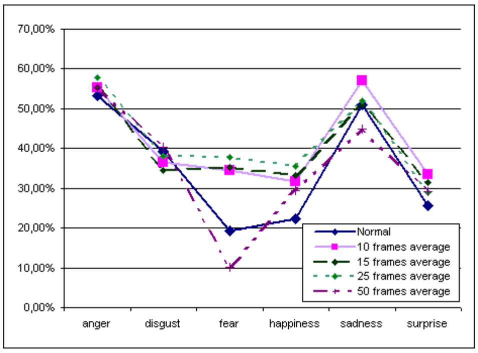


Fig. 18. Emotion estimation: temporal average

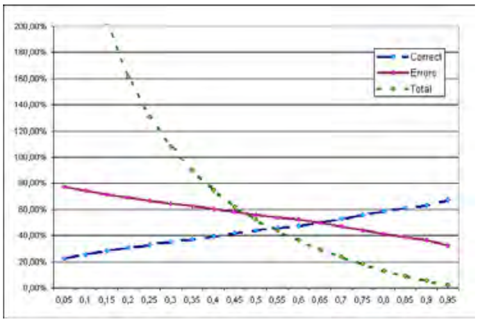


Fig. 19. Effect of thresholding on average recognition

In Figure 19 we show the result of a thresholding of the data for which every detection probability superior to the threshold (x axis) is transformed to 1. By doing this we actually increase the number of detected emotion and we remove the constraint of having one emotion at time, but we detect also the emotions which are not the most likely but are, nevertheless, more probable than the threshold. For low threshold values this technique increases the number of errors but it also improves the score linearly with the threshold up to 67% (from the original 34%). Please note that for low theshold the number of detection is higher than 1 per sample (6 per sample for $th = 0$, 3.6 per sample for $th = 0.05$, etc.).

In Figure 20 we show the result for a thresholding of the data which change to 0 every value smaller than the threshold. In this case the process cuts out emotions which are classified with a low detection probability, compulsorily decreasing the number of detections. The score improves with the threshold up to a maximum of 54.67%.

Both techniques can be exploited to increase the percentage of correctly detected emotion while reducing the total number of detections (note that maximum score is obtained when around 5% frames are tagged).

6. CONCLUSIONS

We have introduced SAMMI: a framework for Semantic Affect-enhanced MultiMedia Indexing. We have overviewed its ar-

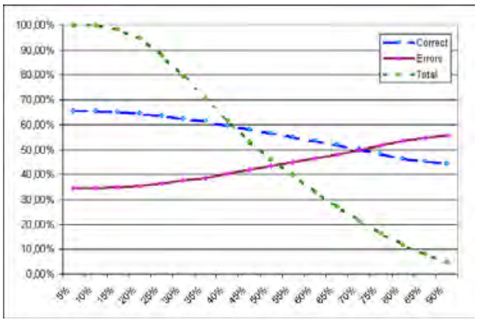


Fig. 20. Effect of thresholding on average recognition (with MAX)

chitecture and presented some scenarios which emphasize the need for a combined affective and semantic tagging.

We have discussed the matter of emotion recognition through speech prosodic features and facial expressions. We have proposed and tested several techniques which can be used to ameliorate the recognition algorithms. Using such techniques we have succeded to double the average recognition rate. Finally, we have reached with a single technique 67% of average recognition rate. The adoption and the fusion of different techniques can lead to better results.

New feature sets needs to be found to better represents the data. Is our impression that video data should be elaborated more to extract more emotionally meaningfull information.

7. REFERENCES

[1] Michael S. Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain, “Content-based multimedia information retrieval: State of the art and challenges,” *ACM Transaction on Multimedia Computing, Communications and Appllications*, vol. 2, no. 1, pp. 1–19, February 2006.

[2] Andrew Salway and Mike Graham, “Extracting information about emotions in films,” in *Proceedings of ACM Multimedia ’03*, 2003, pp. 299–302, Berkeley, CA, USA.

[3] Hisashi Miyamori, Satoshi Nakamura, and Katsumi Tanaka, “Generation of views of TV content using TV viewers’ perspectives expressed in live chats on the web,” in *Proceedings of ACM Multimedia ’05*, 2005, pp. 853–861, Singapore.

[4] Ching Hau Chan and Gareth J. F. Jones, “Affect-based indexing and retrieval of films,” in *Proceedings of ACM Multimedia ’05*, 2005, pp. 427–430, Singapore.

[5] Fang-Fei Kuo, Meng-Fen Chiang, Man-Kwan Shan, and Suh-Yin Lee, “Emotion-based music recommendation by association discovery from film music,” in *Proceedings of ACM Multimedia ’05*, 2005, pp. 507–510, Singapore.

[6] Eun Yi Kim, Soo-Jeong Kim, Hyun-Jin Koo, Karpjoo Jeong, and Jee-In Kim, “Emotion-Based Textile Indexing Using Colors and Texture,” in *Fuzzy Systems and Knowledge Discovery*, L. Wang and Y. Jin, Eds. 2005, vol. 3613/2005 of *LNCS*, pp. 1077–1080, Springer.

[7] Maja Pantic and Lon J.M. Rothkrantz, “Toward an Affect-Sensitive Multimodal Human-Computer Interaction,” in *Proceedings of IEEE*, 2003, vol. 91, pp. 1370–1390.

[8] Carlos Busso, Zhigang Deng, Seldran Yildirim, Murtaza Bulut, Chul M. Lee., Abe Kazemzadeh, Sunbok Lee, Ulrich Neumann, and Shrikanth Narayanan, “Analysis of emotion recognition using facial expressions, speech and multimodal information,” in *Proceedings of ICMI*, 2004, pp. 205–211, State College, PA, USA.

[9] Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas, “The eNTERFACE05 Audio-Visual Emotion Database,” in *Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW’06)*. IEEE, 2006.

[10] IntelCorporation, “Open Source Computer Vision Library: Reference Manual,” November 2006, [http://opencvlibrary.sourceforge.net].

[11] Bruce D. Lukas and Takeo Kanade, “An iterative image registration technique with an application to stereo vision,” in *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 1981, pp. 674–679.

[12] Paul Boersma and David Weenink, “Praat: doing phonetics by computer,” January 2008, [http://www.praat.org/].

[13] James Noble, “Spoken emotion recognition with support vector machines,” *PhD Thesis*, 2003.

[14] Eric Galmar and Benoit Huet, “Analysis of Vector Space Model and Spatiotemporal Segmentation for Video Indexing and Retrieval,” in *ACM International Conference on Image and Video Retrieval*, Amsterdam, The Netherlands, 2007.

[15] Rachid Benmokhtar and Benoit Huet, “Multi-level Fusion for Semantic Video Content Indexing and Retrieval,” in *5th International Workshop on Adaptive Multimedia Retrieval*, LIP6, Paris, France, 2007.

[16] Chih W. Hsu, Chih C. Chang, and Chih J. Lin, “A practical guide to support vector classification,” *Technical report, Department of Computer Science, National Taiwan University*, 2003.

Multimed Tools Appl
DOI 10.1007/s11042-011-0936-5

An ontology-based evidential framework for video indexing using high-level multimodal fusion

Rachid Benmokhtar · Benoit Huet

Abstract This paper deals with information retrieval and semantic indexing of multimedia documents. We propose a generic scheme combining an ontology-based evidential framework and high-level multimodal fusion, aimed at recognising semantic concepts in videos. This work is represented on two stages: First, the adaptation of evidence theory to neural network, thus giving *Neural Network based on Evidence Theory (NNET)*. This theory presents two important information for decision-making compared to the probabilistic methods: belief degree and system ignorance. The NNET is then improved further by incorporating the relationship between descriptors and concepts, modeled by a weight vector based on entropy and perplexity. The combination of this vector with the classifiers outputs, gives us a new model called *Perplexity-based Evidential Neural Network (PENN)*. Secondly, an ontology-based concept is introduced via the influence representation of the relations between concepts and the ontological readjustment of the confidence values. To represent this relationship, three types of information are computed: low-level visual descriptors, concept co-occurrence and semantic similarities. The final system is called *Ontological-PENN*. A comparison between the main similarity construction methodologies are proposed. Experimental results using the TRECVid dataset are presented to support the effectiveness of our scheme.

Keywords Video shots indexing · Semantic gap · Classification · Classifier fusion · Inter-concepts similarity · Ontology · LSCOM-lite · TRECVid

R. Benmokhtar (✉) · B. Huet
Département Communications Multimédia, Eurécom, 2229, route des crêtes,
06904 Sophia-Antipolis, France
e-mail: rachid.benmokhtar@eurecom.fr

B. Huet
e-mail: benoit.huet@eurecom.fr

Published online: 24 December 2011

1 Introduction

The growing amount of image and video available either online or in one's personal collection has attracted the multimedia research community's attention. There are currently substantial efforts investigating methods to automatically organize, analyze, index and retrieve video information. This is further stressed by the availability of the MPEG-7 standard that provides a rich and common description tool for multimedia contents. Moreover, it is encouraged by TRECVid evaluation campaigns which aim at benchmarking progress in video content analysis and retrieval tools developments.

Retrieving complex semantic concepts such as CAR, ROAD, FACE or NATURAL DISASTER from images and videos requires to extract and finely analyze a set of low-level features describing the content. In order to generate a global result from the various potentially multimodal data, a fusion mechanism may take place at different levels of the classification process. Generally, it is either applied directly on extracted features (*feature fusion*), or on classifier outputs (*classifier fusion*).

In most systems concept models are constructed independently [34, 46, 55]. However, the binary classification ignores the fact that semantic concepts do not exist in isolation and are interrelated by their semantic interpretations and co-occurrence. For example, the concept CAR co-occurs with ROAD while MEETING is not likely to appear with ROAD. Therefore, multi-concept relationship can be useful to improve the individual detection accuracy taking into account the possible relationships between concepts. Several approaches have been proposed. Wu et al. [55] have reported an ontological multi-classification learning for video concept detection. Naphade et al. [34] have modeled the linkages between various semantic concepts via a Bayesian network offering a semantics ontology. Snoek et al. [46] have proposed a semantic value chain architecture for concept detection including a multi-concept learning layer called *context link*. In this paper, we propose a generic and robust scheme for video shots indexing based on ontological reasoning construction. First, each individual concept is constructed independently. Second, the confidence value of each individual concept is re-computed taking into account the influence of other related concepts.

This paper is organized as follows. Section 2 reviews existing video indexing techniques. Section 3 presents our system architecture. Section 4 gives the proposed concept ontology construction, including three types of similarities. Section 5 reports and discusses the experimentation results conducted on the TRECVid collection. Finally, Section 6 provides the conclusion of the paper.

2 Review of existing video indexing techniques

This section presents some related works from the literature in the context of semantic indexing. The field of indexing and retrieval has been particularly active, especially for content such as text, image and video. In [2, 11, 45, 50, 52], different types of visual content representation, and their application in indexing, retrieval, abstracting, are reviewed.

Early systems work on the basis of query by example, where features are extracted from the query and compared to features in the database. The candidate images are ranked according to their distance from the query. Several distance functions can be used to measure the similarity between the query and all images in the database. In Photobook [39], the user selects three modules to analyze the query: face, shape or texture. The QBIC system [13] offers the possibility to query on many features: color, texture and shape. VisualSeek [44] goes further by introducing spatial constraints on regions. The Informedia system [53] includes camera motion estimation and speech recognition. Netra-V [58] uses motion information for region segmentation. Regions are then indexed with respect to their color, position and motion in key-frames. VideoQ [9] goes further by indexing the trajectory of regions. Several papers touch upon the semantic problem. Nephade et al. [33] built a probabilistic framework for semantic video indexing to map low-level media features with high-level semantic labels. Dimitrova [11] presents the main research topics in automatic methods for high-level description and annotation. Snoek et al. [45] summarize several methods aiming at automating this time and resource consuming process as state-of art. Vembu et al. [52] describe a systematic approach to the design of multimedia ontologies based on the MPEG-7 standard and sport events ontology. Chang et al. [20] exploit the audio and visual information in generic videos by extracting atomic representations over short-term video slices.

However, models are constructed to classify video shots in semantic classes. Neither of these approaches satisfy holistic indexing, where a user wants to find high level semantic concepts such as an OFFICE or a MEETING for example. The reason is, that there is a semantic gap [52] between low-level features and high-level semantics. While it is difficult to bridge this gap for every high level concept, multimedia processing under a probabilistic framework and ontological reasoning facilitate, bridging this gap for a number of useful concepts.

3 System architecture

The general architecture of our system can be summarized in five steps as depicted in Fig. 1: (1) features extraction, (2) classification, (3) perplexity-based weighted descriptors, (4) classifier fusion and (5) ontological readjustment of the confidence values. Let us detail each of those steps:

3.1 Features extraction

Temporal video segmentation is the first step toward automatic annotation of digital video for browsing and retrieval. Its goal is to divide the video stream into a set of meaningful segments called shots. A shot is defined as an unbroken sequence of frames taken by a single camera. The MPEG-7 standard defines a comprehensive, standardized set of audiovisual description tools for still images as well as movies. The aim of the standard is to facilitate quality access to content, which implies efficient storage, identification, filtering, searching and retrieval of media [31]. Our system employs five types of MPEG-7 visual descriptors: Color, texture, shape, motion and face descriptors. These descriptors are briefly defined as follows:

3.1.1 Scalable Color Descriptor (SCD)

is defined as the hue-saturation-value (HSV) color space with fixed color space quantization. The Haar transform encoding is used to reduce the number of bins of the original histogram with 256 bins to 16, 32, 64, or 128 bins [17].

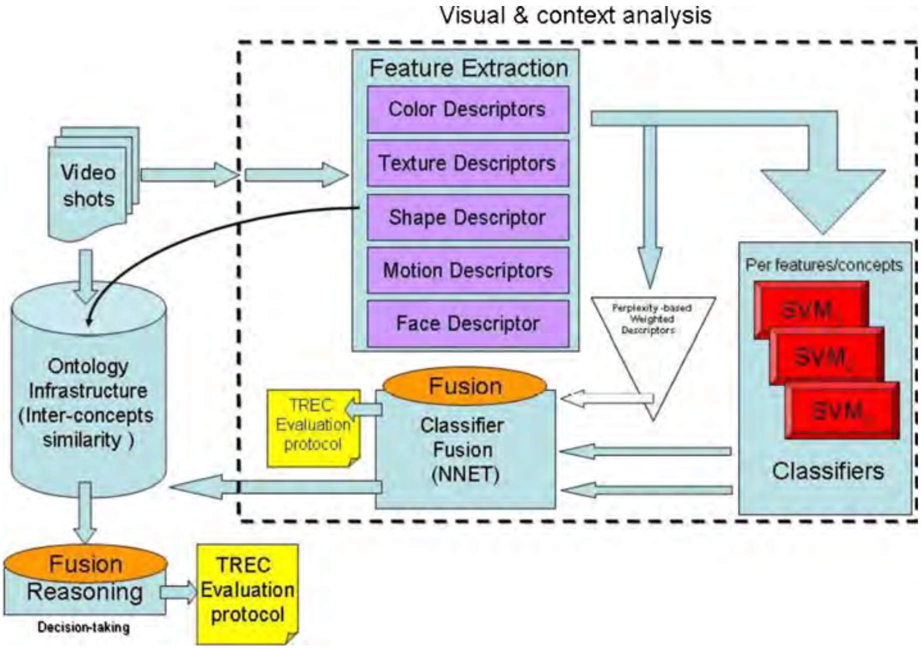


Fig. 1 General indexing system architecture

3.1.2 Color Layout Descriptor (CLD)

is a compact representation of the spatial distribution of colors [21]. The color information of an image is divided into (8×8) block. The blocks are transformed into a series of coefficient values using dominant color descriptor or average color, to obtain $CLD = \{Y, Cr, Cb\}$ components. Then, the three components are transformed by 8×8 DCT (Discrete Cosine Transform) to three sets of DCT coefficients. Finally, a few low frequency coefficients are extracted using zigzag scanning and quantized to form the CLD for a still image.

3.1.3 Color Structure Descriptor (CSD)

encodes local color structure in an image using a structuring element of (8×8) dimension. CSD is computed by visiting all locations in the image, and then summarizing the frequency of color occurrences in each structuring element location on four HMMD color space quantization possibilities: 256, 128, 64 and 32 bins histogram [32].

3.1.4 Color Moment Descriptor (CMD)

provides some information about color in a way which is not explicitly available in other color descriptors. It is obtained by the mean and the variance on each layer of the LUV color space of an image or region.

3.1.5 Edge Histogram Descriptor (EHD)

expresses only local edge distribution in the image. An edge histogram in the image space represents the frequency and the directionality of the brightness changes in the image. The EHD basically represents the distribution of 5 types of edges in each local area called a sub-image. Specifically, dividing the image into (4×4) non-overlapping sub-images. Then, for each sub-image, we generate an edge histogram. Four directional edges (0°, 45°, 90°, 135°) are detected in addition to non-directional ones. Finally, it generates a 80 dimensional vector (16 sub-images, 5 types of edges). We make use of the improvement proposed by [38] for this descriptor, which consist in adding global and semi-global levels of localization of an image.

3.1.6 Homogeneous Texture Descriptor (HTD)

characterizes a region's texture using local spatial frequency statistics. HTD is extracted by Gabor filter banks (6 frequency times, 5 orientation channels), resulting in 30 channels in total. Then, computing the energy and energy deviation for each channel to obtain 62 dimensional vector [31, 56].

3.1.7 Statistical Texture Descriptor (STD)

is based on statistical methods of co-occurrence matrix such as: energy, maximum probability, contrast, entropy, etc [1], to model the relationships between pixels within a region of some grey-level configuration in the texture; this configuration varies rapidly with distance in fine textures, slowly in coarse textures.

3.1.8 Contour-based Shape Descriptor (C-SD)

presents a closed 2D object or region contour in an image. To create *Curvature Scale Space (CSS)* description of contour shape, N equidistant points are selected on the contour, starting from an arbitrary point and following the contour clockwise. The contour is then gradually smoothed by repetitive low-pass filtering of the x and y coordinates of the selected points, until the contour becomes convex (no curvature zero-crossing points are found). The concave part of the contour is gradually flattered out as a result of smoothing. Points separating concave and convex parts of the contour and peaks (maxima of the CSS contour map) in between are then identified. Finally, eccentricity, circularity and number of CSS peaks of original and filtered contour are should be combined to form more practical descriptor [31].

