

Video Exploration: From Multimedia Content Analysis to Interactive Visualization

ML Viaud, O Buisson, A Saulnier, C. Guenais

INA

4 avenue de l'Europe

94366 Bry Sur Marne, France

{mlviaud,obuisson,asaulnier,cguenais}@ina.fr

ABSTRACT

This paper presents 3 interfaces to access video contents. The stream explorer allows to explore and to segment video streams. The video explorer shows a synthetic view of structured TV programmes. The collection explorer proposes cartographies of large video collections. Based on visual and textual automatic processing, proximities and redundancies are analyzed, allowing the emergence of different levels of structure. This is made possible thanks to the volume of data considered: 7 channels during 100 days, ie 16000 hours or 20 Millions key frames. These three tools allow efficient exploration of video contents at different levels of interest: image, shot and sequence, programme and collection level.

Categories and Subject Descriptors

I.4 [Image processing and computer vision]: *feature measurement, feature representation*

H.3.3 [Information Search and Retrieval]: *Content analysis and Indexing, indexing methods Information Search and retrieval Clustering*

H.5 [Information interfaces and presentation]: *User Interfaces, Graphical user interfaces*

General Terms *Algorithms, Theory.*

Keywords

Visualization, Graphic Interfaces, Video Browsing, Video Summary, Cartographic Representation, Content analysis and Indexing

1. INTRODUCTION

The role of the French Audiovisual Institute (INA) is to store and preserve the French audiovisual heritage, to ensure its exploitation and to make it more readily available. Moreover, under the terms of the French law voted on 20 June 1992, Inathèque de France is responsible for collecting and preserving radio and television broadcast. To achieve this mission, 150 INA's media librarians annotate the audiovisual documents collected daily, at different

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Conference'10, Month 1-2, 2010, City, State, Country.

Copyright 2010 ACM 1-58113-000-0/00/0010...\$10.00.

levels of precision. INA's archives holds 1 500 000 pictures and 1 300 000 hours of video (600 000 digitized) and more than one million documentary notes. From 1998 to 2010, Inathèque de France has increased the scope of its collections and 100 TV channels and 20 radio channels are being collected. The entire collection of INA's archives is made available online for professionals by subscription www.inamediapro.com. Since April 2006, INA's public web site has assembled more than 20000 hours of Radio and TV broadcasts, representing about 85000 programmes or excerpts. Documenting these ever-growing contents with constant human resources makes the use of new technologies a necessity. Moreover, a study on INA's professional clients shows that digitization of video contents leads to new uses and needs. Clients expect more precise and diversified access to contents, and ask for a finer description and segmentation of the resources. Then, needs for new tools are clearly expressed to assist archivists in the creation of video and radio excerpts, and to provide users with new types of search and navigation in video contents. In INA's professional context, semi-automatic tools are promoted to assist archivists in their tasks without loss of control high added value, which is to produce coherent and normalized annotations at high semantic level. On the other hand, the INA online site addresses the general public, which opens new paradigms for resource access. Discovering TV archives becomes a recreation process: search strategies are totally different, mainly based on the same rules as web search. Users will follow advices from friends, have butterfly navigation behaviors in the resources or try loose queries. So, open interfaces and navigation modes are required. This paper presents 3 tools to easily explore video contents. The Video Stream Explorer goal is to explore in detail TV streams visual content to browse, search or achieve quick and precise segmentation. The Video Programme Explorer gives an overview of structured TV programmes and allows intuitive browsing of content. The Collection Explorer presents a map of collections based on semantic proximities. These interactive visualization prototypes are based on image or textual automatic processing.

