INTERACTIVE VISUALIZATION TOOL WITH GRAPHIC TABLE OF VIDEO CONTENTS

H. Goeau[1,2], *J. Thièvre*[1], *M-L. Viaud*[1], *D. Pellerin*[2]

[1]Institut National de l'Audiovisuel
INA
4, avenue de l'Europe
94366 Bry-sur-marne Cedex, France
{hgoeau, jthievre, mlviaud}@ina.fr

[2]Grenoble Image Parole Signal Automatique
GIPSA - lab (ex LIS - INPG)
46, avenue Félix Viallet
38031 Grenoble, France
{denis.pellerin, herve.goeau}@lis.inpg.fr

ABSTRACT

We present an interactive visualization, called *Table Of Video Contents (TOVC)*, for browsing structured TV programs such as news, magazines or sports. In these telecasts, getting a good segmentation can be very time-consuming, especially in an annotating context. This visualization, connected with a classical media player, offers a very handy video browser. This system allows a global overview by showing the temporal structure and by giving some semantic information. The drawn structure enables a non linear video access by suggesting relevant frames. The TOVC is created from a graphic framework designed for computing similarities on visual content, and displaying the associated proximities in a 2D map with graph representation. TOVC is one of its first applications and shows interesting capabilities.

1. INTRODUCTION

The French National Audiovisual Institute (INA) is in charge of the preservation of French audiovisual heritage and of its access. In this paper, we focus on what we call composite TV programs which represent 30% of INA's data. These documents are structured videos such as news or magazines, composed of sequences such as reports or interviews inserted in a main sequence generally animated by an anchorman on a TV floor. Getting a good segmentation and an overview of these videos is very time-consuming especially in the manually annotating context of INA.

We propose a tool to help professionals of documentation to segment and browse efficiently these videos. A framework has been designed in order to visualize numerous data with various similarity measures. The idea is to create a generic framework to fulfill various tasks in video indexing. The paradigm is that it is very difficult to elaborate a full automatic system efficient for all program genres. Moreover, in a telecast collection, along the years, fundamental evolutions can appear in the editing, the studio background, or the visual effects. An expert user will design different profiles adapted to a specific collection. For instance, in sport programs, it is relevant to favorize similarity measures on color and motion features, but, in commercials, the cutting rhythm and the sound intensity are more appropriate.

Making Table Of Video Contents visualizations involves 4 stages in which the expert user can interact. Low-level features are first computed to create similarity matrices on key frames. A 2D map of image proximity is built using a graph approach. Clusters of similar visual content are identified, and finally reorganized with time constraints to produce the final view.

First part of the paper is a short state of the art on video browsing. Then, we describe the framework and discuss some experimental results. Finally, perspectives are given for future improvements.

2. STATE OF THE ART

Various approaches exist for multimedia navigation by content. Temporal slider is the most common and easiest way to get a quick overview of video content. The main disadvantages of this tool are the lack of scalability for long documents and the unpleasant aspect due to the jerky visual feedback. A vertical mouse movement can be applied to change the slider scalability for browsing video at different granularities [1]. In the CueVideo system [2], users can modulate the video speed with a time scale modification which prevents from noticeable distortion on speech comprehension.

Other approaches are focused on shots visualization. For shots with camera motion, mosaicking is useful for stitching overlapped frames and condense them in one global picture [3]. For static shots without camera motion such as security video, only motion events can be overlayed in one global image [4].

Checking off judicious time codes or key frames is an interesting way for orientating users in a time scale. In [5], Hidden Markov Models on multimodal features are applied with rule priors for linking game's points of video tennis to a hierarchical table of contents. 2D map of shots can be established from low level similarities for shots clustering with MultiDimensional Scaling (MDS) methods [6][7]. Temporal proximities are combined with low level similarity to create graph visualizations where nodes are clusters of video shots and edges are links between them [8].

Based on some of these previous ideas, we propose an original 2D vizualisation where the document structure appears clearly by stressing notable time entries for video browsing.

3. FRAMEWORK DESCRIPTION

Our framework is an experimental system for manipulating multimedia documents with graph representation. The framework (see figure 1) is composed of two main parts, a computing unit for producing similarity matrices of low level features on videos, and a Graphical User Interface (GUI) for exploring and managing 2D visualizations.