3.1.9 Camera Motion Descriptor (CM)

details what kind of global motion parameters are present at what instance in time in a scene provided directly by the camera, supporting 7 camera operations: fixed, panning (horizontal rotation), tracking (horizontal transverse movement), tilting (vertical rotation), booming (vertical transverse movement), zooming (change of the focal length), dollying (translation along the optical axis), and rolling (rotation around the optical axis) [31].

3.1.10 Motion Activity Descriptor (MAD)

shows whether a scene is likely to be perceived by a viewer as being slow, fast paced, or action paced [48]. Our MAD is based on intensity of motion. The standard deviations are quantized into five activity values. A high value indicates high activity and the low value of intensity indicates low activity.

3.1.11 Face Descriptor (FD)

detects and localizes frontal faces within the keyframes of a shot and provides some face statistics (e.g, number of faces, biggest face size), using the face detection method implemented in OpenCV. It uses a type of face detector called a Haar Cascade classifier, that performs a simple operation. Given an image, the face detector examines each image location and classifies it as "face" or "not face" [37].

3.2 Classification

The classification consists in assigning classes to videos given some description of its content. The literature is vast and ever growing [24]. This section summarizes the classifier method used in the work presented here: "Support Vector Machines".

SVMs have become widely employed in classification tasks due to their generalization ability within high-dimensional pattern [51]. The main idea is similar to the concept of a neuron: Separate classes with a hyperplane. However, samples are indirectly mapped into a high dimensional space thanks to its kernel function. In this paper, a single SVM is used for each low-level feature and is trained per concept under the *"one against all"* approach. At the evaluation stage, it returns for every shots a normalized value in the range [0, 1] using (1). This value denotes the degree of confidence, to which the corresponding shot is assigned to the concept.

$$y_i^j = 1/\left(1 + \exp\left(-\alpha d_i\right)\right) \tag{1}$$

Where (i, j) represents the ith concept and jth low-level feature, d_i is the distance between the input vector and the hyperplane and α is the slope parameter which is obtained experimentally.

3.3 Perplexity-based weighted descriptors

Each concept is best represented or described by its own set of descriptors. Intuitively, the color descriptors should be quite appropriate to detect certain concepts such as: SKY, SNOW, WATERSCAPE, and VEGETATION, while inappropriate for STUDIO, MEETING, MEETING, CAR, etc.

For this aim, we propose to weight each low-level feature according to the concept at hand, without any feature selection (Fig. 2). The variance as a simple second order vector can be used to give the knowledge of the dispersion around the mean between descriptors and concepts. Conversely, the entropy depends on more parameters and measures the quantity of informations and uncertainty in a probabilistic distribution. We propose to maps the visual features onto a term weight vector via entropy and perplexity measures. This vector is then combined with the original classifier outputs[1] to produce the final classifier outputs. As presented in Fig. 2, we shall now define the four steps of the proposed approach [6].

[1] We can also use the weight in the feature extraction step.

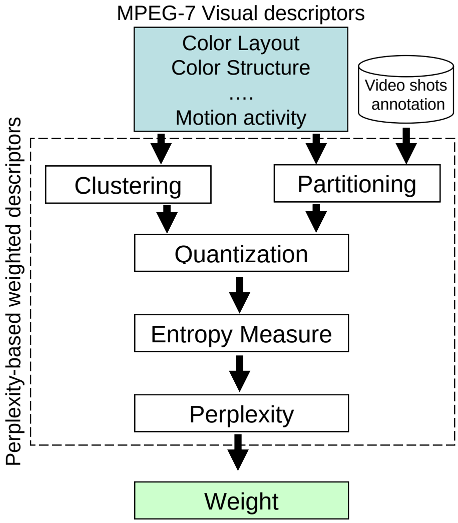


Fig. 2 Perplexity-based weighted descriptors structure

3.3.1 K-means clustering

It computes the k centers of the clusters for each descriptor, in order to create a "visual dictionary" of the shots in the training set. The selection of k is an unresolved problem, and only tests and observation of the average performances can help us to make a decision. In Souvannavong et al. [47], a comparative study of the classification results *vs* the number of clusters used for the quantization of the region descriptors of TRECVid 2005 data, shows that the performances are not deteriorated by quantization of more than 1,000 clusters. Based on this result, our system will employ $k_r = 2{,}000$ for the clustering the MPEG-7 descriptors computed from image regions, and $k_g = 100$ for the global ones. This presents a good compromise between efficiency and a low computation times.

3.3.2 Partitioning

Separating data into positive and negative sets is the first step of the model creation process. Typically, based on the annotation data provided by TRECVid, we select the positive samples for each concept.

3.3.3 Quantization

To obtain a compact video representation, we vector-quantize features. Based on the vocabulary size $k_r = 2{,}000$ (number of visual words) which has empirically shown good results for a wide range of datasets. All features are assigned to their closest vocabulary word using Euclidean distance.

3.3.4 Entropy measure

The entropy H (2) of a certain feature vector distribution $P = (P_0, P_1, ..., P_{k-1})$ gives a measure of concepts distribution uniformity over the clusters k [27]. In [22], a good model is such that the distribution is heavily concentrated on only few clusters, resulting in low entropy value.

$$H = -\sum_{i=0}^{k-1} P_i \log(P_i) \tag{2}$$

where P_i is the probability of cluster i on the quantized vector.

3.3.5 Perplexity measure

In [15], perplexity (PPL) or normalized perplexity value ($\overline{PPL}$) (3) can be interpreted as the average number of clusters needed for an optimal coding of the data.

$$\overline{PPL} = \frac{PPL}{PPL_{\max}} = \frac{2^H}{2^{H_{\max}}} \tag{3}$$

If we assume that k clusters are equally probable, we obtain $H(P) = \log(k)$, and then $1 \leq \overline{PPL} \leq k$.

3.3.6 Weight

In speech recognition, handwriting recognition, and spelling correction [15], it is generally assumed that lower perplexity/entropy correlates with better performance, or in our case, to a very concentrated distribution. So, the relative weight of the corresponding feature should be increased. Many formula can be used to represent the weight such as Sigmoid, Softmax, Gaussian, etc. In our paper, we choose Verhulst's evolution model (4). This function is non exponential, it allows a brake rate α_i to be defined, as well as reception capacity (upper asymptote) K, and β_i defines the decreasing speed of weight function.

$$w_i = K \frac{1}{1 + \beta_i \exp\left(-\alpha_i (1/\overline{\mathrm{PPL}}_i)\right)} \tag{4}$$

$$\beta_i = \begin{cases} K \exp\left(-\alpha_i^2\right) & \text{if } Nb_i^+ < 2 * k \\ 1 \quad \text{Otherwise} & \end{cases} \tag{5}$$

β_i is introduced to decrease the negative effect of the training set limitation, due to the low number of positive samples ($Nb_i^+ << k$) of certain concepts such as WEATHER, DESERT, MOUNTAIN,... (see Table 2). We observe a lower perplexity value, which could not be interpreted as a relevant relation between descriptor and concept. So, we increase β_i (5) to obtain a rapid weight decrease for each concept presenting less than $2 * k$ positive samples.

The relevance of the various descriptors at identifying high level concepts can be obtained through the perplexity distribution (see Fig. 3). The Boxplot provides a good visual summary of many important aspects of a distribution. The lower and upper lines express the data range, the lower and upper edges of the box indicate the 25th and 75th percentile. The line inside the box indicates the median value of the data. Figure 3 shows the normalized perplexity for each descriptor and its best concept presented by the minimum observation, such as: SCD is more effective to detect the concept SKY "13", EDH for ROAD "12", etc. The first observation concerns the same value of median perplexity obtained for SCD, CLD, CMD, CSD, where

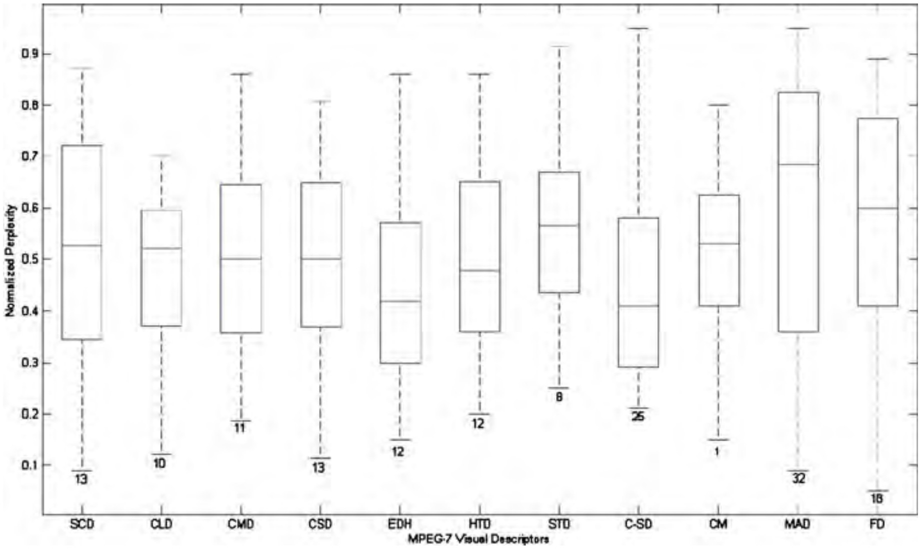


Fig. 3 Normalized perplexity Boxplot

color is more discriminant. Secondly, C-SD gives the smallest 25th percentile of normalized perplexity for all data, followed by EDH and SCD. Thirdly, it seems that EHD is very useful in the detection of the contour as in the SPORT and ROAD concepts. Identical observation is given for C-SD. Conversely, MAD presents a large interval of perplexity but gives small value for the concepts WALKING-RUNNING, PEOPLE-MARCHING where the motion activity can be detected. Finally, FD is a relevant descriptor to detect FACE and PERSON concepts which was to be expected from the very nature of this descriptor.

This approach is proposed to weight each low-level feature per concept, within an adaptive classifier fusion step (Section 3.4). The combination provides a new classification system that we call PENN "Perplexity-based Evidential Neural Network". We will now present the classifier fusion step.

3.4 Classifier fusion

Classifier fusion is an important step of the classification task. It improves recognition reliability by taking into account the complementarities between classifiers, in particular for multimedia indexing and retrieval. Several schemes have been proposed in the literature according to the type of information provided by each classifier as well as their training and adaptation capacity. The state of the art and the comparison study about the effectiveness of the classifier fusion methods are given in [4].

In [12], Duin et al. have distinguished the combination methods of different classifiers and the combination methods of weak classifiers. Another kind of grouping using only the type of classifiers outputs (class, measure) is proposed in [57]. Jain [18] built a dichotomy according to two criteria of equal importance: the type of classifiers outputs and their capacity of learning. This last criteria is used by [25, 26] for grouping

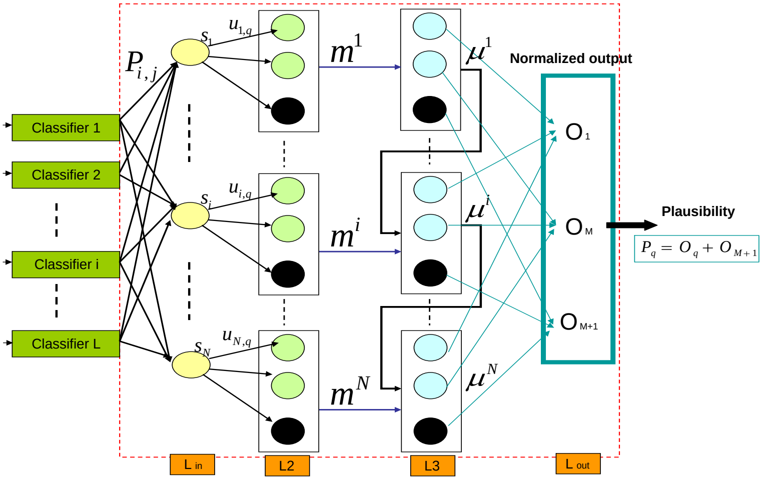


Fig. 4 NNET classifier fusion structure

the combination methods. The trainable combiners search and adapt the parameters in the combination. The non trainable combiners use the classifiers outputs without integrating another *a priori* information of each classifiers performances.

In this part, we describe our proposed neural network based on evidence theory (NNET) [5] to address classifier fusion (Fig. 4).

1. **Layer** L_{input}: Contains N units. Identical to the RBF (Radial Basis Function) network input layer with an exponential activation function ϕ. d: distance computed using training data. $\alpha \in [0, 1]$ is a weakening parameter associated to unit i.

$$\begin{cases} s^i = \alpha^i \phi(d^i) \\ \phi(d^i) = \exp\left(-\gamma^i (d^i)^2\right) \end{cases} \tag{6}$$

2. **Layer** L_2: Computes the belief masses m^i (7) associated to each unit. The units of module i are connected to neuron i of the previous layer.

$$\begin{cases} m^i(\{w_q\}) = \alpha^i u^i_q \phi(d^i) \\ m^i(\Omega) = 1 - \alpha^i \phi(d^i) \end{cases} \tag{7}$$

where u^i_q is the membership degree to each class w_q, q class index $q = \{1, ..., M\}$.

3. **Layer** L_3**:** The Dempster–Shafer combination rule combines N different mass functions in one single mass. It is given by the conjunctive combination (8):

$$m(A) = (m^1 \oplus ... \oplus m^N) = \sum_{B_1 \bigcap ... \bigcap B_N = A} \prod_{i=1}^{N} m^i(B_i) \tag{8}$$

The activation vector of modules i is defined as $\overrightarrow{\mu^i}$. It can be recursively computed using:

$$\begin{cases} \mu^1 = m^1 \\ \mu^i_j = \mu^{i-1}_j m^i_j + \mu^{i-1}_j m^i_{M+1} + \mu^{i-1}_{M+1} m^i_j \\ \mu^i_{M+1} = \mu^{i-1}_{M+1} m^i_{M+1} \end{cases} \tag{9}$$

4. **Layer** L_{output}: In [10], the output is directly obtained by $O_j = \mu^N_j$. The experiments show that this output is very sensitive to the number of prototype, where a small modification in the number can change the classifier fusion behavior. To resolve this problem, we use normalized output (10). Here, the output is computed taking into account the activation vectors of all prototypes to decrease the effect of an eventual bad behavior of prototype in the mass computation.

$$O_j = \frac{\sum_{i=1}^{N} \mu^i_j}{\sum_{i=1}^{N} \sum_{j=1}^{M+1} \mu^i_j} \tag{10}$$

$$P_q = O_q + O_{M+1} \tag{11}$$

The different parameters (Δu, $\Delta \gamma$, $\Delta \alpha$, ΔP, Δs) can be determined by gradient descent of output error for an input pattern x. Finally, the maximum of plausibility P_q of each class w_q is computed.

Therefore, the combination between perplexity-based weighted low-level feature per concept, within the adaptive NNET classifier fusion provides a novel system that we call PENN "Perplexity-based Evidential Neural Network".

4 Concept ontology construction

The ontology has been historically used to achieve better performance in the multimedia retrieval system [8]. It defines a set of representative concepts and the inter-relationships among them. It is therefore important to introduce some constraints to the development of the similarity measures before proceeding to the presentation of our method. Psychology demonstrates that similarity depends on the context, and may be asymmetric [30]. However, when ontologies have been defined for multimedia they have not been extensively used at the decision making stage of high level concept detection.

Most indexing models are based on binary classification, ignoring possible relationships between concepts. However, concepts do not exist in isolation and are interrelated by both their semantic interpretations and co-occurrence. Wu et al. [55] have reported an ontological multi-classification learning for video concept detection in the NIST TREC-2003 Video Retrieval Benchmark.[2] Ontology-based multi-classification learning consists of two steps. At the first step, each single concept model is constructed independently based on SVM (Support Vector Machine). At

[2] NIST TREC-2003 Video Retrieval Benchmark defines 133 video concepts, organized hierarchically and each video data belong to one or more concepts [35].

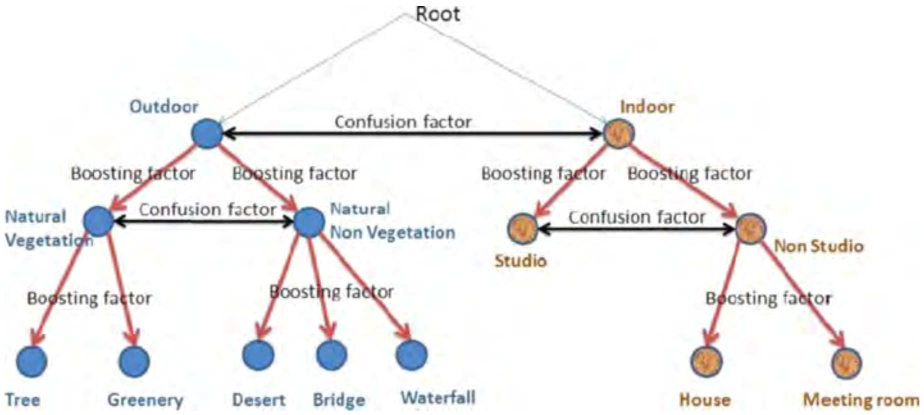


Fig. 5 An example of ontology used by Wu et al. [55]

the second step, ontology-based concept learning improves the accuracy of individual classifiers based on updating confidence scores from single concept models. Two kinds of influences have been defined: *confusion factor* β and *boosting factor* λ. The *confusion factor* β is the influence between concepts that can not be co-existent. The *boosting factor* λ is the top-to-down influence from big to small concepts in the ontology hierarchy. A small example of such influence is presented in Fig. 5. The factors are obtained using a correlation study of the training data. Then, an update of the novel confidence is applied, as shown in (12) and (13).

$$\begin{cases} \underline{p}(x/C_i) = p(x/C_i) + \sum_{j\in\psi} \lambda^i_j p(x/C_j) \\ \lambda^i_j = \dfrac{A}{B + \exp\left(C|p(s/C_i) - p(s/C_j)|\right)} \end{cases} \tag{12}$$

$$\begin{cases} \underline{p}(x/C_i) = \dfrac{p(x/C_i)}{\beta} \\ \beta = \dfrac{1}{f(p(x/C_i) - \max_{j\in\theta}(p(x/C_j)))} \end{cases} \tag{13}$$

The parameters A, B and C of (12) are empirically obtained as described in the works of Li et al. [28]. $f(.)$ is a positive and increasing function for the (13).