2. State of the art: interfaces for video access

Video analysis is a time consuming process because it implies time linear reading. For a few years, many interfaces have been dedicated to video content summarization for exploration. Most of these works are based on image and/or audio analysis to extract the most representative content a video or video stream, and propose adapted visualisation and interaction mode. The Manga approach [1] and the panoramic mosaic [2] propose a static linear layout of the most important events of the video content. The storyboard approach [3] and the Silver systems [4] introduce a

notion of hierarchy in the video representation. Campanella and all [5] have created multi-view environment to analyze video content for annotation. Dynamic Summarization techniques based on audio content has been applied to TV stream in [6, 22]. The Lean system [8], which runs with tabletPC, allows dynamic exploration of video, thanks to the “twist lens”. More detailed state of the art about video exploration may be found in [9, 10]. Browsing efficiently collections of video remains a challenge [23]. The MediaMill video search engine [11, 12, 13] proposes several browsing tools based on visual and/or concepts, and on 2D or 3D layouts. Most of these methods rely on intra- programme analysis and work on limited amount of data. The strengths of this paper rely on the volume of video resources considered, the granularity of the proposed access (stream, programme and collection) and the originality of the related visualizations. In fact, proximities and redundancies are analysed at the stream, programme, and collection level, allowing the emergence of different levels of structure that are exploited in the interfaces.

3. Multimodal Similarities for video access

3.1 Visual Features and shot representation

The segmentation process of videos relies on visual content analysis. Global visual features are extracted from each frame of the stream. Image descriptors based on histograms in different color spaces (RGB, LUV and HSV) and gradient orientation histograms are processed. Then, a robust clustering method, Rock [14], is applied on these visual descriptors. To reduce the main computational costs of the clustering process (similarity measures between each descriptor), we developed a specific version of Rock method based on our index structure: PMH (Probabilistic Multidimensional Hashing) [15]. Our video segments differ slightly from the classical definition of shots because they are based on visual content and not only on editing cuts. In fact, we define a shot as the result of visual descriptors clustering: consecutive frames belonging to the same cluster, are considered in the same shot. This strategy has two advantages: representative images of shots summarize precisely visual contents and search and data mining functions may then be processed at the shot level. Clusters’ generation and characterization may be improved with learning techniques. We use a machine learning method based on Evidential KNN Classifier [16] both to reduce the number of descriptors per shot and to select the representative visual descriptors. The classifier learns thanks to positive and negative samples. In our case, a class corresponds to a cluster, the positive samples are the descriptors of the current cluster, while negative samples are the descriptors of other clusters. The model obtained for each cluster is composed of a concatenation of descriptors, and a radius. The radius is defined in the feature space with the distance used by the clustering method (L2 or L1). The use of the PMH index structure allows very efficient search processes for very large database. The experimentation presented here involved 7 French channels during 100 days, which represents 16000 hours of video stream or 20 million of images.

3.2 Shot labeling and Data Mining

TV is built on viewers regular “Rendez-vous” and presents multiple visual redundancies: jingles, programme openings, TV programme announcements, weather forecast, commercials, TV shows... The model search for shots may also be used to detect similar shots in order to analyze and label video streams’ contents. The simplest process for shot labeling corresponds to the near-duplicate detection. We use this method to locate the beginning of

TV programmes, specific announcements or advertising. The strategy is based on a semi-supervised method: an archivist provides a single shot or image example and the title of the programme. The shot/image is visually described and the search process uses its representative descriptors to detect each near-duplicate in the TV stream. To extend the shot labeling to shots that are similar but not near copies, like sets or backgrounds of programme collections, we provide a limited set of examples (ground truth) for each category. Each set of ground truth is retrieved in the entire corpora by our previous detection method. Results are analyzed by data-mining algorithms to extract temporal statistical characteristics for each category: frequency, time slot and main days of broadcast... These characteristics are used to classify or label automatically shots in classes such as frequentProgramme, periodicalProgramme, rareProgramme, weather forecast, ads... We define decision results, which are automatically learned by standard Decision Tree learning algorithms [17]. This strategy allows for automatic labeling of shots on very large-scale corpora of TV streams and with very little manual work.

3.3 Textual indexing and search

TV programmes are associated with metadata inherited from TV channels or guideline, or created by archivists. We select fields that describe the resource content, such as keywords, title, and summary, when it exists. These textual metadata fields are indexed with Lucene/Apache. Then, for each documents we perform a weighted multi fields search and store the K-nearest neighbors and their rank. This process has been applied to the collection of 80000 video excerpts available on INA’s web site www.ina.fr to build a “semantic” distance between documents.