Graph representation as a data structure has several advantages.

- It is efficient for expressing a distance matrix: a node is an entity to compare, and an edge, valued or not, indicates if a link exists between two nodes.

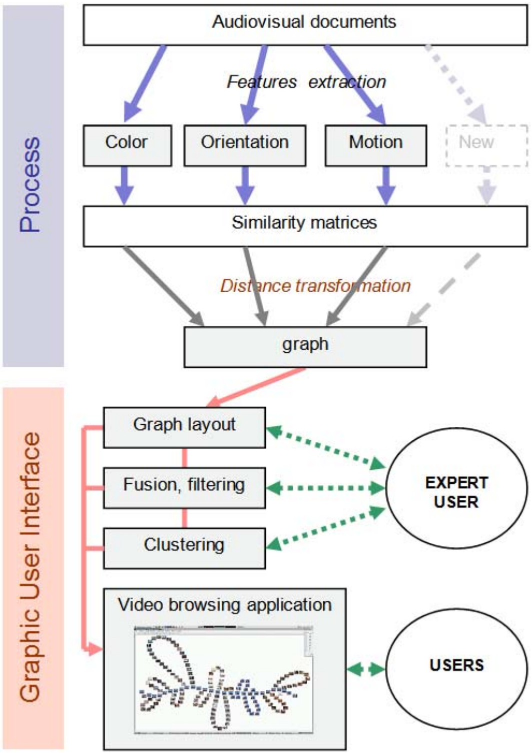

Fig. 1. The framework: a computing unit produces similarity matrices on video key frames, an expert user manipulates data and designs a profile to allow a video browsing in a collection.

- Numerous algorithms exist for drawing graphs more or less adapted to the data.
- Remove links promotes a topology emergence in the data set.
- A valued graph, through the distance matrix, allows MDS placements.

The GUI provides a fully operational interface for users with basic functionalities, and advanced settings for expert users. First of all, an expert user has an access to several setting tabs like data table, layout management, clusters identification or filtering operations. He customizes and experiments settings to design a profile adapted to one video collection. Then, users automatically apply this profile on the entire video collection.

3.1. Feature Extraction and similarity

The elementary units are frames selected at constant frequency. Low level features are extracted such as histograms in different colors spaces (RGB, LUV and HSV), gradient orientation histogram, or affine motion model estimation from corner detection.
The system compares each element with the whole other items and fills up full matrices of similarities for each measure. These similarity matrices are transformed into distance matrices and finally exported in a complete graph representation: a node is an image with its descriptors, and an edge between two frames symbolizes a link multivalued by all the distances computed.
The main difficulty here is to choose a good distance related to the task we are willing to achieve. It is possible to compute various distances on color histograms as Bhattacharyya, classical Euclidean and Manhattan distances. Some are more noise-robust, others are faster to compute, some are closer to human perception, and at least their effectiveness depend on data distribution. In an archive context, video formats are very numerous, with potential deteriorations, and so that color quality or resolution may influence the distance choice.

3.2. Graph Layout and clustering

We are now aiming to provide a 2D visualization of these n dimensional data. Two approaches are possible: MDS approaches usually applied for complete graphs with valued edges, and topological placements for non valued links.
There exists a multitude of variants of MDS for reducing n dimensions of the feature space to a 2D space such as the nonlinear projection of Sammon or the Curvilinear Component Analysis (see [7] for an overview of visualization of high-dimensional data). Our approach is to use a customized energy-force model algorithm [9] [10] to achieve the MDS. This model considers 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.
Topological layouts are also useful for drawing graphs with a limited number of edges. The principle is to consider non valued links and to make more readable views (see part 3.4).
Finally, a standard agglomerative hierarchical clustering algorithm is used [11]. A linkage metrics based on the minimum distance between objects is used to obtain clusters of arbitrary shapes. These distances can be parametrized in the interface, allowing the user to obtain relevant results.

3.3. Merging similarities

Users have several distances available, they can switch from one to another and experiment each proximity associated. For a homogenous collection, one or two distances will be adapted in practice, depending of the data set. Moreover, measures can be orthogonal as texture and color. They can bring complementary informations, and users may combine some of them in order to gather relevant data. Two kinds of methods are possible for creating edges merging the values available:

- in the continuous space, by a weighted normalized sum,
- in the discrete space, with image classification algorithms and edge filtering.