Naphade et al. [34] have modeled the linkages between various semantic concepts via a Bayesian network offering a semantics ontology. The central theme to this approach is the concept of Multijects or Multimedia Objects. A Multiject has a semantic label and summarizes a time sequence of low level features of multiple modalities in the form of a probability. It has 3 main aspects: The first aspect is the semantic label. The second aspect of a Multiject is, that it summarizes a time sequence of low level features. The detection of a certain Multiject can increase or decrease the probability of occurrence of other Multiject. For example, if the Multiject BEACH is detected with a very high probability, then the probability of occurrence of the Multiject YACHT or the Multiject SUNSET increases. This is the third aspect of Multijects, i.e. their interaction in a network.

Authors assume that all concepts have the same semantic level, related by the conditional dependence relation with the associated low-level descriptors.

Fan et al. [14] have proposed a hierarchical classification for image annotation.[3] This approach introduce the contextual dependences of the WordNet ontology and the co-occurrence relationship, as presented by the following equation:

$$\begin{cases} \lambda(C_m, C_n) = \rho(C_m, C_n)\pi(C_m, C_n) \\ \text{ou} \quad \rho(C_m, C_n) = log\left(\dfrac{P(C_m, C_n)}{P(C_m)P(C_m)}\right) \\ \text{et} \quad \pi(C_m, C_n) = -log\left(\dfrac{dist(C_m, C_n)}{2D}\right) \end{cases} \tag{14}$$

with $\rho(C_m, C_n)$ is the joint probability between two concepts. It is obtained by the computation of the frequency for the co-occurrence of the relevant C_m and C_n. $\pi(C_m, C_n)$ is the contextual dependency, extracted in the ontology structure (*dist* is the length of the shortest path between two concepts, and D is the maximum depth of the WordNet).

Hauptmann et al. [16] have presented a comparison between the unimodal and the multimodal indexing. The multimodal system learn the dependence between concepts using the following graphical models: *Conditional Random Field "CRF"* and *Bayesian network*. The two models provide closer results in term of precision but are better than the unimodal approach. Koskela et al. [22] have exploited the correlations between the concepts to build a clustering method.

In another development, Li et al. [29] have proposed a study of various linear and non-linear functions $S = f(f_1, f_2, f_3)$ depending on the shortest path length l, depth of subsumer concept in the hierarchy h, and the local semantic density d, as shown in (15).

$$\begin{cases} f_1 = \exp(-\alpha l) \\ f_2 = \dfrac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} \\ f_3 = \dfrac{e^{\delta d} - e^{-\delta d}}{e^{\delta d} + e^{-\delta d}} \end{cases} \tag{15}$$

where α is a constant, $\beta > 0$ is a smoothing factor. $\delta = max_{c \in CS(c_m, c_n)}(-\log p(c))$ represents the semantic similarity measured by the information content.

Several combinations have been applied and evaluated such as: $S_1 = f_1$, $S_2 = f_1 f_2$, $S_3 = S_2 f_3$, $S_4 = S_2 + f_3$, etc. The obtained results with different parameters (α and β) indicate that different functions have satisfactory performances, particularly those that use the three influences.

Discussion The work of Wu et al. [55] uses a confidence update using the correlation of data, and a fixed ontology structure. Naphade et al. [34] have trained the low-level features, and the co-occurrence between concepts. Koskela et al. [22] have included the co-occurrence and visual information in the construction of

[3] Three image datasets are used: *Corel Images, Google Images, and LabelMe.*

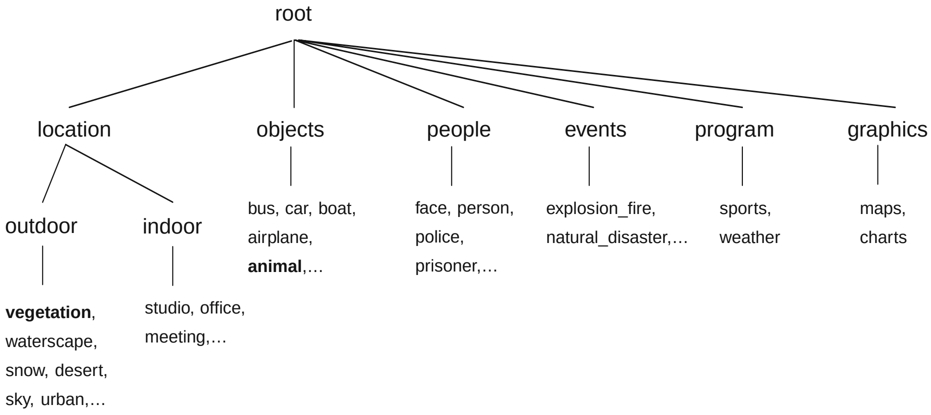

root
location
objects
people
events
program
graphics
outdoor
indoor
bus, car, boat, airplane, **animal**,…
face, person, police, prisoner,…
explosion_fire, natural_disaster,…
sports, weather
maps, charts
vegetation, waterscape, snow, desert, sky, urban,…
studio, office, meeting,…

Fig. 6 Fragment of the hierarchical LSCOM-Lite

this relationship. Fan et al. [14] as in Li et al. [29] have incorporated contextual dependencies of the WordNet ontology, and co-occurrences. This paper extends preceding works in term of the inter-concepts similarity construction. We use the co-occurrence in the corpus, the visual information outcome from low-level description, and finally the hybrid semantic similarity obtained from the ontology architecture.

In LSCOM-lite ontology[4] [35], we notice positive relationships such as (ROAD, CAR), (VEGETATION, MOUNTAIN), and negative relationships like (BUILDING, SPORTS), (SKY, MEETING).

Here, we will investigate how the relationship between different semantic concepts can be extracted and used. One direct method for similarity calculation is to find the minimum path length of connecting two concepts [40]. For example, Fig. 6 illustrates a fragment of the semantic hierarchy of LSCOM-Lite. The shortest path between VEGETATION and ANIMAL is VEGETATION-OUTDOOR-LOCATION-ROOT-OBJECTS-ANIMAL. The minimum length of a path is 5. Or, the minimum path length between VEGETATION and OUTDOOR is 1. Thus, we could say in LSCOM-lite ontology, OUTDOOR is more similar semantically to VEGETATION than ANIMAL. But, we should not say ANIMAL is more similar to CAR. In an other way, OUTDOOR contains many different concepts such as "DESERT, URBAN, ROAD,etc" each with different colors and textures scene descriptions. Therefore, the linking of concepts can infer new and more complex concepts, or improve the recognition of concepts previously detected. Thus, the presence or absence of certain concepts suggests a high or low capability to find other concepts (e.g. detection of SKY and SEA increases the probability of the concept BEACH and reduces the likelihood of DESERT). For this, more information between the concepts are introduced, so that it becomes a function of attributes

[4] The LSCOM-lite (Large-Scale Concept Ontology for Multimedia) [36] annotations include 39 concepts, which are interim results from the effort in developing a LSCOM. The dimensions consist of program category, setting/scene/site, people, object, activity, event, and graphics. A collaborative effort among participants in the TRECVid benchmark was completed to produce the annotations. Human subjects judge the presence or absence of each concept in the video shots.

"co-occurrence, low-level visual descriptors, path length, depth and local density" to boost the performance of specific indexing system, as:

$$\lambda(C_m, C_n) = \sum_{i=\{\text{Cos, Vis, Sem}\}} (Sim_i(C_m, C_n)) \tag{16}$$

Below, we explain with more details the similarity forms used in our architecture.

4.1 Co-occurrence

The first similarity is obtained by considering the co-occurrence statistics between concepts, where the presence or absence of certain concepts may predict the presence of other concepts. Intuitively, documents (video shots) that are "close together" in the vector space relate to similar things. Many methods are proposed in literature to represent this proximity such as: Euclidean, Hamming, Dice, etc. Here, we use Cosine similarity because it reflects similarity in terms of relative distributions of component. Cosine is not influenced by one document being small compared to others like the Euclidean distance tends to be [23]:

$$Sim_{\cos}(P^m, P^n) = \frac{\sum_{i=0}^{k-1} P_i^m P_i^n}{\sqrt{\sum_{i=0}^{k-1} (P_i^m)^2 \sum_{i=0}^{k-1} (P_i^n)^2}} \tag{17}$$

4.2 Visual similarity

The second similarity is based upon low level visual features. In Section 3.3, we have used perplexity to build a weighted descriptor per concept. Now, in order to compute the visual similarity d_{vis}, we are interested in mutual information presented as a measure of divergence. To this end, several measures are proposed in the literature: *Jensen–Shannon (JS), Kullback–Leibler (KL)*, etc. We decided to use d_{JD} *Jeffrey divergence* [23] which is like d_{KL}, but is numerically more stable.

$$d_{JD}(P^m, P^n) = \sum_{i=0}^{k-1} \left(P_i^m \log \frac{P_i^m}{\hat{P}_i} + P_i^n \log \frac{P_i^n}{\hat{P}_i} \right) \tag{18}$$

where $\hat{P}_i = \frac{P^m + P^n}{2}$ is the mean distribution. The visual distance between two concepts is:

$$Sim_{\text{vis}}(C_m, C_n) = \frac{1}{\sum_{i=1}^{Nb\ \ features} \frac{1}{2}(w_i^m + w_i^n) d_{JD}^i(P^m, P^n)} \tag{19}$$

where w_i^m is the ith perplexity-based weighted descriptors for the concept m.

4.3 Semantic similarity

The semantic similarity between the concepts has been widely studied in the literature and can be classified in three major approaches [43]:

4.3.1 Distance-based approach

It estimates the distance (edge length) between nodes which correspond to the concepts being compared. Two concepts C_m and C_n are similar if their path is short,

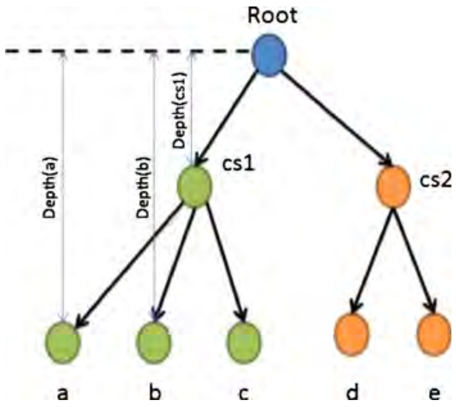


Fig. 7 The concept similarity measure

presented by the minimum number of edges that separates the two concepts. Rada et al. [40] propose the following equation:

$$Sim_{sem}(C_m, C_n) = 1/\left(1 + dist_{Rada}(C_m, C_n)\right) \tag{20}$$

Wu and Palmer [54] propose a similarity-based (see Fig. 7) on the depth of the concept subsumes CS[5] and the two concepts (21).

$$Sim_{sem}(C_m, C_n) = \frac{2 * depth(CS)}{depth(C_m) + depth(C_n)} \tag{21}$$

The drawbacks of this approach are its dependence on the concepts position in the hierarchy, and that all edges have the same weight, which imposes difficulties in defining and controlling the distance edges.

4.3.2 Information content-based approach

It takes into account the information shared by the concepts in terms of entropy measure. Two methods exist. The first uses a learning corpus and compute the probability $p(C_i)$ to find the concept C_i or one of its descendants. For Resnik [41], the semantic similarity can be obtained per the frequency of appearance in the corpus, and defined by:

$$Sim_{sem}(C_m, C_n) = \max\left(IC(CS(C_m, C_n))\right) \tag{22}$$

with $IC(C_i) = -\log(p(C_i))$ is the information content of the concept C_i (i.e, the entropy of a class C_i). The probability $p(C_i)$ is computed by dividing the number of instances of C_i by the total number in the corpus. This measure does not seem complete and precise because it depends on the specific subsumed concept only.

The second method computes the information content of nodes from WordNet instead of a corpus. Seco et al. [42] use descendant hyponyms of the concepts to obtain the information content. This approach can produce a similarity between two neighbor concepts of an ontology, exceeding the value of two concepts contained in the same hierarchy. This is inadequate in the context of information retrieval.

[5] The concept subsumes is the most common specific concept.

4.3.3 Hybrid approach

The hybrid approach combines the two previous approaches. Often, it reuses the information content of nodes and the smallest common ancestor, as with the equation of Lin et al. [30], or with the distance of Jiang and Conrath $dist_{J\&C}$ [19].

$$Sim_{sem_{Lin}}(C_m, C_n) = \frac{2 * \log P(CS)}{\log P(C_m) + \log P(C_n)} \tag{23}$$

$$\begin{cases} dist_{J\&C}(C_m, C_n) = IC(C_m) + IC(C_n) - 2 * IC(CS(C_m, C_n)) \\ Sim_{sem_{J\&C}}(C_m, C_n) = 1/\left(dist_{J\&C}(C_m, C_n)\right) \end{cases} \tag{24}$$

For the ontology presented in the Fig. 8, we compare the last two hybrid approaches with the novel one as presented in the (26), that it is the combination of Rada [40] and J&C [19].

$$\begin{cases} Sim_{sem_{J\&C}}(C_m, C_n) = 1/d_{J\&C}(C_m, C_n) \\ d_{J\&C}(C_m, C_n) = IC(C_m) + IC(C_n) - 2 * IC(CS(C_m, C_n)) \end{cases} \tag{25}$$

$$Sim_{sem}(C_m, C_n) = 1/\left(d_{Rada}(C_m, C_n) + d_{J\&C}(C_m, C_n)\right) \tag{26}$$

where $d_{Rada}(C_m, C_n)$ is the length of the shortest path between C_m and C_n.

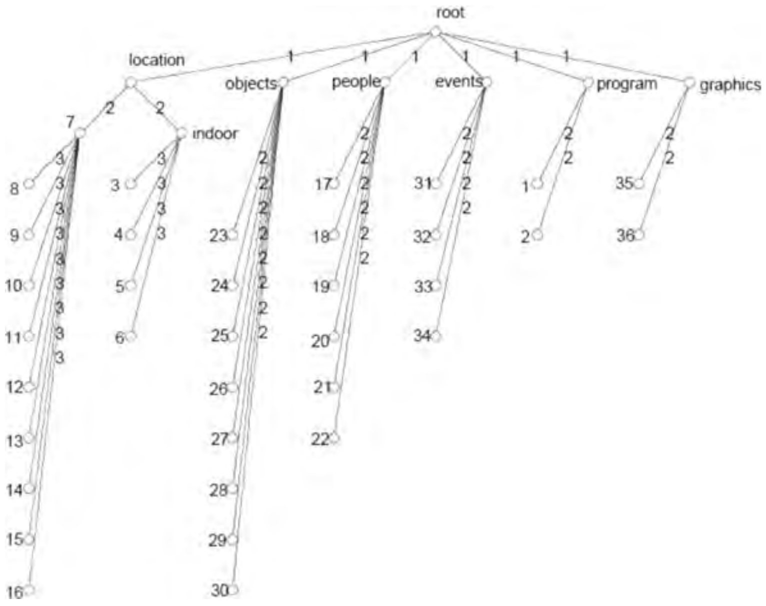


Fig. 8 Hierarchical ontology model

4.4 Concept-based confidence value readjustment (CCVR)

The proposed framework (Fig. 1) introduces a *reranking* or confidence value readjustment to refine the PENN results for concept detection [7], and is computed using:

$$\underline{P(x/Ci)} = P(x/Ci) + \frac{1}{Z}\sum_{j=1}^{Nb\ arc} \lambda_{i,j}(1-\zeta_j)P(x/Cj) \tag{27}$$

where $\underline{P(x/Ci)}$ corresponds to the multi-modal PENN result, $\lambda_{i,j}$ is the causal relationship between concepts C_i and C_j, ζ_j is the classifier error in the validation set and Z is a normalization term.

5 Experimentations

The experiments provided here are conducted on the TRECVid 2007 dataset [49] containing science news, news reports, documentaries, etc. Of the 100 hours of video segmented into shots and annotated [3] with semantic concepts from the 36 defined labels. Half is used to train the feature extraction system and the other half is used for evaluation purposes. The evaluation is realized in the context of TRECVid using mean average precision MAP in order to provide a direct comparison of the effectiveness of the proposed approach with other published work using the same dataset. Precision provides a measure of the ability of a system to present only relevant sequence.

$$\text{AP} = \frac{\left(\frac{\text{number of relevant video sequences retrieved}}{\text{total number of video sequences retrieved}}\right)}{\text{total number of relevant video sequences}} \tag{28}$$

Other metrics are introduced in our evaluation to have a global comparison: F-measure, classification rate CR, and balanced error rate BER.[6] The classifier results can be represented in a confusion matrix (Table 1), where a, b, c and d represent the number of examples falling into each possible outcome:

$$\text{F-measure} = 2\frac{P.R}{P+R} \tag{29}$$

$$\text{BER} = \frac{1}{2}(\frac{b}{a+b} + \frac{c}{c+d}) \tag{30}$$

Figure 9 shows the variation of average precision results *vs* semantic concepts, for three systems: NNET,[7] PENN,[8] and Onto-PENN.[9] First, we observe that PENN

[6] The balanced error rate is the average of the errors on each class. BER is used in "Performance Prediction Challenge Workshop".

[7] NNET: Neural Network based on Evidence Theory.

[8] PENN: Perplexity-based Evidential Neural Network.

[9] Onto-PENN: Ontological readjustment of the PENN. The results presented in the rest of paper for the Onto-PENN, are given by (26) for the semantic similarity computation.

Table 1 Confusion matrix representation

		Prediction	
		Class 0	Class 1
Real	Class 0	a	b
Class	Class 1	c	d

and Onto-PENN systems have the same performance on average for several concepts, and present a significant improvement compared to NNET for the concepts 4,6,17,18,19,23,31 and 32. This is not surprising considering the manner the MAP (Mean Average Precision) is computed (using only the first 2,000 returned shots as in TRECVid) (see Table 2). Furthermore, low performances on several concepts can be observed due to both numerous conflicting classification and limited training data regardless of the fusion system employed. This also explains the rather low retrieval accuracy obtained for concepts 3, 22, 25, 26, 33 and 34.