4. Stream Explorer

The goal of the stream explorer is to allow quick and precise access to visual contents of video streams. Our constraint here is to build a view able to show the segments corresponding to one hour of video, the shortest segment being a single frame segment. The interface uses the metaphor of the 60’s tape recorder, where the double spirals of the tape represent the timeline of the detected video segments. The number of turns of the spirals has been studied to satisfy the constraint of visibility for single frame segments. The video player is synchronized within the middle of the line between the two spirals, which corresponds intuitively to the tape head of the recorder. Metadata relative to the current segment (keywords, transcription, archives’ notes, summaries) are displayed in the central area. We chose to map length and angles to produce an automatic zoom effect: the further a segment is from the current segments, the smaller it appears on the timeline. Moreover, the temporal windows displayed may have fixed or shifting boundaries, allowing continuous exploration of the stream. The video stream is segmented in classes that are represented by colored areas on the timeline. The caption appears on the left of the interface. Figure 1 represents the French public channel F2, 27th of July 2009, from 7 to 8 pm. For this specific part of the TV Stream, 6 generic classes (advertising, inter advertising, rare content, frequent shots, recurrent programmes) and 7 specific labels (FR2-transition, FR2 credits, FR2 “Bonne soirée” credits, FR2 adverting generic, FR2 programmes announcement, weather broadcast, News jingle, “On aime la chanson” generic) are detected and marked. The class name appears when moving the mouse over the segments. The *play*, *pause*, *rewind*, *forward* buttons of the video player behave like

classic recorders, rolling or unrolling the tape and displaying the central current segment. Clicking on a segment makes it the current segment. This interface has been developed in FLEX, with a QuickTime video player. Our first assessments from professional users are enthusiastic because it covers real needs to assist video segmentation. In fact, precise content visualization is needed to build excerpts such as commercials, songs or political citations.



Figure 1. Interactive Video Stream Explorer: F2 programme, on 27 July 2009 from 7 to 8 pm

5. TV Programme Explorer

TV news, games, sport programmes, magazines or debates are built on interlaced sequences presenting often visual and/or audio redundancies: anchorman, journalists, actors, setting views, highlight in sport competition... The hypothesis of the TV Programme Explorer is that these visual redundancies, which are the result of some editing process, reflect most of the time higher levels of semantic. The idea behind the TV programme explorer [18] is to structure the programme visualization with a backbone that gathers either the biggest cluster of similar images or a chosen set of clusters of similar images. The images of this backbone are lined up on the horizontal axis and a new graph is built. Two time constraints are applied on edges: a chronological order is applied in the backbone cluster, and the basic temporality is preserved for other frames. Finally, a topological custom algorithm based on Fruchterman Rhein-gold spreads out thread loops on both sides of the backbone. The layout has to be read from left to right, in chronological order Figure 2 shows a news programme.

The final visualization emphasizes the video structure and gives a quick overview of its content. Loop length indicates sequence duration. Generally, news edition balances reports length, then length exception is a good indication to notice a important event. Images of the loop are zoomed to enhance the perception of the most informative content. The video player is synchronized with the image on which the mouse is positioned and the loop currently displayed is marked by colored frames on the images. Metadata (description and/or transcription) is available under the video player. One click in the image freezes the video, a double click plays it again. Textual search is available on metadata and the retrieved segments are marked. Textual search is available on textual metadata.

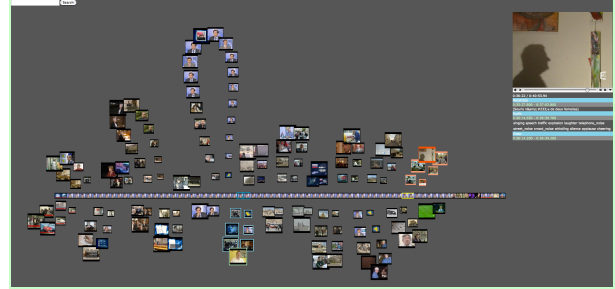


Figure 2. Programme Explorer: News of the French public channel FR2, on 4 April 2004

6. Collection Explorer

The Video Collection Explorer aims at giving users an overview of the size and richness of the resources available. The goal of this visualization is to gather documents semantically similar on the map.