3.4. Building a Table Of Video Contents

The idea is to extract a series of successive images at regular interval from a TV program, and manipulate it in the 2D space like a thread or a rosary of pictures.
The first step is to identify the images belonging to the video *backbone*. A *backbone* is a set of frames making up the unifying thread of the video narration. The expert user experiments 2D placements on frames with available measures in order to optimize clusters formation. The *backbone* cluster(s) may be identified in a supervised way or automatically:

- an expert user appoints one or several clusters,
- the cluster which contains the most time covering frames is labeled as the *backbone*.

After this identication, a new graph is built with two time constraints on the temporal proximities values of 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 Rheingold develops the time axis horizontally and spread out thread loops on both sides of the *backbone*. The final visualization emphasizes the video structure and gives a quick overview of its content. The layout has to be read from the left to the right, in the chronological order like in figure 2.

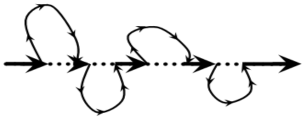

Fig. 2. Time reading of TOVC: horizontal line represents the *backbone*, and successive frames of same sequences are brought together in loops.

4. EXPERIMENTATION

4.1. Backbone identification

Figure 3 shows a layout of a news broadcast of 27 minutes with frames taken every 10 seconds. The distance applied in the MDS algorithm is a weighted normalized sum between two euclidean distances on histograms, one on a 64 binned LUV histogram, and one on a 12 binned gradient orientation histogram (3 amplitudes and 4 directions). If only the color distance is applied, the 2D clustering algorithm merges them in 5 ”TV floor” clusters, corresponding to the camera angles: the anchorman alone, with title announcements, in a wide-shot, in a two-shot, and a last one of a specialist interviewed. With a combination of nearly 80% of orientation and 20% of color, user identifies now two main ”TV floor” clusters: the anchorman (a) and the specialist (b). Anchorman cluster can constitute the *backbone* for creating the final visualization.

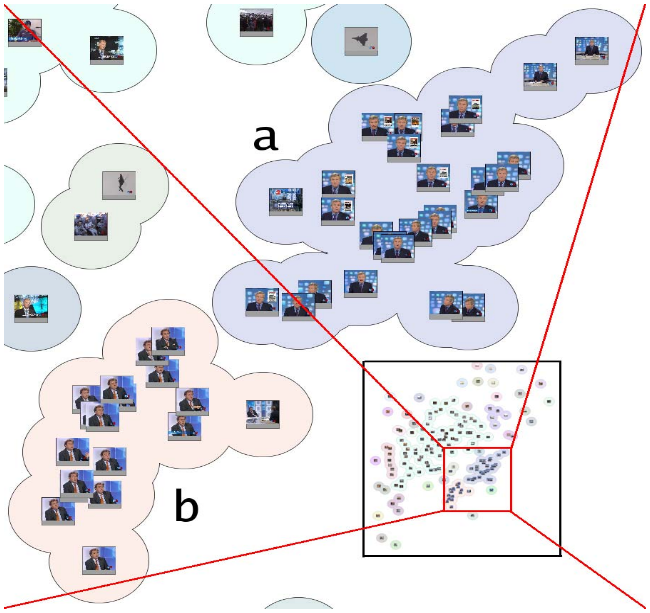

Fig. 3. Zoom of a layout applied on 161 frames of a newscast. The MDS algorithm uses a weighted normalized sum of color and orientation distances in order to improve the *backbone* clusters identification.

4.2. Table Of Video Contents

This part shows several examples and discuss on advantages to use these visualizations. Figure 4 is the TOVC version of previous data set used in figure 3. The figure 5 is a short newscast of nearly 3 minutes with frames taken every 3 seconds. Figure 6 is a TOVC of a deaf and hard of hearing news programs with a duration of 4 minutes with one frame per second.