To evaluate the inter-concepts similarity contribution in the video shots indexing system, we need to study the results in all test set. For this, the comparisons of the detection performances are carried out by thresholding the soft-decisions at the shot-level before and after using the inter-concepts similarity via F-meas, CR$^+$ and BER. Note that the MAP is not sensitive to *Threshold* values τ. Figure 10 compares the three experimental systems along with the variation of $\tau \in [0.1, 0.9]$, by step of 0.1. We can clearly see that for any τ value the Onto-PENN dominates and obtains higher performances for F-meas, CR$^+$ as well as lower BER comparing to PENN and NNET. The BER$_{min}$ = 40.38% is given by $\tau = 0.2$, for F-meas= 16.98% and CR$^+$ = 34.48%. The best results are obtained for $\tau \in [0.2, 0.5]$. With $\tau = 0.40$, the

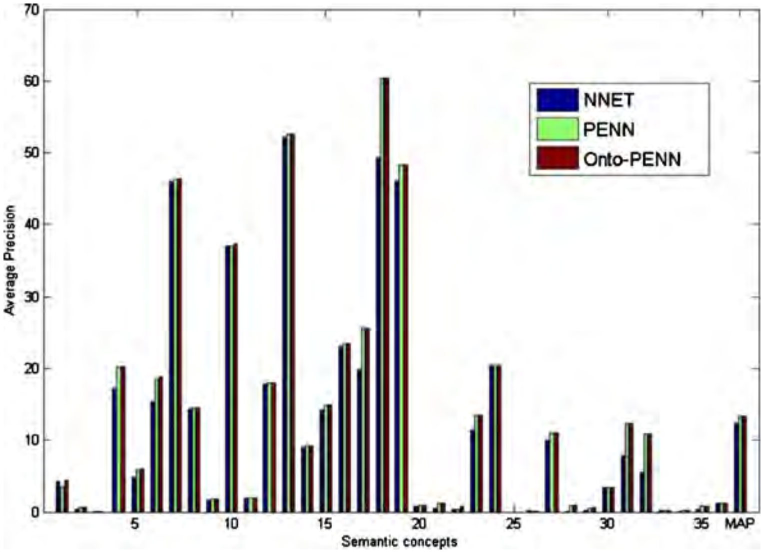


Fig. 9 Average precision evaluation

Table 2 Id of the TRECVid 2007 concepts

Id	Concepts	Neg.train	Pos.train	Pos.test
1	Sports	11,974	106	42
2	Weather	12,029	51	34
3	Court	11,967	113	5
4	Office	11,159	921	453
5	Meeting	11,532	548	270
6	Studio	11,722	358	468
7	Outdoor	8,643	3,437	1,812
8	Building	10,964	1,116	477
9	Desert	12,019	61	15
10	Vegetation	10,615	1,465	499
11	Mountain	12,004	76	17
12	Road	11,420	660	297
13	Sky	10,777	1,303	853
14	Snow	12,044	36	91
15	Urban	10,746	1,334	537
16	Waterscape	11,725	355	414
17	Crowd	11,159	921	552
18	Face	6,596	5,484	2,325
19	Person	4,981	7,099	2,972
20	Pol. security	11,824	256	63
21	Military	11,848	232	74
22	Prisoner	12,067	13	7
23	Animal	11,675	405	271
24	Computer Tv	11,617	463	202
25	US Flag	12,070	10	0
26	Airplane	12,052	28	7
27	Car	11,663	417	187
28	Bus	12,033	47	40
29	Truck	11,985	95	19
30	Boat/ship	11,979	101	151
31	Walk. running	11,221	859	385
32	Peop. marching	11,960	120	82
33	Exp. fire	11,068	12	19
34	Nat. disaster	12,061	19	21
35	Maps	12,030	50	31
36	Charts	11,954	126	80

CR$^+$ is improved by 10.14% to achieve 22.07%, and decreasing the BER of 2.91% compared to NNET.

Figure 11 presents the performance evolution per concepts using $\tau = 0.4$. Some points can be noticed: The three systems produce a certain non-detection (F-meas $=$ 0, CR$^+ = 0$) for the concepts 2, 3, 9, 11, 25, 26, 28, 29, 33, 34, and 36. Then, NNET can not detect any of the following concepts 1, 5, 6, 20, 21, 22, 31, 32, and 35. Identically, for PENN in 5,20,22, and 35. Finally, Onto-PENN resolves the limitation previously mentioned and achieves a high improvement for the concepts 1, 4, 7, 8, 10, 12, 13, 15, 16, 17, 18, 19, 22, 23, 24, and 31, due to the strong relationship between the connected concepts, allowing for better, more accurate decision-making.

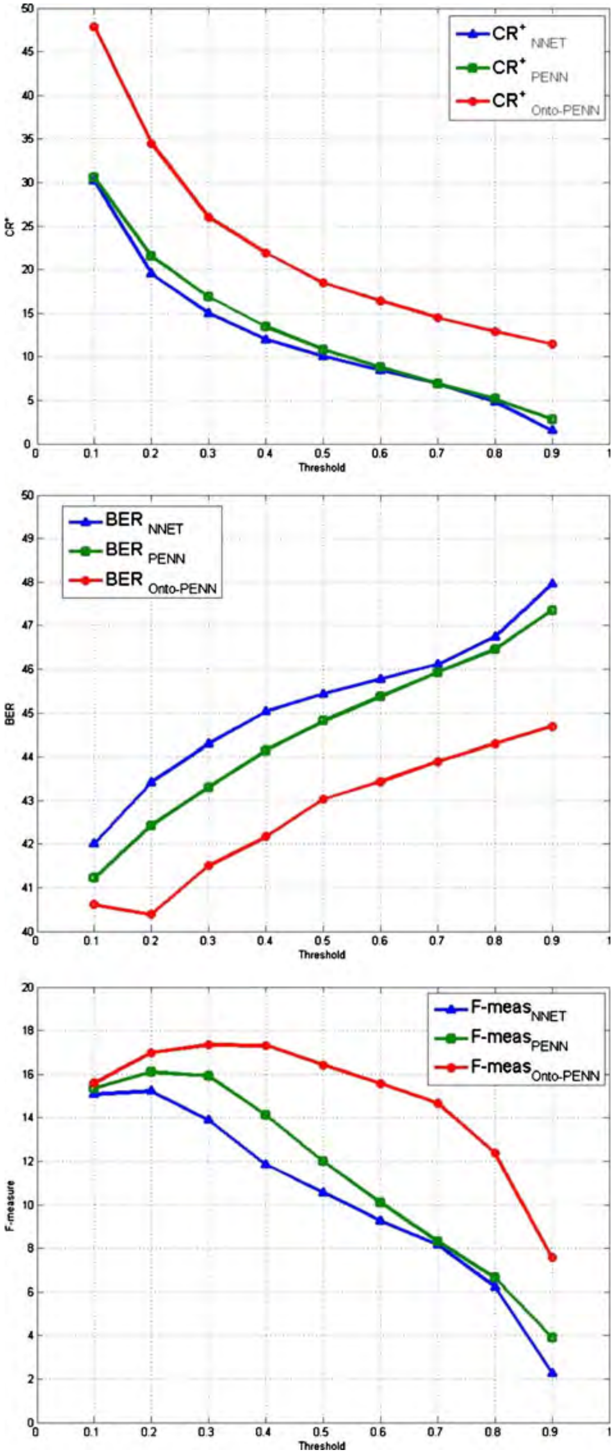


Fig. 10 Evaluation of the metrics (CR$^+$, BER and F-measure) vs *Threshold* $\tau \in [0.1, 0.9]$

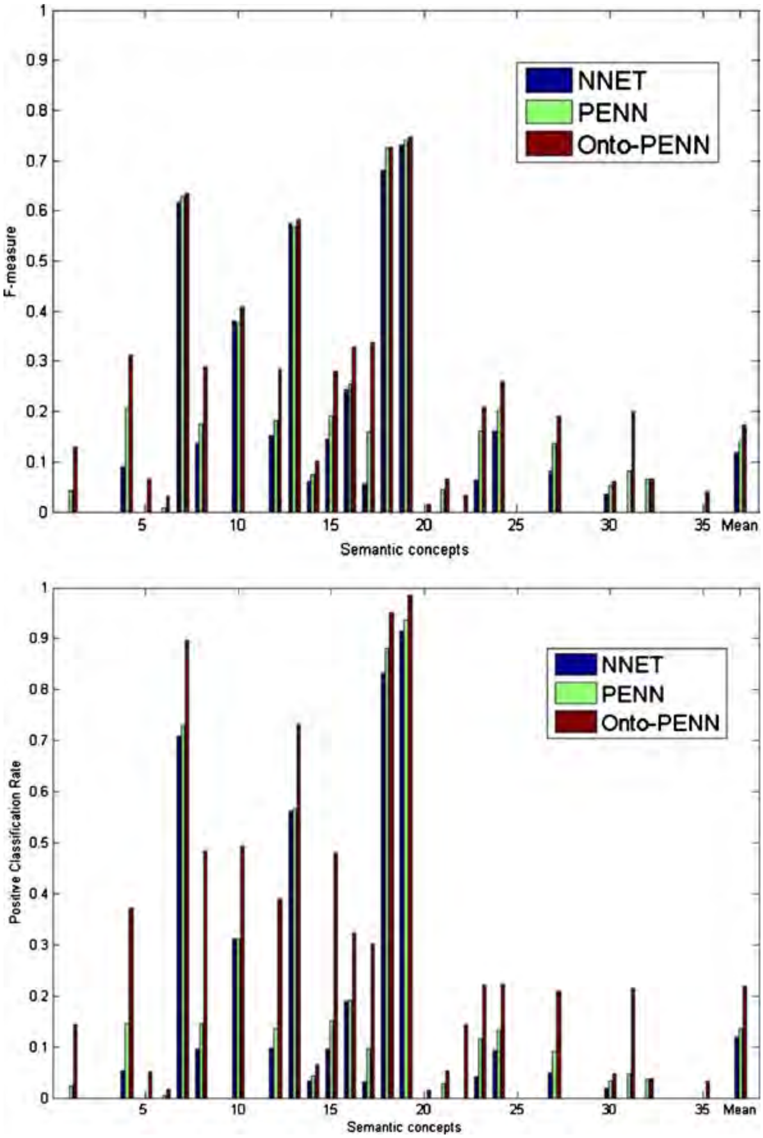


Fig. 11 F-measure and CR$^+$ evaluation

As an example, to detect FACE, PERSON, MEETING, or STUDIO concepts, PENN gives more importance to *FaceDetector, ContourShape, ColorLayout, ScalableColor, EdgeHistogram* than others descriptors. For the “PERSON” concept, the improvement was as high as 11%, making it the best performing run. The Onto-PENN system introduces the relationship between the connected concepts (i.e. concepts that are likely to co-occur in video shots), increasing the performance in term of accuracy (see

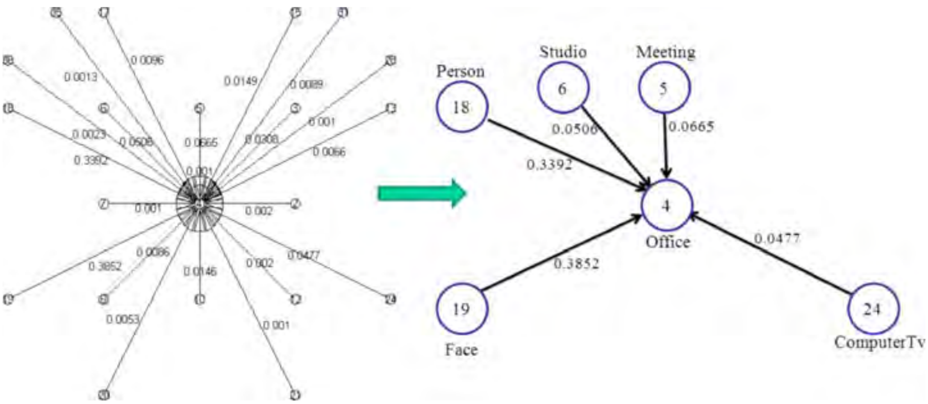


Fig. 12 Inter-concept connections graphical model for the concept OFFICE. We observe that 20 concepts are connected with OFFICE, but only 5 are strong and significant (MEETING: 6.65%, STUDIO: 5.06%, FACE: 33.92%, PERSON: 38.52%, and COMPUTERTV: 4.77%) presenting 88.92% of the global information

Fig. 12). The co-occurring concept constitute some type of contextual information about the content of the shot under consideration.

Table 3 summarizes the overall performances for the content-based video shots classification systems using a fixed $Threshold(\tau = 0.4)$. We compute the above mentioned statistics for all concepts, and for a subset composed of the 10 most frequent concepts in the dataset. All hybrid semantic similarities-based Onto-PENN allow an overall improvement of the system and a significant increase of F-meas and CR^{+}. They achieve a respectable result for MAP, and significantly decrease the balanced error rate "BER" compared to NNET and PENN. Finally, the results given by the two equations (25 and 26) are very close, with a slight advantage for the (26). However, it can be observed that the MAP declines using the equations of Rada,

Table 3 Performance comparisons between the three experimental systems: NNET, PENN and Onto-PENN

Methods/	NNET	PENN	Onto-PENN			
eval. (%)			Rada	Lin	J&C	B&H
MAP	12.70	13.29	12.94	13.01	13.31	13.37
MAP@10	33.70	35.30	34.12	34.91	35.30	35.36
F-meas	11.84	14.10	15.97	16.17	17.07	17.30
F-meas@10	38.75	40.79	41.83	43.41	44.67	44.74
CR^{+}	11.93	13.43	18.12	20.58	21.76	22.07
CR^{+}@10	40.69	41.74	53.76	57.80	59.45	59.71
BER	45.02	44.13	43.93	43.62	42.32	42.11
BER@10	38	36.52	36.02	35.45	34.03	33.96

We present in term of accuracy the effect of each similarity method (Rada (20), Lin (23), J&C (25)), and our proposed method B&H (26) in the Onto-PENN system, for $\tau = 0.4$

and Lin compared to the two equations used, which underlines the importance of the semantic similarity choice.

6 Conclusions

In this paper, we have presented a generic and robust ontology-based video shots indexing scheme. One of the particular aspect of the proposed framework is to employ contextual information during the classification phase. To learn the influence of the relation between concepts, three types of influence are computed: co-occurrence, visual descriptors and hybrid semantic similarity. A comparison of some approaches to automatically construct the semantic similarity has been presented. Based on the newly defined simulated user principle, we evaluate the results of four alternative methodologies. We demonstrate through statistical study and empirical testing the potential of multimodal fusion, to be exploited in video shots retrieval. In TRECVid 2007 benchmark, a significant improvement is obtained with our system, about 18.75% in terms of correct positive recognition rate (CR^+), 5.99% for the F-measure, 1.66% for the mean average precision (MAP), and decreases the balanced error rate of 2.91% on average. Our proposed "Onto-PENN" method outperforms clearly both the NNET and PENN methods which are not using any contextual information. In addition, we have shown that perplexity-based weighted vector integration in the indexing papeline increases the performances of our system.

In the future works, we plan to extend application to WordNet instead of a corpus, integration of richer semantics and broader knowledge.

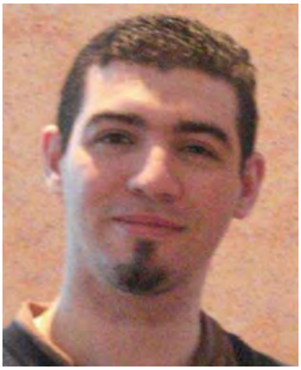
Acknowledgements This research was supported by Eurecom's industrial members: Ascom, Bouygues Telecom, Cegetel, France Telecom, Hitachi, ST Microelectronics, Motorola, Swisscom, Texas Instruments and Thales.