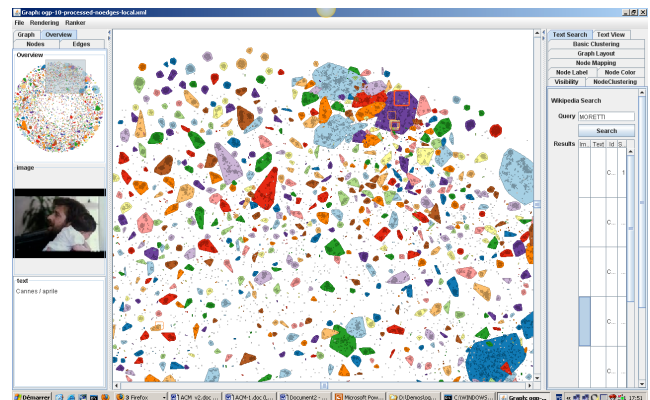


Figure 3. Maps of the 80000 TV excerpts available on INA's public web site (www.ina.fr) The central window displays the overview of the collection. Zoom and translation functions are available to discover clustered contents. The upper left window displays the entire collection and the zoomed area. The left window presents functionalities. In this picture, the search "MORETTI" is launched and the documents are marked with colored squares. The selected document is surrounded by the biggest red square and its thumbnail and textual description are displayed on the left.

The process is based on graph models. Each programme or excerpts has been described by archivists and distances between resources are processed from keywords and summaries. The graph of resources is created using the distance matrix to generate valuated edges. The layout uses a customized energy force model algorithm [19] [20] to achieve the layout. Within this model, we consider a repulsive force between nodes, and a spring force between connected nodes. Each edge is seen as a spring characterized by its resting length and its stiffness coefficient. The resting length of each edge is linearly correlated to its distance attribute. For graphs based on similarity matrices, the highest similarities of each node appear to be the most significant criteria to elaborate the layout. Then, we generate a graph issued from knn filtered matrices. Moreover, we have implemented generic filtering methods of nodes and edges based on their inner or topological attributes (centrality, degree, hub/authority values...). In such cases, the radius of a cluster and the distance between two

clusters are related to the inverse of the edge density (normalized edge-cut) [21]. Finally, we use a standard agglomerative hierarchical clustering algorithm to identify and label clusters in the display area. To obtain clusters of arbitrary shapes, we choose a linkage metrics based on the minimum distance between objects. These distances can be parameterized in the interface, allowing the user to *control* the view.

The map of the collection appears in the central part of the interface. Clusters' exploration may be done either by pointing the mouse over a cluster, which displays the list of the main words associated with the cluster on the right window, or by zooming in the view. The zoomed area is shown on the top left window of the interface. A thumbnail and a light description of the document pointed by the mouse are shown in the middle and bottom left windows. A search function highlights results according to their rank.

7. Conclusion

INA legal deposit stores and describes manually about 100 TV channels and 20 radios. In this context, automatic processing allowing an intuitive and precise browsing of video contents becomes a real challenge. To reach INA objectives in terms of evolutions, we have created links between different scientific fields (image description, index structuring, information visualization and IHM). These links enabled us to develop modules based on these different fields to achieve prototypes adapted to INA's needs (search, segmentation and annotation help, analysis & monitoring). The evaluation with professional users shows a real interest from users to try the tools in real context. Visual descriptions are being integrated in INA general system, making such objective a near future reality. We are now trying to focus on the temporal continuity of the video streams so as to enhance the quality of the segmentation. The next main improvements will be the integration of audio analysis to produce audio segments (classification of sounds, speaker detection...) and to merge visual and audio information in order to obtain more robust "meta-segments". We thanks Jérôme Thièvre, Hervé Goëau, and Laurent Joyeux for their participation to this work, The European commission and the Vitalas Project which supported this work; Special thanks to Alexis Joly and the INRIA Imedia Project.