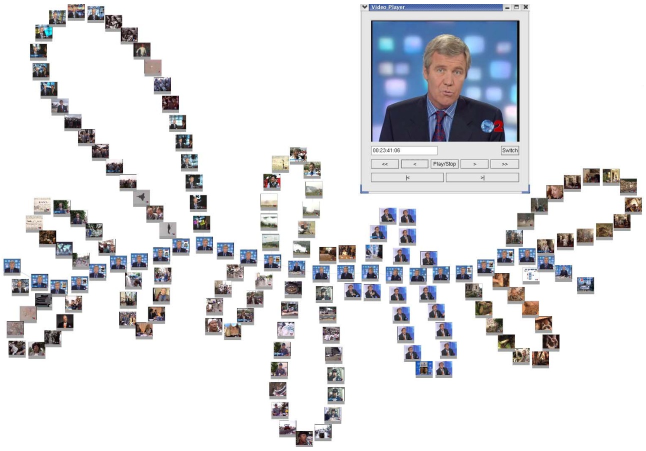

Fig. 4. TOVC of the same news program in figure 3: frames are reorganized in an esthetic visualization connected with a player for a video browsing

These overviews allow very readable video content and quick comparisons between these news programs.

- The *backbone*, the horizontal line, represents the studio floor frames with one or two newscasters like in figures 4 and 6. Our system is generic enough for identifying the video main thread even visual format are very different like in figure 5 where the *backbone* part contains exclusively on-screen graphics.
- Each loop is a report sequence or an interview shot. Users can count easily their insertions in the TV floor sequences. For the standard news there is almost the same number of reports, but for the deaf and hard of hearing news, reports are less numerous. This indicates a typical structure where the sign language interpreter on stage and the speech newscasters limit visual content.
- Loop length indicates sequence duration. Generally, news edition balances the report durations, like in the short newscast of figure 5 for making a fast overview of the daily events. But duration exception can be a good indication to notice a particular event, like in figure 4 where the 4^{th} loop long length shows a link-up broadcasting on this day.
- Dominant colors gives interesting information: for instance, the blue at the 10^{th} to 12^{th} loops in figure 4 identifies an interview on TV floor. Rich colors in the two last reports indicate probably a less dramatic content than the top news headline. In figure 5 the very red and yellow colors in the first report indicate a disorder or a fire.

The method can be applied in another kind of program as sport. Figure 7 displays a result on a video extract with frames taken every 5 seconds during 14 minutes of pole vaulting events. Here, the difference with the previous visualization is that the expert user does not

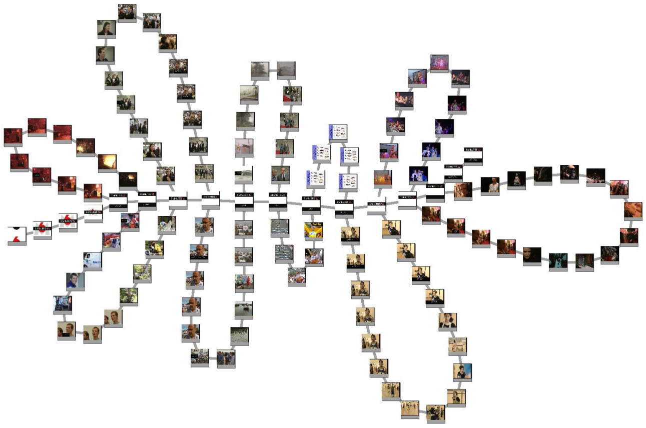

Fig. 5. TOVC of a short newscast of 3 minutes with frames taken every 3 seconds. There is no anchorman.

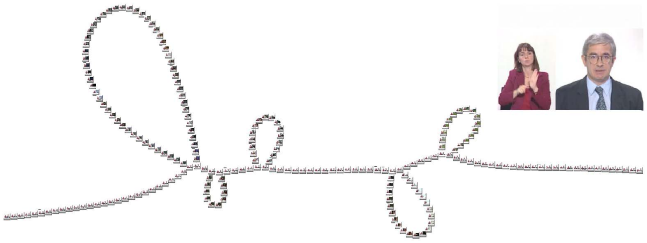

Fig. 6. TOVC of a deaf and hard of hearing news about 4 minutes with frames taken every second. Fiew visual reports are present in the *backbone*.

identify a backbone cluster but ”loops” clusters. A global motion estimation on successive frames allows to associate 6 local signals from an affine model for each frame. The measure applied is the cross-correlation on the x translational coefficient. The structure of the final view is very similar to previous TOVC: 7 pole vaulting trials are represented and the zoom part illustrates one jump event and its slow motion playback.