References

1. Adamek T (2007) Extension of MPEG-7 low-level visual descriptors for TRECVid07. Kspace Technical Report, FP6-027026
2. Aigrain P, Joly P (1994) The automatic real-time analysis of film editing and transition effects and its applications. Comput Graph 18(1):93–103
3. Ayache S, Quènot G (2007) TRECVid 2007 collaborative annotation using active learning. In: TRECVid, 11th international workshop on video retrieval evaluation, Gaithersburg, USA
4. Benmokhtar R, Huet B (2006) Classifier fusion: combination methods for semantic indexing in video content. In: International conference on artificial neural networks, pp 65–74
5. Benmokhtar R, Huet B (2007) Neural network combining classifier based on Dempster-Shafer theory for semantic indexing in video content. In: International multimedia modeling conference, pp 196–205
6. Benmokhtar R, Huet B (2008) Perplexity-based evidential neural network classifier fusion using MPEG-7 low-level visual features. In: ACM international conference on multimedia information retrieval, pp 336–341
7. Benmokhtar R, Huet B (2009) Hierarchical ontology-based robust video shots indexing using global MPEG-7 visual descriptors. In: Proceedings of the international workshop on content-based multimedia indexing, pp 195–200
8. Berners T, Hendler J, Lassila O (2001) The semantic web. Scientific American, pp 29–37

9. Chang S-F, Chen W, Meng H, Sundaram H, Zhong D (1998) A fully automated content-based video search engine supporting spatiotemporal queries. In: IEEE transactions circuits and systems for video technology, pp 602–615
10. Denoeux T (1995) An evidence-theoretic neural network classifer. In: International conference on systems, man and cybernetics, vol 31, pp 712–717
11. Dimitrova N (2003) Multimedia content analysis: the next wave. In: International conference on image and video retrieval. Lecture notes in computer science, vol 25, pp 8–17
12. Duin R, Tax D (2000) Experiements with classifier combining rules. In: Proc. first int. workshop MCS 2000, vol 1857, pp 16–29
13. Faloutsos C, Barber R, Flickner M, Hafner J, Niblack W, Petkovic D, Equitz W (1994) Efficient and effective querying by image content. JIIS 3(3), 231–262
14. Fan J, Gao Y, Luo H (2007) Hierarchical classification for automatic image annotation. In: Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, pp 111–118
15. Gao J, Goodman J, Li M, Lee K (2001) Toward a unified approach to statistical language modeling for chinese. In: ACM transactions on Asian language information processing
16. Hauptmann A, Christel M, Concescu R, Gao J, Jin Q, Lin W, Pan J, Stevens S, Yan R, Yang J, Zhang Y (2005) CMU informedia's TRECVid 2005 skirmishes. In: TREC video retrieval evaluation online proceedings
17. ISO/IEC 14496-2. Information Technology (2001) Coding of moving pictures and associated audio information.
18. Jain A, Duin R, Mao J (2000) Statistical pattern recognition: a review. IEEE Trans Pattern Anal Mach Intell 20(1), 4–37
19. Jiang J, Conrath D-W (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In: International conference research on computational linguistics
20. Jiang W, Cotton C, Chang S-F, Ellis D, Loui A (2009) Short-term audio-visual atoms for generic video concept classification. In: MM '09: proceedings of the seventeen ACM international conference on multimedia, pp 5–14
21. Kasutani E, Yamada A (2001) The MPEG-7 color layout descriptor: a compact image feature description for high-speed image/ video retrieval. In: Proceedings of the IEEE international conference on image processing, vol 1, pp 674–677
22. Koskela M, Smeaton A (2006) Clustering-based analysis of semantic concept models for video shots. In: Proceedings of the international conference on multimedia and expo, pp 45–48
23. Koskela M, Smeaton A, Laaksonen J (2007) Measuring concept similarities in multimedia ontologies: analysis and evaluations. IEEE Trans Multimedia 9:912–922
24. Kotsiantis S-B (2007) Supervised machine learning: a review of classification techniques. Informatica 31:249–268
25. Kuncheva L (2003) Fuzzy versus nonfuzzy in combining classifiers designed by bossting. IEEE Trans Fuzzy Syst 11(6),729–741
26. Kuncheva L, Bezdek JC, Duin R (2001) Decision templates for multiple classifier fusion: an experiemental comparaison. Pattern Recogn 34:299–314
27. Laaksonen J, Moskela M, Oja E (2004) Class distributions on SOM surfaces for feature extraction and object retrieval. Neural Netw 17:1121–1133
28. Li B, Goh K (2003) Confidence-based dynamic ensemble for image annotation and semantics discovery. In: Proceedings of the eleventh ACM international conference on multimedia, pp 195–206
29. Li Y, Bandar ZA, Mclean D (2003) An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans Knowl Data Eng 15(4):871–882
30. Lin D (1998) An information-theoretic definition of similarity. In: Proceedings of the 15th international conference on machine learning. Morgan Kaufmann, pp 296–304
31. Manjunath B, Salembier P, Sikora T (2002) Introduction to MPEG-7: multimedia content description interface. Wiley, New York
32. Messing D-S, Beek PV, Errico J-H (2001) The MPEG-7 color structure descriptor: image description using color and local spatial information. In: Proceedings of the IEEE international conference on image processing, vol 1, pp 670–673
33. Naphade MR, Kozintsev I, Huang T (2000) Probabilistic semantic video indexing. In: Proceedings of neural information processing systems, pp 967–973

34. Naphade M, Kristjansson T, Frey B, Huang T (1998) Probabilistic multimedia objects (multijects): a novel approach to video indexing and retrieval in multimedia systems. In: Proceedings of the IEEE international conference on image processing, pp 536–540
35. Naphade M, Kennedy L, Kender J, Chang S, Smith J, Over P, Hauptmann A (2005) A light scale concept ontology for multimedia understanding for TRECVid 2005 (LSCOM-Lite). IBM Research Technical Report
36. Naphade M, Kennedy L, Kender J, Chang S, Smith J, Over P, Hauptmann A (2005) A light scale concept ontology for multimedia understanding for trecvid 2005. IBM Research Technical Report
37. OpenCV (2010) Intelcorporation: open source computer vision library: reference manual. http://opencvlibrary.sourceforge.net
38. Park D, Jeon YS, Won CS (2000) Efficient use of local edge histogram descriptor. In: Proceedings of ACM workshop on multimedia, pp 51–54
39. Pentland A, Picard R, Sclaroff S (1994) Photobook: content-based manipulation of image databases. In: Proceedings of SPIE conference on storage and retrieval for image and video databases
40. Rada R, Mili H, Bicknell E, Blettner M (1989) Development and application of a metric on semantic nets. IEEE Trans Syst Man Cybern 19(1):17–30
41. Resnik P (1995) Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th international joint conference on artificial intelligence, pp 448–453
42. Seco N, Veale T, Hayes J (2004) An intrinsic information content metric for semantic similarity in WordNet. In: Proceedings of European conference on artificial intelligence
43. Slimani T, BenYaghlane B, Mellouli K (2007) Une extension de mesure de similarité entre les concepts d'une ontologie. In: International conference on sciences of electronic, technologies of information and telecommunications, pp 1–10
44. Smith J-R, Chang S-F (1996) VisualSEEk: a fully automated content-based image query system. In: Proceedings of ACM international conference on multimedia, pp 87–98
45. Snoek C-M, Worring M (2005) Multimodal video indexing: a review of the state-of-the-art. Multimedia Tools and Applications 25:5–35
46. Snoek C, Worring M, Geusebroek J-M, Koelma D-C, Seinstra F-J (2004) The mediamill TRECVid 2004 semantic viedo search engine. In: TREC video retrieval evaluation online proceedings
47. Souvannavong F (2005) Indexation et recherche de plans vidéo par le contenu sémantique. PhD thesis, Eurécom, France
48. Sun X, Manjunath B, Divakaran A (2002) Representation of motion activity in hierarchical levels for video indexing and filtering. In: Proceedings of the IEEE international conference on image processing, pp 149–152
49. TRECVID (2010) Digital video retrieval at NIST. http://www-nlpir.nist.gov/projects/trecvid/
50. Tsinaraki C, Polydoros P, Christodoulakis S (2004) Interoperability support for ontology-based video retrieval applications. In: Proceedings of the third international conference on image and video retrieval
51. Vapnik V (2000) The nature of statistical learning theory. Springer, New York
52. Vembu S, Kiesel M, Sintek M, Baumann S (2006) Towards bridging the semantic gap in multimedia annotation and retrieval. In: Proceedings of the 1st international workshop on semantic web annotations for multimedia
53. Wactlar H, Kanade T, Smith MA, Stevens SM (1996) Intelligent access to digital video: the informedia project. In: IEEE computer, vol 29
54. Wu Z, Palmer M (1994) Verbs semantics and lexical selection. In: Proceedings of the 32nd annual meeting on association for computational linguistics, pp 133–138
55. Wu Y, Tseng B, Smith J (2004) Ontology-based multi-classification learning for video concept detection. In: Proceedings of the international conference on multimedia and expo, vol 2, pp 1003–1006
56. Xu F, Zhang Y (2006) Evaluation and comparison of texture descriptors proposed in MPEG-7. J Vis Commun Image Represent 17:701–716
57. Xu L, Krzyzak A, Suen C (1992) Methods of combining multiple classifiers and their application to hardwriting recognition. IEEE Trans Syst Man Cybern 22:418–435
58. Yining D, Manjunath B (1998) Netra-V: toward an object-based video representation. In: Proceedings of IEEE conference of multimedia and expo, vol 8, no 5, pp 616–627



Rachid Benmokhtar received the engineering degree in 2004 from USTHB University of Algiers-Algeria, and the Master diploma in June 2005 from University of technology of Compiègne, France. In October 2005, he joined the Multimedia Communications Department at Eurecom to study toward the Ph.D. degree under the supervision of *Benoit Huet* and *Bernard Mérialdo*. He defended his Ph.D thesis in June 2009, with a Very Honorable mention. During 2009–2011, he worked with *Ivan Laptev* as a research engineer under the Quaero project at Willow-TexMex INRIA research teams. In June 2011, he moved to TexMex-INRIA of Rennes in the *Hervé Jégou* group. His research interests include multimedia indexing, content-based video retrieval, object and scene recognition from video and still images, event detection, tracking, multi-level fusion, classification, and machine learning.

Benoit Huet received his BSc degree in computer science and engineering from the Ecole Superieure de Technologie Electrique (Groupe ESIEE, France) in 1992. In 1993, he was awarded the MSc degree in Artificial Intelligence from the University of Westminster (UK) with distinction, where he then spent two years working as a research and teaching assistant. He received his DPhil degree in Computer Science from the University of York (UK) for his research on the topic of object recognition from large databases. He is currently working as a research and teaching assistant in the multimedia information processing group of the Eurecom (France). His research interests include computer vision, content-based retrieval, multimedia data mining and indexing (still and/or moving images) and pattern recognition. He has published over 80 papers in journals, edited books and refereed conferences. He is a member of IEEE, ACM and ISIF. He has served in many international conference organization and technical program committee. He is regularly invited to serves as reviewer for prestigious scientific journals as well as expert for project proposal at national, European and International level.

Concept Detector Refinement Using Social Videos

Xueliang Liu
EURECOM France
2229 Route Des Cretes
Sophia Antipolis, France
xueliang.liu@eurecom.fr

Benoit Huet
EURECOM France
2229 Route Des Cretes
Sophia Antipolis, France
benoit.huet@eurecom.fr

ABSTRACT

The explosion of social video sharing sites gives new challenges on video search and indexing techniques. Because of the concept diversity in social videos, it is very hard to build a well annotated dataset that provides good coverage over the whole meaning of concepts. However, the prosperity of social videos on the internet also make it easy to obtain a huge number of videos, which gives an opportunity to mine the semantic content from an infinite amount of video entities. In this paper, we focus on improving the performance concept detectors and propose a refinement framework based on a semi-supervised learning technique. In our framework, the self-training algorithm is employed to expand the training dataset with automatically labeled data. The contribution of this paper is to demonstrate how to utilize the visual feature and text metadata to enhance the performance of concept classifier with a lot number of unlabeled videos. By experimenting on a social video dataset with 21,000 entities, it is shown that after expanding the training set with automatically labeled shots, the concept detectors' performance can be significantly improved.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Measure,Performance,Experimentation

Keywords

Social Video, Semantic Analysis, Semi-supervised Learning

1. INTRODUCTION

With the advances of digital capture equipment and multimedia storage, the recent explosion in video shared web technologies make it possible to upload the videos by the web users. The last five years have witnessed rapidly growing popularity of social video sites, such as YouTube[4], DailyMotion[1]. It results in the explosion of social video documents, with diverse concepts and sparse text descriptions.

The prosperity of social video gives new challenge on the traditional text-based search engine, though they have gained a remarkable success on text retrieval on the internet. Nowadays video search engine still adopt standard text retrieval technologies to index and search social videos according the accompanied metadata, such as tags, description, comments. This is an efficient way for video query but cannot index video data semantically unless video data are well annotated by hand. Obviously it is becoming an emergency to develop video search techniques that can mine the semantic concept without requiring extensive manual labeling.

In parallel, semantic video analysis techniques based on statistical models have made great progress in recent years. These research are currently focusing on the analysis and mining the visual content of video by modeling the low level features extracted from video shots. These learning based techniques succeed in traditional video but face new challenges on dealing with social video. Due to the infinite amount of video and the diversity of concepts, it is very hard to build a well labeled dataset with a good semantic concept coverage for training.

Both the text-based technique and the visual content modeling approach have their strengths and limitations on video indexing. It gives possibility to fuse these two kinds of features together for better semantic analysis. In this paper, we aim to integrate both the text and visual feature of social video entities to improve the performance of concepts detection, and propose a semi-supervised learning based framework to obtain a group of concept detectors with better coverage though exploitation of unlabeled data.

The organization of the rest of this paper is as followed. In Section 2, we provide a brief review on the related work. In Section 3, we introduce our refinement framework based on visual feature and tags. In Section 4, we demonstrate the experiments and show the results. Finally, we conclude the paper in Section 5.

2. RELATED WORK

In recent years, there already have been lots of studies on video concept analysis thrust of video storage and machine learning techniques. To investigate the problem, the multimedia community has built many benchmark dataset, such as TrecVID[3], Caltech[10]. Semantic video analysis has traditionally involved these known datasets with fixed

VLS-MCMR'10, October 29, 2010, Firenze, Italy.
Copyright 2010 ACM 978-1-4503-0166-4/10/10 ...$10.00.

and limited sets of keywords and semantic concepts. Based on those well annotated datasets, many related framework are proposed with the advance of machine learning algorithm. A straightforward way to achieve video analysis is to adapt the image search and indexing techniques directly. In those approaches, the visual feature obtained from shots keyframes are employed to model the concept underlying in the video contents. For example, the authors of [9] have modeled image keywords using a multiple Bernoulli distribution for image annotation. To apply their method on videos they simply build their model for visual features within rectangular regions of the keyframes of a video and achieve better results.

Instead of adapted image indexing techniques, there are some works much more specifically designed for videos, which focus on mining on audio and spatio-temporal properties of the video. Motivated by the success of SIFT feature in image indexing, some similar work to represent shots with a spatio-temporal feature have been done[17, 13]. The authors of [17] proposed a local space-time features to capture local events. This technique have been shown effectiveness on people event recognition, such as people running or jumping. In [13], the authors studied a shape-based feature for event recognition in crowded videos. To exploit the power of audio feature for video concept detection, Jiang et. al [12] investigated the joint audio and visual analysis for semantic concept detection, and propose a novel visual feature with background audio representation to improve concept detection.

Recently, due to social sharing website such as YouTube[4] , Flickr[2], where a large number of multimedia documents are available benefited from the contribution of web users, it has become possible to attempt multimedia search and indexing on a large size of dataset. However, working on real world web datasets provides new challenge on the traditional video concept detection techniques. Clearly, it is very time-consuming and tedious work to build a well labeled dataset for the training purpose. To address the problem of semantic web video analysis, some large scale datasets s have been built using multimedia data crawler from shared portals [8, 6]. Beside those web video datasets built very recently, a number of research works in the image domain have shown acceptable results by investigation on semi-supervised learning techniques[19, 11]. Similarly to the work as we propose in this paper, the authors of [18] have developed an active learning based concept classifier refinement system on a large scale of image dataset, the concept classifiers can be reinforced by updating the positive and negative training samples iteratively, but it still need users to review the output of the classifier. In addition, [14] have proposed an automatically collected dataset using the huge resources of web by object recognition techniques. However, the authors did not take the rich information in metadata into consideration. All of these work inspire our research on video dataset annotation.

In this paper, we report an automatic framework based on semi-supervised learning for concept refinement and expect a similar result on social videos. Our aim is to leverage on the information provided by authors and users rather than requiring for extra human annotation efforts.

3. OUR PROPOSED SCHEME

Visual features such as SIFT, color moments, are commonly used in previous research [9, 17, 13]. They are thought as the representative feature to video content. In social video shared site, the videos entities are always uploaded along with some text metadata. Though these metadata are labeled manually by the video user arbitrarily, and are not always accurate, these text also give a rough indication to the real concept behind the video shots. They are therefore of importance for video content representation. The proper combination of textual and visual feature can boost the performance of video analysis technique greatly. In this paper, we take a web video entity as composition of a group of shots along with a tag set, and focus on refine the concept detectors by exploration into unlabeled pools based on semi-supervised self-training approach.

3.1 Refinement with Self-training Strategy

At first, let us introduce the general semantic video concept analysis problem briefly. It is the process by which a computer system automatically assigns caption or keywords to a digital video shot, which can often be regarded as a classification problem. Suppose we have a well annotated dataset $X = \{x_i\}$ along with its label $Y = \{y_i\}$. Our goal is to find out a group of classifiers

$$\mathcal{F} = \{f_i | f_i : X \to Y\} \tag{1}$$

Their parameters $\{\lambda_i\}$ can be obtained from

$$\{\lambda_i\}^* = \arg\max_{\lambda_i} P(Y|X, \{\lambda_i\}) \tag{2}$$

For the video annotation problems, the dataset X can be the shot, which is represented by the visual feature vectors, and the label Y is the concept to be annotated. Associated with each concept, there is a model f_i to compute the probability with which a shot belongs to this concept. In order to obtain the prediction model, lots of data should be labeled well for the training process. However, it is unaffordable to annotate a large-scale video corpus with a good coverage over the whole meaning of concepts because of the work intensity and time consumption. On another side, it becomes easier to obtain a huge number of unlabeled data with the booming of video sharing website. This makes it possible to capture more underlying meaning of the concepts. There are also some related work done with a semi-supervised learning framework to mine semantic meaning among data pool without labels . Semi-supervised learning is a group of algorithms that make use of the labeled and unlabeled data. The one we used in this work is called self-training[20]. Besides the labeled dataset X, other dataset U containing unlabeled videos is available. In self-training, the classifiers are firstly trained from the small amount of labeled data X as shown in Equation 1, then used to predict label for the unlabeled data, and the most likely unlabeled data items are added to the training set.

$$X^* = \{x | \max_i(P(x|f_i)) > \theta, x \in U\} \tag{3}$$

And

$$X^{'} = X + X^* \tag{4}$$

The classifiers are then re-trained on the extended training set $X^{'}$ with the same method as Equation 2.

$$\{\lambda_i^{'}\}* = \arg\max_{\lambda_i^{'}} P(Y, Y^*|X^{'}, \{\lambda_i^{'}\}) \tag{5}$$

Where Y^* is the label of X^* predicted by model $\mathcal{F}$, and $\lambda_i^{'}$ is the parameters of updated model $\mathcal{F}^{'}$.

For social video, the useful information we can explore for extending the labeled training set automatically are visual features and textual metadata. We consider them for our refinement algorithm as follows.

3.2 Visual Feature Based Refinement

In most of previous annotation research, visual features are used to represent the content of video shot. The intuitional approach of refinement is to utilize the visual feature directly. As shown in Figure 1, we first initiate the training of the concept detectors with the manually labeled subset. Then, those newly trained detectors are run on the unlabeled video collection to predict labels. The video shots that have a high similarity to a concept, in other words, when the probability estimation of the concept detector exceeds a given threshold, will be added to the training set. The concept detectors are then re-trained on the automatically extended training set. Both the original concept detector and the re-trained ones are then evaluated on the testing dataset which has been held out for performance evaluation only.