REFERENCES

- [1] S. Uchihashi, J. Foote, A. Girgensohn, and J. Boreczky. Video Manga: Generating semantically meaningful video summaries. In *Proceedings of Multimedia'99*, pages 383–392. ACM, 1999.
- [2] D. B. Goldman, B. Curless, S. M. Seitz, and D. Salesin. Schematic storyboarding for video visualization and editing. *ACM Transactions on Graphics (Proc. SIGGRAPH 2006)*, 25(3), July 2006.
- [3] D. Zhong, H.J. Zhang, and S.F. Chang. Clustering methods for video browsing and annotation. In *Proc. SPIE*, volume 2670, pages 239–246, 1996.
- [4] B.A. Myers., D. Yocum, , S. Stevens, L. Dabbish, A. Corbett, J. Casares, *First ACM/IEEE Joint Conference on Digital Libraries (JCDL'01)*, Roanoke, VA, pp. 106-115, June 24-28, 2001
- [5] M. Campanella, R. Leonardi, P. Migliorati, Interactive visualization of video content and associated description for semantic annotation, *SIViP(3)*, No. 2, June 2009, pp. xx-yy. Springer DOI Link 0903
- [6] M.A. Smith, T. Kanade, Video skimming and characterization through the combination of image and language understanding techniques, *CVPR97*, pp.775-781, 1997
- [7] G Ramos, R. Balakrishnan. (2003). Fluid interaction techniques for the control and annotation of digital video. *ACM CHI Letters*, ACM UIST 2003 Symposium on User Interface Software & Technology.
- [8] Klaus Sch'offmann, "Enabling Explorative Search in Videos for Instantaneous Use by Fast Content Analysis and Integration of Users' Expertise", thesis, mai 2009.
- [9] R. Hammoud. *Interactive Video: Algorithms and Technologies (Signals and Communication Technology)*. Springer-Verlag New York, Inc. Secaucus, NJ, USA, 2006.
- [10] J. Graham, JJ Hull, R.I. Inc, and CA Menlo Park. A paperbased interface for video browsing and retrieval. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, volume 2, 2003.
- [11] C. Snoek, M. Worring, D. Koelma, and A.W.M. Smeulders. Learned lexicon-driven interactive video retrieval. In *International Conference on Image and Video Retrieval (CIVR 2006)*, volume LNCS 4071, pages 11{20, Tempe, AZ, USA, July 2006. Springer Berlin / Heidelberg.
- [12] M. Worring, C.G.M. Snoek, O. de Rooij, G.P. Nguyen, and W.M. Smeulders. The Mediamill semantic video search engine. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (invited paper)*, Honolulu, Hawaii, USA, 2007.
- [13] M. Worring, C.G.M. Snoek, D.C. Koelma, G.P. Nguyen, and O. de Rooij. Lexicon-based browsers for searching in news video archives, *ICPR 2006*, pages 1256 Los Alamitos, CA, USA, 2006. IEEE Computer Society.
- [14] S. Guha , R. Rastogi And Kyuseok Shim, 2000, ROCK: A Robust Clustering Algorithm for Categorical Attributes.
- [15] A. Joly, O. Buisson, "A Posteriori Multi-Probe Locality Sensitive Hashing", in *Proc. of ACM Multimedia*, 2008
- [16] H. Goeau, O. Buisson and M.L. Viaud, "Image collection structuring based on evidential active learner", at *CBMI*, 2008.
- [17] T. Hastie, R. Tibshirani, J.H. Friedman, *The elements of statistical learning : Data mining, inference, and prediction*. New York: Springer Verlag, 2001.
- [18] H Goeau, O Buisson and ML Viaud, "Image collection structuring based on evidential active learner", at *CBMI*, 2008.
- [19] T. M. J. Fruchterman and E. M. Rheingold, "Graph drawing by force directed placement," in *Software - Practice and Experience*, 1991, vol. 21, pp. 1129–1164.
- [20] A. Noack. An energy model for visual graph clustering. In G. Liotta, editor, *11th International Symposium on Graph Drawing (GD 2003)*, LNCS 2912, pages 425{436, Berlin, 2004. Springer-Verlag.
- [21] A. Noack. Energy-based clustering of graphs with non uniform degrees. In *14th International Symposium on Graph Drawing (GD 2006)*, pages 309-320, Limerick, Ireland, 2006.
- [22] M. Delest, A. Don and J. Benois-Pineau, DAG based visual interfaces for navigation in index vidéo content, *Journal Multimedia Tools and Applications*, Volume 31, Number 1, page 51--72 - Oct 2005.
- [23] T. Meiers, T. Sikora and I. Keller, Hierarchical image database browsing environment with embedded relevance feedback. *ICIP (2)* 2002: 593-596, 2002.