5. CONCLUSION AND FUTURE WORK

In this paper, we have presented a framework for multimedia content visualization and navigation. We propose a Table Of Video Contents adapted for structured programs as news, magazines or sport events. Connected with a player, users can have handy navigation through a global view. The visualization shows the video structure and gives almost instantaneously relevant informations such as sequence number, relative durations, dominant colors.... Users can zoom on loops and play video at corresponding time code by clicking on frames. This visualization can be seen as a visual signature of a video program and allows quick comparisons between documents of a homogenous collection (newscasts of the same channel or from various channels the same day ...).
In future work, we want to converge to a full automatic process for newscasts, maybe with learning algorithms. Usuability test with users are in progress and will be presented soon. For less structured video, like fictions or documentaries, user supervision is important to lead more advanced clustering for identifying multiple *backbones* or narrative threads. In these cases, structures may be more complex and may induce other graph topologies. So new visualizations will be needed and experimented, that is yet conceivable within our framework.

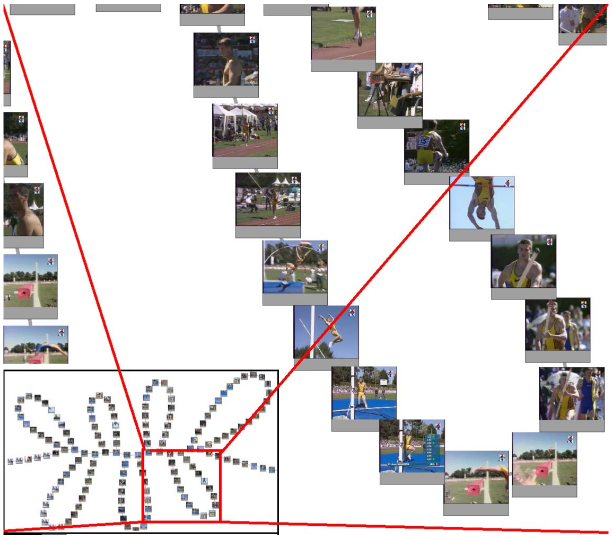

Fig. 7. Zoom on a TOVC of pole vaulting events created from a measure on motion estimation: each loop represents an event and its overcrank.

6. REFERENCES

[1] W. Hurst and P. Jarvers, “Interactive, dynamic video browsing with the zoomslider interface,” in *IEEE ICME*, 2005.

[2] A. Amir, D. Ponceleon, B. Blanchard, D. Petkovic, S. Srinivasan, and G. Cohen, “Using audio time scale modification for video browsing,” in *HICSS*, 2000.

[3] Y. Taniguchi, A. Akutsu, and Y. Tonomura, “Panoramaexcerpts: extracting and packing panoramas for video browsing,” in *ACM MM*, 1997, pp. 427–436.

[4] C. Pal and N. Jojic, “Interactive montages of sprites for indexing and summarizing security video,” in *IEEE CVPR*, 2005.

[5] E. Kijak, G. Gravier, L. Oisel, and P. Gros, “Audiovisual integration for tennis broadcast structuring,” *Multimedia Tools and Applications*, vol. 30, pp. 289–311, 2006.

[6] M. Campanella, R. Leonardi, and P. Migliorati, “Future viewer: an efficient framework for navigating and classifiying audio-visual document,” in *IEEE ICME*, 2005.

[7] J. X. Li, “Visualization of high-dimensional data with relational perspective map,” in *Information Visualization*, 2004, vol. 3, pp. 49–59.

[8] M. M. Yeung and B. L. Yeo, “Time-constrained clustering for segmentation of video into story units,” in *ICPR*, 1996.

[9] T. M. J. Fruchterman and E. M. Rheingold, “Graph drawing by forcedirected placement,” in *Software - Practice and Experience*, 1991, vol. 21, pp. 1129–1164.

[10] P. A. Eades, “A heuristic for graph drawing.,” in *Congressus Numerantium*, 1984, vol. 42, pp. 149–160.

[11] S. Guha, R. Rastogi, and K. Shim, “CURE: an efficient clustering algorithm for large databases,” in *ACM SIGMOD*, 1998.