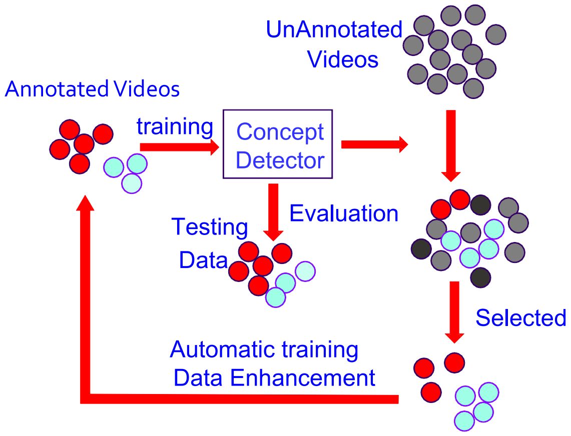


Figure 1: Visual Feature Refinement

3.3 Tags-Based Visual Supervised Refinement

Compared with traditional videos, social videos are commonly accompanied with metadata such as tags, description, script, etc…, which are uploaded by the users themselves. Though the textual information is usually erroneous and sparse, and not accurate enough to provide the required knowledge for effective content-based retrieval, the analysis of the auxiliary text shows possibility of improving the performance of traditional multimedia information analysis approaches. Additionally, when correctly tagged by their authors or other contributors, such information could really benefit concept modeling.

Before we utilize the text metadata for semantic analysis, there are some problems that we need to be dealt with. In video annotation, the concepts should be labeled on each shot. However, the web video tags are given to the whole video entity, so we cannot use the tags as a kind of weak labels directly. Additionally, the synonymy and polysemy problems make it more complicated. Synonymy is used to describe the fact that there are many ways to refer to the same object and polysemy means that most words have more than one distinct meaning. Considering the synonymy and polysemy, it is hard to mine the meaning from so brief and sparse text description. For example, the video shots tagged with “boat” and “ship” should be annotated with the same concept, and if a video shot is tagged as “Apple”, the user’s intuition may be “a kind of fruit”, “A kind of electrical device”, or “a famous company”. In *nature language processing*(NLP) there are some techniques that can be used to solve the synonymy[16]. The method we use here is query expanding: we expand the concept with keywords that have similar meaning in semantic level. In case of polysemy, it should be noticed that for each word, each different meaning has a different visual appearance, and therefore will lead to a special distribution in feature space.

With this in mind, we propose our tag-based visual supervised (TBVS) refinement framework as shown in Figure 2. We query with keywords for each concept from our dataset, and initialize the annotation of all the *Shots* with such concept for each returned video entity. Then we use the same strategy as Section 3.2. A group of trained visual concept detectors are run on those shots, and sort the result by visual similarity. Those whose probability exceed a given threshold are reserved to the training set.

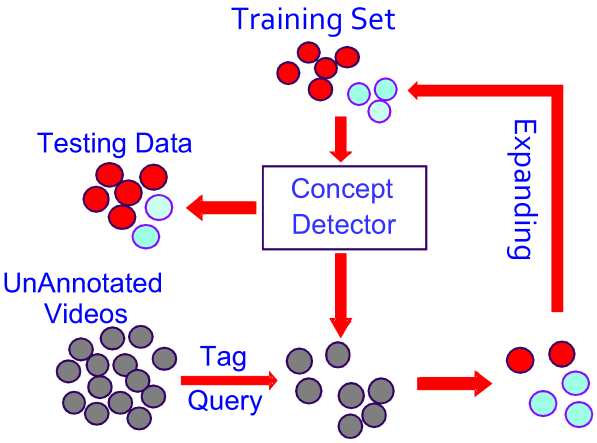


Figure 2: Tags-Based Visual Supervised Refinement

Though it seems that similar strategies are used in the two refinement methods, they are essentially quite different from each other. For visual feature refinement, only visual similarity is taken into consideration. However for the tag-based visual supervised refinement, we obtain the expanding video shots from tag query result, and it is more possible to get relevant shots by semantic meaning.

In both refinement approaches, obviously the selection metric chosen is crucial. This issue will be studied further in Section 4.3.

4. EXPERIMENTS

To validate our proposed approach on social videos, we use a social video dataset and conduct a group of experiments.

4.1 Dataset

A well designed dataset is very important for our concept detector refinement problems. Here we use a subset of our dataset [15] built by the data crawled from YouTube[4].

The subset used in the experiments contains about 21,000 videos along with their text metadata. All of the videos are segmented into shots and a keyframe is extracted for each shot. This results in 240,000 shots and a text corpus with 300 keywords, built from the video tags and titles, upon which we will evaluate the proposed approach. For each keyframe, multiple types of low level visual feature (64-D color histogram, 225-D color moment, 250-D Bag-Of-Words) are computed. For the sake of simplification, only the 225-D color moment feature, which has have been shown efficient and effective in generic concept detection [5], is used in the experiments.

Besides the visual feature, we also build a keywords dictionary from the text metadata along the videos. From video title and tags we obtain 562K textual words. We sort them by frequency after removing the stop-word and words stemming. We also remove some meaningless words such as "video", "music" manually and reserve the top 300 words as our keywords corpus.

In this experiments, we manually choose five visual concept: *Airplane*, *Animal*, *Boat_Ship*, *Person*, *Snow*, which have recognizable appearance and good distribution in our dataset. We expand those concept semantically with the keywords in our corpus for synonymy as shown in Table 1.

Table 1: Concept Expanding

Concept	Keywords
Airplane	Airplane, Flight
Animal	Animal, Dog, Tiger,Lion
Boat_Ship	Boat, Ship
Person	Person, People, Girl, Boy
Snow	Snow

4.2 Learning Process and Evaluation

The self-learning is a practical wrapper approach, there still need a baseline machine learning algorithm. Here *support vector machine* (SVM) is employed for the learning process. SVM is an effective method to solve binary-class or multi-class classification problems. A classification problem is considered on a given a set of labeled training data $(\vec{x_i}, y_i)$ where samples $\vec{x_i} \in R^d$ and binary labels are given as $y_i \in$ {1,-1} for binary-class problems and $y_i \in Z$ for multi-class problems. In the case of multimedia information retrieval, we can consider R^d the d-dimensional space of low-level visual features so that each image or video has a unique feature vector descriptor. The labels y_i are used to indicate which concept examples are relevant with. The solution of SVM is to construct a hyperplane or set of hyperplane in a high feature dimensional space, which can be used for classification, as well as regression or other tasks. The hyperplanes have the largest distance to the nearest training data points of any class and make the classification error of the classifier to be lower. The implementation used here is the latest LIBSVM [7] with a *Radial Basis Function* (RBF) kernel. We use the cross-validation methods to determinate the parameters in the SVM models.

In this experiment, the *average precision* (AP) and *mean average precision* (MAP) are used as criteria to measure the performance. AP is a standard performance measure for image and video semantic concept search and indexing, which can be calculated by

$$AP = \frac{\sum_r Precision(r) \times Relevance(r)}{Number of Relevant Documents} \qquad (6)$$

Here $Precision(r)$ means the precision of rank r, and the $Relevance(r)$ equals to 1 if documents at rank r is relevant and equals to 0 for else. The value of AP is almost the same as the area under the un-interpolated precision-recall curve. And MAP is the arithmetic mean of average precision values across all of the concepts.

4.3 Parameters Setting

The key issue in self-learning is how to find the proper metric to decide which examples to add to the training set. In our refinement process, we use a threshold to decide the amount of new adding shots for simplification. It is obvious that the threshold plays a crucial role in this model. On one hand,a high value threshold will lead to fewer shots reserved and the training set are still far to reach a good coverage in the feature space, which will lead to the new trained concept detectors' performance will not improved much. On the other hand, it should be noticed that classifier performances can be degraded if there are many incorrectly labeled sample in the new training set. If the threshold is too small, more shots will be inserted to the training set, and the number of shots labeled incorrectly will increase, This will contaminate the training set with potentially noisy data and directly bring down the performance of concept detector in the subsequent training process. Figure 3 shows the percent of reserved shots in both refinements strategy. From this figure, we can see that with the increasing of threshold, the percent of shots used for expanding decreases.

For the two process, we use different threshold because of two reasons. First, we know that there are some kind of underlying truth in tags labeled by users, so we can forecast that in the tag-based refinement it is no need to use a threshold with the same value as visual feature refinement. Secondly, because of the sparsity of keywords in text metadata, there is no large amount of shots return queried by keywords. So a high probability threshold value will block too many shots into the refinement process. In our experiment, we find the optimal threshold of 0.92 for visual feature based refinement and 0.72 for tag based refinement can achieve a better result.

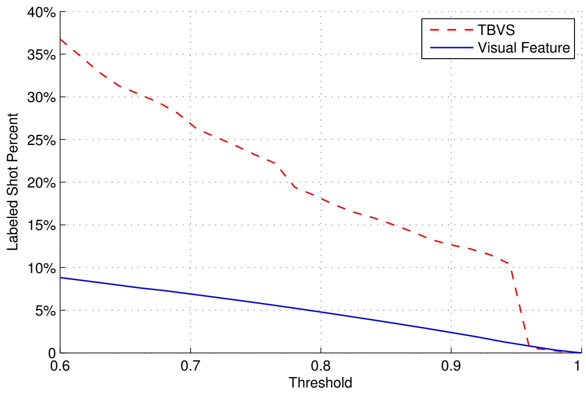


Figure 3: retrieved shots with respect to selection threshold

4.4 Results

To validate our methods, a group of experiments are peformed in our dataset: a) training with annotated shots; b) training with on all shots from tag query result; c) refinement based on visual feature; d) refinement based on tags query and visual supervision. All of the detectors are tested on the same data that are labeled well.

Figure 4 gives the detector performance measure result on the experiments. From the figure, we can see that detectors trained on the data queried by tags gain the worst performance in all of the concept, as we expect, because of the noise among user tags. Compared with the performance of classifiers trained on labeled data, both the visual feature refinement and tag-based visual supervised refinement achieve better results. In visual feature refinement, the detection accuracy is improved when new shots are added automatically through self-learning scheme for most of concept. Significant AP gains are achieved for “Boat-Ship” by 43.1%, “Person” by 28.5%, “Airplane” by 19.0%. The overall MAP is improved by 21.7% after a single iteration.

Figure 4 also show the remarkable improvement on tag-based visual supervised refinement. Similar to visual feature refinement, this group of concept detectors is also enhanced by coming of the new shots. With an overall MAP improved by 23.5%, concept detectors also gained significant advance, such as “Boat-Ship” by 53.5%, “Airplane” by 27.7%, and “Person” by 17.7% respectively.

In Figure 4, it can be observed that for some concepts the visual refinement approach performs better than the tag refinement one. However, there are also concepts for which the tag refinement works best. Furthermore, the mean average precision over all concepts studied in this work shows a small advantage for tag refinement. Nonetheless, the intrinsic distribution of both the training data and the new automatically selected training samples should be studied further in order to identify which how to choose between tag based and visual based refinement.

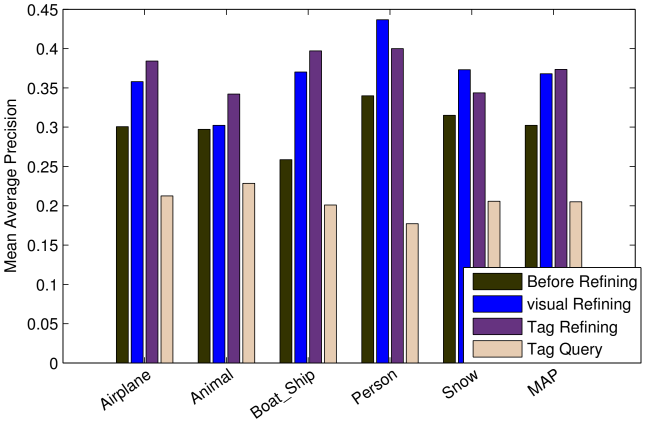


Figure 4: The AP of concept detectors

In general, we observe that the concept coverage in this dataset can be enhanced by the automatically annotated data, and concept detector’s performance achieve a remarkable improvement with the two refinement processes. Further more, we can see from the results that although the video tags are sparse and erroneous for a single video, there are indeed semantic truth in a groups of shots. As shown in Figure 5 even though less expanding shots are added into the training data in the tag-based visual supervised refinement, the two refinement process performs almost the same.

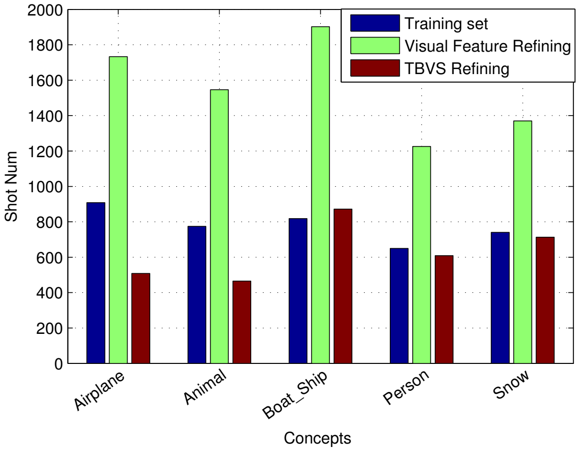


Figure 5: Shot Number for training and refinement

5. CONCLUSIONS

As the amount of social video content available online continues to increase exponentially everyday, there is an immediate need for automatic tools for semantic video search and indexing. While the multimedia concept detectors require large amount of samples to learn models, it is very hard to provide a well labeled datasets because of the infinite and diversity of the concepts. In this paper, we focus on improving the performance of concept detection with a semi-supervised learning process. The proposed refinement framework utilize both the visual feature and text meta data and dynamically extend the set of training examples available for learning.Using a large dataset of 21,000 videos, we have shown the ability of our proposed approach to enhance the performance of semantic detectors to accurately identify content.

In the future, we will consider how to build concept detectors in a fully automated fashion. In the refinement framework proposed in the paper, pre-trained models are necessary to initiate the refinement process. We intend to study the possibility to analyze and mine from online shared multimedia content in order to learn semantic concepts automatically from web documents and analyze them semantically in our future work.

6. REFERENCES

[1] *DailyMotion*. `http://www.dailymotion.fr`.

[2] *Flickr*. `http://www.flickr.com/`.

[3] *TrecVID*. `http://trecvid.nist.gov/`.

[4] *YouTube*. `http://www.youtube.com`.

[5] A. Amir, J. O. Argillander, M. Berg, and et al. IBM research TRECVID-2004 video retrieval system. In

NIST TRECVID 2004 Workshop, Gaithersburg, MD, 2004.

[6] J. Cao, Y. Zhang, Y. Song, Z. Chen, X. Zhang, and J. Li. MCG-WEBV: A benchmark dataset for web video analysis. Technical report, May. 2009.

[7] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[8] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. NUS-WIDE: A real-world web image database from national university of singapore. In *Proc. of ACM Conf. on Image and Video Retrieval*, Santorini, Greece.

[9] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1002–1009, 2004.

[10] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical report, California Institute of Technology, 2007.

[11] R. Hong, G. Li, L. Nie, J. Tang, and T.-S. Chua. Explore large scale data for multimedia QA. In *ACM conference on Image and Video Retrieval*, Xi'an, China, 2010.

[12] W. Jiang, C. Cotton, S.-F. Chang, D. Ellis, and A. C. Loui. Short-term audio-visual atoms for generic video concept classification. In *Proceeding of ACM international conference on Multimedia (ACM MM)*, October 2009.

[13] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *International Conference on Computer Vision*, 2007.

[14] L.-J. Li, G. Wang, and L. Fei-Fei. Optimol: automatic online picture collection via incremental model learning. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 0:1–8, 2007.

[15] X. Liu and B. Huet. Automatic concept detector refinement for large-scale video semantic annotation. In *IEEE International Conference on Semantic Computing*, Pittsburgh, USA, September 2010.

[16] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 1 edition, July 2008.

[17] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *International proceeding of Pattern Recognition*, 2004.

[18] M. yu Chen, M. Christel, E. Hauptmann, and H. Wactlar. Putting active learning into multimedia applications: Dynamic definition and refinement of concept classifiers. In *Proceedings of ACM Multimedia*, pages 902–911. ACM Press, 2005.

[19] Z.-J. Zha, T. Mei, J. Wang, Z. Wang, and X.-S. Hua. Graph-based semi-supervised learning with multi-label. *ACM Trans. Program. Lang. Syst.*, 20(5):97–103, 2009.

[20] X. Zhu. Semi-supervised learning literature survey. Technical report, CMU, 2006.

Finding Media Illustrating Events

Xueliang Liu
EURECOM
Sophia Antipolis, France
xueliang.liu@eurecom.fr

Raphaël Troncy
EURECOM
Sophia Antipolis, France
raphael.troncy@eurecom.fr

Benoit Huet
EURECOM
Sophia Antipolis, France
benoit.huet@eurecom.fr

ABSTRACT

We present a method combining semantic inferencing and visual analysis for finding automatically media (photos and videos) illustrating events. We report on experiments validating our heuristic for mining media sharing platforms and large event directories in order to mutually enrich the descriptions of the content they host. Our overall goal is to design a web-based environment that allows users to explore and select events, to inspect associated media, and to discover meaningful, surprising or entertaining connections between events, media and people participating in events. We present a large dataset composed of semantic descriptions of events, photos and videos interlinked with the larger Linked Open Data cloud and we show the benefits of using semantic web technologies for integrating multimedia metadata.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information System**]: Audio, Video and Hypertext Interactive Systems; I.7.2 [**Document Preparation**]: Languages and systems, Markup languages, Multi/mixed media, Standards

General Terms

Languages, Hyperlinks, Web, URI, HTTP

Keywords

Events, LODE, media ontology, multimedia semantics

1. INTRODUCTION

Events are a natural way for referring to any observable occurrence grouping persons, places, times and activities that can be described [16]. Events are also observable experiences that are often documented by people through different media (e.g. videos and photos). We explore this intrinsic connection between media and experiences so that people can search and browse through content using a familiar event perspective. We are aware that web sites already exist that provide interfaces to such functionality, e.g. eventful.com, upcoming.org, last.fm/events, and facebook.com/events to name a few. These services have sometimes explicit connection with media sharing platforms, have often overlap in terms of coverage of upcoming events and provide social networks features to support users in sharing and deciding upon attending events. However, the information about the events, the social connections and the representative media are all spread and locked in amongst these services providing limited event coverage and no interoperability of the description [5].

Our goal is to aggregate these heterogeneous sources of information using linked data, so that we can explore the information with the flexibility and depth afforded by semantic web technologies. Furthermore, we investigate the underlying connections between events to allow users to discover meaningful, entertaining or surprising relationships amongst them. We also use these connections as means of providing information and illustrations about future events, thus enhancing decision support. In this paper, we present a method for finding automatically medias hosted on Flickr and YouTube that can be associated to a public event. We show the benefits of using linked data technologies for enriching semantically the descriptions of both events and media.

The remaining of this paper is structured as follow. In Section 2, we briefly describe the LODE event model and how we scrap large event directories. In Section 3, we present the dataset on which we will evaluate our method. We then detail our approach for associating media with events (Section 4). We discuss our results in Section 5 and present some related work in Section 6. Finally, we give our conclusions and outline future work in Section 7.

2. LODE AND EVENT DIRECTORIES

Large numbers of web sites contain information about scheduled events, of which some may display media captured at these events. This information is, however, often incomplete and always locked into the sites. In previous research, we carried out user studies in order to collect end-user experiences, opinions and interests while discovering, attending and sharing events, and user insights about potential web-based technologies that support these activities. The results of this study support the development of an environment that merges event directories, social networks and media sharing platforms [5]. We argue that linked data technologies is suitable for doing this integration at large scale given they naturally based on URIs for identifying objects

ICMR '11, April 17-20, Trento, Italy

and a simple triple model (RDF) for representing semantic descriptions. In this section, we present the LODE event model and how we populate this ontology by scraping three large event directories: last.fm, eventful and upcoming.

2.1 LODE by Example

The LODE ontology[1] is a minimal model that encapsulates the most useful properties for describing events [11]. The goal of this ontology is to enable interoperable modeling of the "factual" aspects of events, where these can be characterized in terms of the *four Ws*: *What* happened, *Where* did it happen, *When* did it happen, and *Who* was involved. LODE is not yet another "event" ontology *per se.* It has been designed as an *interlingua* model that solves an interoperability problem by providing a set of axioms expressing mappings between existing event ontologies. Hence, the ontology contains numerous OWL axioms stating classes and properties equivalence between models such as the Event Ontology [10], CIDOC-CRM, DOLCE, SEM [15] to name a few.

Figure 1 depicts the metadata attached to the event identified by `350591` on last.fm according to the LODE ontology. More precisely, it indicates that an event of type `Concert` has been given on the `13th of July 2007 at 20:30 PM` in the `Nouveau Casino` theater in Paris featuring the Irish singer `Róisín Murphy` known for electronic style.

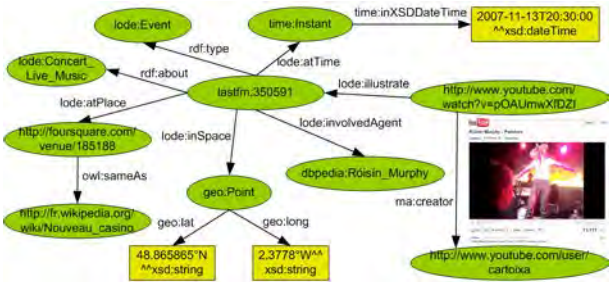


Figure 1: *Róisín Murphy at the Nouveau Casino in Paris* described with LODE

2.2 Scraping Event Directories

We use the Last.fm, Eventful and Upcoming APIs to convert each event description into the LODE ontology. We mint new URIs into our own namespace for events (`http://data.linkedevents.org/event/`), agents (`http://data.linkedevents.org/agent/`) and locations (`http://data.linkedevents.org/location/`). A graph representing an event is composed of the type of the event, a full text description, the agents (e.g. artists) involved, a date (instant or interval represented with OWL Time [7]), a location in terms of both geographical coordinates and a URI denoting the venue and users participation. A graph representing an agent or a location is composed of a label and a description (e.g. the artist's biography).

Event directories have overlap in their coverage. We interlink these events descriptions when they involve the same agents at the same date or when they happen at the same venue at the same date. We invoke additional semantic web lookup services such as dbpedia and freebase, or foursquare and geonames in order to enrich the descriptions of the agents and the locations. Hence, the venue has been converted into a foursquare URI (`http://foursquare.com/venue/185188`) that provides additional information such as the number of different users that have *check in* at this place and the current virual mayor while the wikipedia URI (`http://fr.wikipedia.org/wiki/Nouveau_casino`) provides the history of this venue in French.

The agent URI, which has for label "Róisín Murphy" has also been interlinked with the dbpedia URI (`http://dbpedia.org/resource/RóisínMurphy`) which provides additional information about the solo singer such as its complete discography. This URI is declared to be `owl:sameAs` another identifier from Freebase (`http://rdf.freebase.com/ns/guid.9202a8c04000641\-f80000000004a1685`) which provides information about the 2 bands she has been part of. The linked data journey can be rich and long. One of the challenges we want to address is how to visualize these enriched interconnected datasets while still supporting simple user tasks such as searching and browsing enriched media collections.

3. DATASET

Explicit relationships between scheduled events and photos hosted on Flickr can be looked up using special machine tags such as `lastfm:event=XXX` or `upcoming:event=XXX`. In a previous work, we explored the overlap in metadata between four popular web sites, namely Flickr as a hosting web site for photos and Last.fm, Eventful and Upcoming as a documentation of past and upcoming events [14]. Hence, we have been able to convert the description of more than 1.7 million photos which are indexed by nearly 110.000 events (Table 1).

	Event	Agent	Location	Photos	User
`Last.fm`	57,258	50,151	16,471	1,393,039	18,542
`Upcoming`	13,114		7,330	347,959	4,518
`Eventful`	37,647	6,543	14,576	52	12

Table 1: Number of event/agent/location and photo/user descriptions in the dataset published in [14]

In this paper, we consider a subset of this events dataset that corresponds to the intersection of Last.fm, Flickr and YouTube. In other words, we consider the set of last.fm events for which there is at least one photo and one video shared respectively on Flickr and YouTube that has been tagged with the `lastfm:event=xxx` machine tag. The number of YouTube videos that actually contains such a machine tag is unsurprisingly much smaller. Hence, this intersection yields a dataset of 110 events, 4790 photos and 263 videos.

The Ontology for Media Resource currently developed by W3C is a core vocabulary which covers basic metadata properties to describe media resources[2]. It provides properties for describing the duration of a video, its target audience, copyright, genre, rating or the various renditions of a photos. Media fragments can also be defined in order to have a smaller granularity and attach keywords or formal annotations to parts of a media. The ontology contains a formal set of axioms defining mapping between different metadata formats for multimedia. We use this vocabulary together with properties from SIOC, FOAF and Dublin Core to convert into RDF the Flickr photo and YouYTube video descriptions

[1] `http://linkedevents.org/ontology/`

[2] `http://www.w3.org/TR/mediaont-10/`

(Figure 2). The link between the media and the event is realized through the `lode:illustrate` property, while more information about the `sioc:UserAccount` can be attached to his URI. In Figure 2, we see that both the video hosted on YouTube and the photo hosted on Flickr has the same `ma:creator`: the user `cartoixa`.

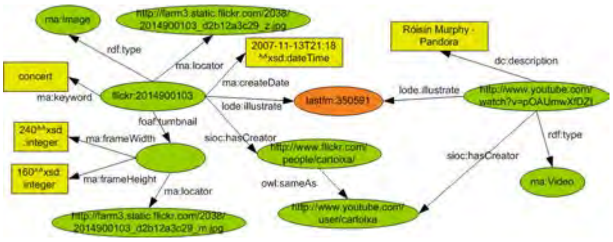


Figure 2: A photo and a video taken by the same user at the *Róisín Murphy Concert* described with the Media Ontology

4. FIND MEDIA ILLUSTRATING EVENTS

The set of photos and videos available on the web that can be explicitly associated to a Last.fm event using a machine tag is generally a tiny subset of all media that are actually relevant for this event. Our goal is to find as much as possible media resources that have **not** been tagged with a `lastfm:event=xxx` machine tag but that should still be associated to an event description. In the following, we investigate several approaches to find those photos and videos to which we can then propagate the rich semantic description of the event improving the recall accuracy of multimedia query for events.

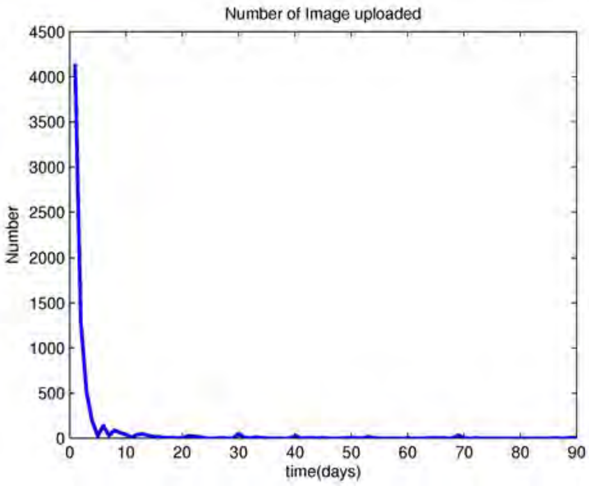


Figure 3: Image uploading tendency along time

Starting from an event description, three dimensions from the LODE model can easily be mapped to metadata available in Flickr and YouTube and be used as search query in these two sharing platforms: the *what* dimension that represents the title, the *where* dimension that gives the geo-coordinates attached to a media, and the *when* dimension that is matched with either the taken date or the upload date of a media. Querying Flickr or YouTube with just one of these dimensions bring far too many results: many events took place on the same date or at nearby locations and the title is often ambiguous. Consequently, we will query the media sharing sites using at least two dimensions. We also find that there are recurrent annual events with the same title and held in the same location, which makes the combination of "title" and "geo tag" inaccurate. In the following, we consider the two combinations "title" + "time" and "geo-tag" + "time" for performing search query and finding media that could be relevant for a given event.

4.1 How Fast Media are Uploaded?

We first investigate the time difference between the start time of an event and the upload time of Flickr photos attached to this event. For the 110 events composing our dataset, we analyze the 4790 photos that are annotated with the Last.fm machine tag in order to compute the time delay between the event start time and the time at which the photos were captured according to the EXIF metadata. Figure 3 shows the result: the y-axis represents the number of photos uploaded on a day to day basis, while the x-axis represents the time (in days) after the event occurred.

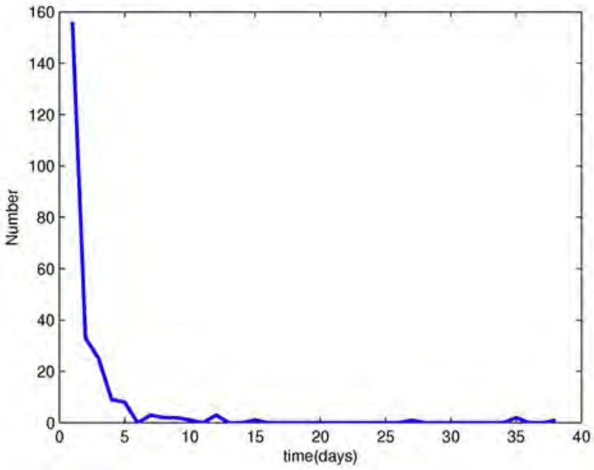


Figure 4: Video uploading tendency along time

The trend is clearly a long-tail curve where most of the photos taken at an event are uploaded during or right after the event took place and within the first 5 days. After ten days, only very few photos from the event are still being uploaded. In the following, we choose a threshold of **5 days** when querying the photos using either the title or the geotag information.

We conduct a similar analysis with the 263 YouTube videos that are annotated with the Last.fm machine tag. The "taken time" being not available for videos from the YouTube API, we use instead the "upload time". Figure 4 shows the results and we observe the same long tail: most the videos are uploaded within the first 5 days following an event.

4.2 Query by Geotag

Geotagging is the process of adding geographical identification metadata to a media and is a form of geospatial metadata. These data usually consist of latitude and longitude coordinates, though they can also include altitude, bearing, distance, accuracy data, and place names. They are extremely valuable for application to structure the data according to location and for users to find a wide variety of location-specific information [1, 17]. Considering that a place is generally a venue, we assume that at any given place and time there is a single event taking place.

For all events of our dataset, we extract the latitude and longitude information from the LODE descriptions and we perform search query using the Flickr API applying a time filter of 5 days following each event date. We perform the same query using the YouTube API although the number of video that are geotagged is much smaller than for photos. Figures 5(a) and 5(b) show the distribution of the number of retrieved photos and videos for the 110 events in our dataset. We observe that the data is centralized in the left bins which means that for most of the events (n=95), the number of photos (resp. videos) retrieved with geotags is within the 0-100 range (resp. 0-20 range). The largest bin is composed of 45 events that have each between 1 and 50 photos retrieved.

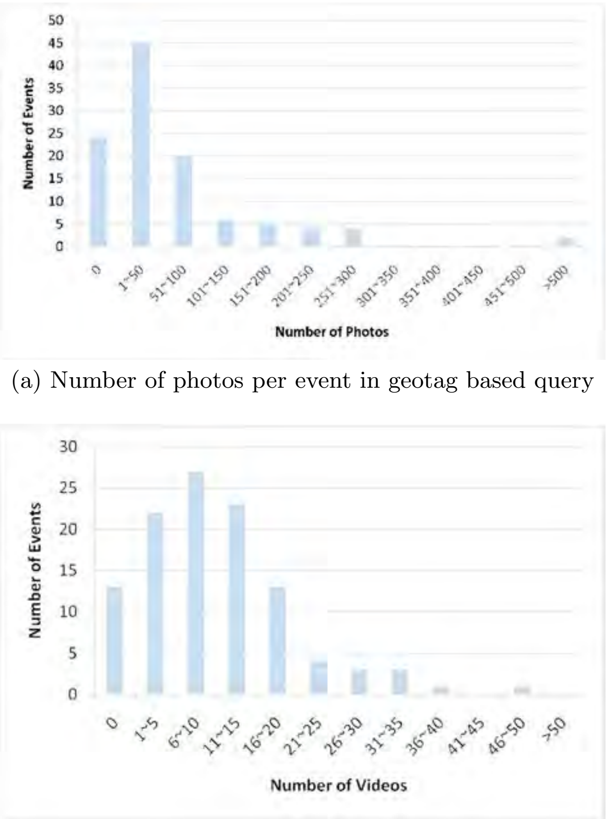
4.3 Query by Title

The title is often the most useful information for describing the events. Similarly to geo-tagged queries, we perform full text search queries on Flickr and YouTube based on the event title that is extracted from the LODE description. The photos and videos retrieved are also filtered using a time interval of five days following the time of the event. When performing search query using the Flickr API query, we use the "text mode" rather than the "tag mode" since the latter is missing in many photos. The number of photos retrieved at this stage is however in an order of magnitude greater than with geo-tagged queries. Due to the well-known polysemy problems of textual-based query, the title-based query brings lots of irrelevant photos. We describe in the Section 4.4 an heuristic for filtering out those irrelevant media.

In contrast, we do not observe this noise when querying the YouTube API with only the event title (filtered by the time of the event) using a strict match mode. Hence, the number of videos retrieved per event is rather small and most of the time relevant. The distribution of the number of retrieved photos and videos for the 110 events in our dataset is depicted in Figures 6a and 6b.

Generally, the results of query by title have a similar distribution than the result of query by geotag. For most of the events, a lower number of photos is obtained. Out of the 110 events under investigation, there are 80 events with less than 150 photos, and 83 events with less than 25 videos. However, for some events, a large number of media is retrieved: 12 events (resp. 15) with more than 500 photos (resp. 50 videos). Compared with Figure 5, we can clearly see that the standard deviation of Figure 6 is larger and that again photos are more readily available than videos.

Table 2 shows the overall number of photos and videos retrieved for each strategy for the 110 events that composed our dataset. We first observe that these two strategies allow to retrieve an order of magnitude more media that using solely machine tags. Hence, while 4790 photos are tagged with the `lastfm:event=xxx` machine tag, 6933 photos can be retrieved using the geo-location of the event and 32583 photos can be retrieved using the event title. After removing the duplicated ones, we obtain 36412 photos that are candidate to illustrate an event which is 7,6 times more than the ones labeled by a machine tag. For the videos, the number of candidates is 19,6 times more than the ones with machine tags. Unsurprisingly, most of the media uploaded and shared on the web do not have machine tag.



(a) Number of photos per event in geotag based query

(b) Number of videos per event in geotag based query

Figure 5: Statistics for geotag based query

4.4 Pruning Irrelevant Media

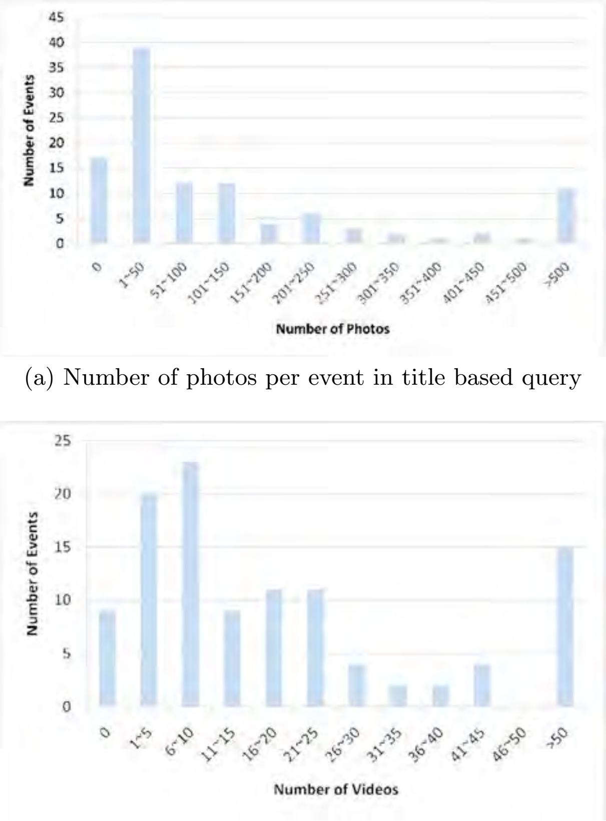
Images and videos with specific machine tags such as `lastfm:-event=207358` can be unconditionally associated to events. We consider that media retrieved with geotag queries during a correct time frame should also be relevant for those events. The problem arises with the media retrieved with text-based queries (using the event title) where one can find many irrelevant media. For example, the event identified by 207358 has for title `Malia`. However, a search on Flickr or YouTube with this keyword returns photos about cities, different people (Malawian singer, French swimmer, daughter of the US president Barack Obama) or even hotels with this name.

In order to filter out this noise and to avoid propagating rich event descriptions to those medias, we propose a method for pruning the set of candidates photos using visual analysis. The photos captured at a single event are already very diverse, depicting the artist, the scene, the audience or even the tickets. The diversity of the data makes it difficult to remove all the noisy images that should not be associated with the event considered, while keeping as much as possible the good ones. We address this issue in two steps to ensure high precision and recall ratio.

First, we build a training dataset composed of the media containing either the event machine tag or a combination of geo-coordinates and time frame corresponding to the event

Table 2: Number of photos and videos retrieved for 110 events using the event machine tag (ID), the geo-coordinates or the event title

	QueryByID	**QueryByTitle**	**QueryByGeo**	**ID∩Title**	**Geo∩Title**	**Geo∩ID**	**Geo∩ID∩Title**
Photos	4790	32583	6933	2350	494	484	405
Videos	263	4237	1163	103	39	115	29



(a) Number of photos per event in title based query

(b) Number of videos per event in title based query

Figure 6: Statistics for title based query

dimensions. The photos resulting from query by title compose the testing dataset. The visual features employed are 225D color moments in Lab space, 64D Gabor texture, and 73D Edge histogram. For each image in the training data, the nearest neighbors using the ***L1*** distance measure in the training set are found and the smallest distance taken as threshold. Second, images originating from the title query are compared with training images. Images for which the distance to images in the test set is below the threshold are candidates for illustrating the event. The algorithm can be formalized as followed:

Algorithm 1 Prune function

```
INPUT: $TrainingSet$,$TestingSet$
OUTPUT: $PrunedSet$

for each $img$ in $TrainingSet$ do
    $D = [\,]$
    for each $imgj$ in $TrainingSet$-$\{img\}$ do
        $D$.append(dist_L1($img$,$imgj$))
    end for
    $Threshold = \min(D)$
    for each $imgt$ in $TestingSet$ do
        if dist_L1($imgt$,$img$)$<$ $Threshold$ then
            $PrunedSet$.append($imgt$)
        end if
    end for
end for
return $PrunedSet$
```

We adopted an adaptive threshold because of the visual diversity within the training dataset. Even for the images belong to the same event, the concept can vary from the musicians, singer to venue, or event ticket. In order to remove noisy images in the testing data, the threshold should adjust respectively. Figure 7 shows the value of threshold used in the experiments which range from 0.01 to 0.346.

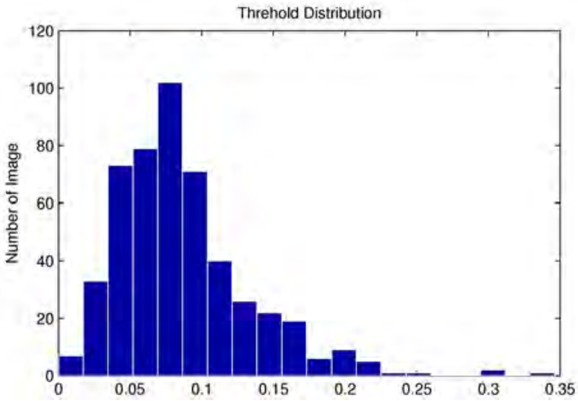


Figure 7: The distribution of threshold

Table 3: Number of photos for 20 events, results of the pruning algorithm and results of the simple heuristic extension

ID	DataSet (nb of photos)			Pruning Result			Extended Heuristic	
	TrainingData	TestingData	GroundTruth	Pruned	Precision	Recall	Extend	NewRecall
346054	2	24	2	1	1	0.500	1	0.500
158744	3	48	48	23	1	0.479	44	0.917
371981	4	16	6	4	1	0.667	4	0.667
341832	7	0	0	0	1	1.000	0	1.000
362195	7	0	0	0	1	1.000	0	1.000
235445	10	1	1	0	1	0.000	0	0.000
42644	13	85	81	13	1	0.160	13	0.160
165697	23	1	1	0	1	0.000	1	1.000
137530	24	9	4	0	1	0.000	1	0.250
517159	24	0	0	0	1	1.000	0	1.000
222241	36	204	180	33	0.97	0.183	72	0.400
234649	45	35	4	1	1	0.250	1	0.250
207358	54	68	4	4	1	1.000	4	1.000
429517	60	171	169	27	1	0.160	41	0.243
437747	65	144	142	8	1	0.056	13	0.092
117886	68	99	97	4	1	0.041	11	0.113
150390	71	16	16	1	1	0.063	1	0.063
350591	79	85	85	6	1	0.071	66	0.776
472733	93	500	478	8	1	0.017	18	0.038
176257	97	260	255	47	1	0.184	147	0.576

4.5 Experiments

For evaluating our pruning algorithm, we take the top 20 events from our 110 events dataset. For these 20 events, there are 785 images in the training set (photos containing either an event machine tag or a geotag) and 1766 photos in the testing set (photos retrieved by event title). We build manually the ground truth for those 1766 photos selecting which ones should be attached to an event and which ones should not (Table 3). The 20 events were all concert events and photos are often depicting artists, venues, stages or audience. Some photos were, however, sometimes hard to judge but the manual assessor used all metadata available around each photo such as the entire list of tags or the albums in which the photos were gathered to decide whether the photo should be discarded or not. In the end, we manually remove 193 irrelevant images by their visual appearance and metadata. The remaining 1593 images are used as ground truth dataset.

The results of the pruning algorithm detailed in the Section 4.4 applied to the 1766 photos are shown in the Table 3. The threshold used is quite strong in order to guarantee a precision of **1** for most of the events. However, this causes about 80% of the candidate images to be excluded, including many relevant photos.

In order to increase the recall ratio, we extend the selected images by our prunning algorithm with all the ones uploaded by the same uploader. The rationale is that if one photo can reliably be attached to an event, we infer that this person indeed attended this event and that all the others photos taken by this person during this time frame are likely to be illustrative media for this event. This simple heuristic allows to significantly improve the recall ratio without sacrifying to the precision.

5. DISCUSSION

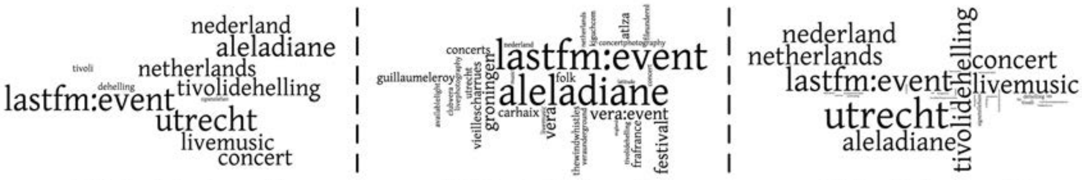
Event directories are largely overlapping, providing multiple identifiers for the same venues, artists, and events. We argued that linked data technologies help to integrate at large scale all data sources because of the use of URIs for identifying objects and a simple triple model for representing all metadata yielding a giant graph. Rich semantic descriptions of events can then be propagated to the media to which they are attached. Hence, for the dataset[3] presented in the Section 3, 1,248,021 photos (that is 73 %) have been geotagged for free since Flickr had no geo-tagged information for those photos but only knowledge of an event machine tag that points to a rich description of an event including venues that are geo-localized. Similarly, the propagation of semantic metadata enables to detect inconsistencies between data sources such as the misplacement of a venue.

We have proposed a method for finding media that are relevant for an event based on queries using several dimensions of the event, and pruning the resulting results using visual similarity. However, we observe that there is limited value in pruning video results yet while it is very necessary for photos. Although the total number of video uploads is still exponentially increasing[4], duplicates or videos with absolute no metadata and a very small number of views are important which prevent their discovery.

We also investigate the concept shift as the set of relevant media increases by looking at the tag cloud associated to these media attached to a particular event. Figure 8 depicts tag cloud examples for the event `1097166` that corresponds to the live concert of Alela Diane which took place on Tuesday 14 July 2009 at 7:30pm at the venue `Tivoli De Helling` in Utrecht, The Netherlands. From the three sub figures, we can clearly see the topic shift when new metadata is added

[3] The entire dataset is composed of more than 30 million RDF triples and is available as a dump at `http://www.eurecom.fr/~troncy/ldtc2010/`

[4] YouTube reported in November 2010 that 35 hours of video content are uploaded every minute.



a) Flickr photos tagged by **"lastfm:event=1097166"**

b) Flickr photos tagged by **"lastfm:event=1097166"** AND by **"Alela Diane"**

c) Flickr photos tagged by **"lastfm:event=1097166"** AND geo-tagged by the venue

Figure 8: Tag Clouds of photos associated to the event 1097166: *Alela Diane at Tivoli De Helling (Utrecht) on 14 Jul 2009*

from a another source. Figure 8(a) is built from all the tags of all photos retrieved when using the event machine tag: the most frequent keyword in this cloud is, as expected, `lastfm:event`. However, in Figure 8(b), when the tags set are enlarged by the metadata from the photos retrieved by query by title (and pruned with our algorithm), the topic shifts to `aleladiane` who is the artist performing during this event. A similar observation can be made after looking at Figure 8(c), where some metadata from query by geotag are added, and the most significant keyword changes to the location of the event `utrecht`.

Finally, we briefly present initial interfaces that we have started to develop for searching and browsing media through an event perspective [13]. Users wish to discover events either through invitations and recommendations, or by filtering available events according to their interests and constraints. Therefore, the interface allows constraining different event properties (e.g. time, place, category). Mechanisms for providing this desired support include restricting a time period through a timeline slider control input (Figure 9). Categories and location can be filtered using hierarchical faceted metadata [6], allowing users to browse through different dimensions of the collection. The hierarchical facets are presented according to a taxonomy of predefined event categories, and through an event's geo-location information. These properties allow the combination of different event types and locations while visually guiding the user through an interactive query refinement process. Faceted browsing also avoids empty results by restricting the available filtering options to display only non-empty results. Since users are likely to revisit information they have viewed in the past [6], we will also support simple history mechanisms, by saving a list of recently viewed events. To aid search, input boxes with dynamic term suggestions (auto-completion) is used to provide user feedback by suggesting a list of matching terms while typing.

After an event is selected, all associated information is displayed. Media are presented to convey the event experience, along with social information to provide better decision support. According to user interests, social proximity should be emphasized while displaying event attendance (e.g. friends attending). Other information that should be presented includes: performers, topic, genre, price. While scraping the data, some events such as popular music festivals were associated with more than 2,000 photos and videos. In order to deal with this large number, pagination is used while ordering media according to different contexts (e.g. by popularity, time, or social proximity). Alternatively, pictures are clustered according to context or visual similarity, and representative images are shown through Treemaps to present a varied sample of associated media.

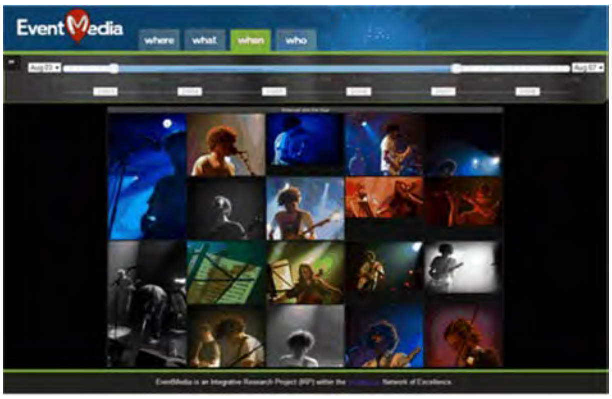


Figure 9: Interface illustrating a set of media associated to an event for a period of time

6. RELATED WORK

In recent years, research on how to better support the end-user experience when searching and browsing multimedia content has drawn lots of attention in the research community. A tremendous amount of work has been done in very different areas. Among the possible directions, the usage of low-level visual features for improving content-based multimedia retrieval systems has made great progress in the past ten years [4]. The drawback of content-based retrieval systems is often the lack of manually labeled data for training systems. Our approach propagates the rich semantic description of events to the media, thus contributes to semi-automatically build reliable large training dataset. In [2, 3], the authors follow a very similar approach, exploiting the rich "context" associated with social media content and applying clustering algorithms to identify social events. In contrast to our work, they do not rely on linked data technologies to realize large scale integration and reconciliation of event directories.

Tagging is popular on media sharing web sites, but tags are also as diverse as there are users. Tags might describe the visual content of media but could also simply refer to emotions of be completely personalized to a user with the sole aim of triggering the user's memory. In [12], the authors take tags as a knowledge source and they studied the problem of inferring semantic concepts from associated noisy tags of social images. Some other work are done to improve the tag quality. In [9], Liu proposed a social image retagging approach that aims to assign better content descriptor to the social images and remove noise description. In [1], Arase et al. propose a method to detect people's trip based on their research of geo-tagged photos.

A natural extension of our work would benefit from [8]. In this paper, the authors proposed a system to present the media content from live music events, assuming a series of concerts by the same artist such as a world tour. By synchronizing the music clips with audio fingerprint and other metadata, the system gives a novel interface to organize the user-contributed content. We did not yet consider audio fingerprint for tracking down series of events but rely only on semantic metadata so far.

7. CONCLUSION AND FUTURE WORK

In this paper, we have shown how linked data technologies can be used for integrating information contained in event and media directories. We used the LODE and Media Ontology respectively for expressing linked data description of events and photos. We described a method for finding as much as possible photos and videos relevant for a given event: we start from the media that contain specific machine tags and that can be used to train classifiers that will prune results from general queries. We evaluated our approach against a manually built gold standard and we show that we are able to increase significantly the recall with a very conservative approach that does not sacrify the precision. Ultimately, we aim at providing an event-based environment for users to explore, annotate and share media and we present an initial user interface (available at `http://eventmedia.cwi.nl/demo`) that we continue to develop.

We are currently consolidating and cleaning our dataset with more sources and more linkage. We intend to provide soon user participation at events from public Foursquare check-in and live Tweets. Our priority is also to express the right licensing and attribution information to the data that has been rdf-ized. We truly believe that multimedia will then be finally added back to the Semantic Web.

Acknowledgments

The research leading to this paper was partially supported by the project AAL-2009-2-049 "Adaptable Ambient Living Assistant" (ALIAS) co-funded by the European Commission and the French Research Agency (ANR) in the Ambient Assisted Living (AAL) programme, and by the projects FP7-216444 "Peer-to-peer Tagged Media" (Petamedia).

8. REFERENCES

[1] Y. Arase, X. Xie, T. Hara, and S. Nishio. Mining People's Trips from Large Scale Geo-tagged Photos. In *18th ACM International Conference on Multimedia (ACM MM'10)*, pages 133–142, Firenze, Italy, 2010.

[2] H. Becker, M. Naaman, and L. Gravano. Event Identification in Social Media. In *12th International Workshop on the Web and Databases (WebDB'09)*, Providence, USA, 2009.

[3] H. Becker, M. Naaman, and L. Gravano. Learning Similarity Metrics for Event Identification in Social Media. In *3rd ACM International Conference on Web Search and Data Mining (WSDM'10)*, pages 291–300, New York, USA, 2010.

[4] R. Datta, D. Joshi, J. Li, James, and Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40, 2008.

[5] A. Fialho, R. Troncy, L. Hardman, C. Saathoff, and A. Scherp. What's on this evening? Designing User Support for Event-based Annotation and Exploration of Media. In *1st International Workshop on EVENTS - Recognising and tracking events on the Web and in real life*, pages 40–54, Athens, Greece, 2010.

[6] M. Hearst. *Search User Interfaces.* Cambridge University Press, 2009.

[7] J. Hobbs and F. Pan. Time Ontology in OWL. W3C Working Draft, 2006. `http://www.w3.org/TR/owl-time`.

[8] L. Kennedy and M. Naaman. Less talk, more rock: automated organization of community-contributed collections of concert videos. In *18th ACM International Conference on World Wide Web (WWW'09)*, pages 311–320, Madrid, Spain, 2009.

[9] D. Liu, X.-S. Hua, M. Wang, and H.-J. Zhang. Image retagging. In *18th ACM International Conference on Multimedia (ACM MM'10)*, pages 491–500, Firenze, Italy, 2010.

[10] Y. Raimond, S. Abdallah, M. Sandler, and F. Giasson. The Music Ontology. In *8th International Conference on Music Information Retrieval (ISMIR'07)*, Vienna, Austria, 2007.

[11] R. Shaw, R. Troncy, and L. Hardman. LODE: Linking Open Descriptions Of Events. In *4th Asian Semantic Web Conference (ASWC'09)*, 2009.

[12] J. Tang, S. Yan, R. Hong, G.-J. Qi, and T.-S. Chua. Inferring semantic concepts from community-contributed images and noisy tags. In *17th ACM International Conference on Multimedia (ACM MM'09)*, pages 223–232, Beijing, China, 2009.

[13] R. Troncy, A. Fialho, L. Hardman, and C. Saathoff. Experiencing Events through User-Generated Media. In *1st International Workshop on Consuming Linked Data (COLD'10)*, Shanghai, China, 2010.

[14] R. Troncy, B. Malocha, and A. Fialho. Linking Events with Media. In *6th International Conference on Semantic Systems (I-SEMANTICS'10)*, Graz, Austria, 2010.

[15] W. van Hage, V. Malaisé, G. de Vries, G. Schreiber, and M. van Someren. Combining Ship Trajectories and Semantics with the Simple Event Model (SEM). In *1st ACM International Workshop on Events in Multimedia (EiMM'09)*, Beijing, China, 2009.

[16] U. Westermann and R. Jain. Toward a Common Event Model for Multimedia Applications. *IEEE MultiMedia*, 14(1):19–29, 2007.

[17] Y.-T. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T.-S. Chua, and H. Neven. Tour the World: building a web-scale landmark recognition engine. In *22nd International Conference on Computer Vision and Pattern Recognition (CVPR'09)*, Miami, Florida, USA, 2009.