Interactive Tools for Constructing and Browsing Structures for Movie Films

Riad Hammoud and Roger Mohr

8th ACM International Conference on Multimedia
November 2000

Abstract: This paper presents a prototype for constructing, browsing and using structures for movie films based on content image analysis only. The goal of the structuring is to facilitate the user's access to the video content (non-linear navigation, etc.). Our prototype provides for the "editor user" advanced tools for structuring the video at its low-level (e.g. shots, key-frames) and high-level structures (e.g. groups of objects, scenes). Also, it provides for the "end user" flexible interfaces for browsing and using constructed video structures.

Against the previous version of this prototype [1], we mainly improved the matching and the clustering of segmented objects. Also, constructing and browsing the high-level scene structure are now available.

1 Constructing structures

Figures 1 and 2 illustrate the system designed for building shots, clusters and scenes structures. It is developed in C++/Ilog-Views and portable under Unix and Linux. In the following, we briefly list its main functionalities. Some modules which implement these functionalities are not completely integrated in the system.

▶ **Basic segmentation.** The partitioning into shots is done firstly using the dominant motion approach. The estimated motion is then used to localize and track mobile objects within shots [2] (figure 1 -bottom). The static objects are manually segmented and tracked.

▶ **Characterizing and matching individual objects.** Individual occurrences of tracked objects are characterized by three different features: global color histograms, color correlograms and local differential invariants. The matching process of individual objects is performed on each descriptor separately. A *linear fusion* of the matching results is adopted when multiple descriptors are used. The weight of each descriptor is fixed by the "editor user" which decides the importance of each descriptor (for example, for this sequence colors are more discriminant than geometric informations).

▶ **Characterizing of tracked objects.** Due to the variable appearance of objects during tracking and the acquisition in poorly constrained dynamic scenes, the matching of individual objects using classical features gives poor results. In order to increase the robustness of existing features, we use the Gaussian mixture densities to model the intra-shot variability of each tracked object [6].

▶ **Clustering of objects.** Both supervised and unsupervised clustering of objects are implemented (figure 2). (1) The user selects by the mouse some tracked objects, considers them as "models" or classes, and then assigns to them all other objects. Currently, this technique is applicable only on modeled tracked objects in the color histogram feature space where the mixture classifier is used to identify classes of individual objects.

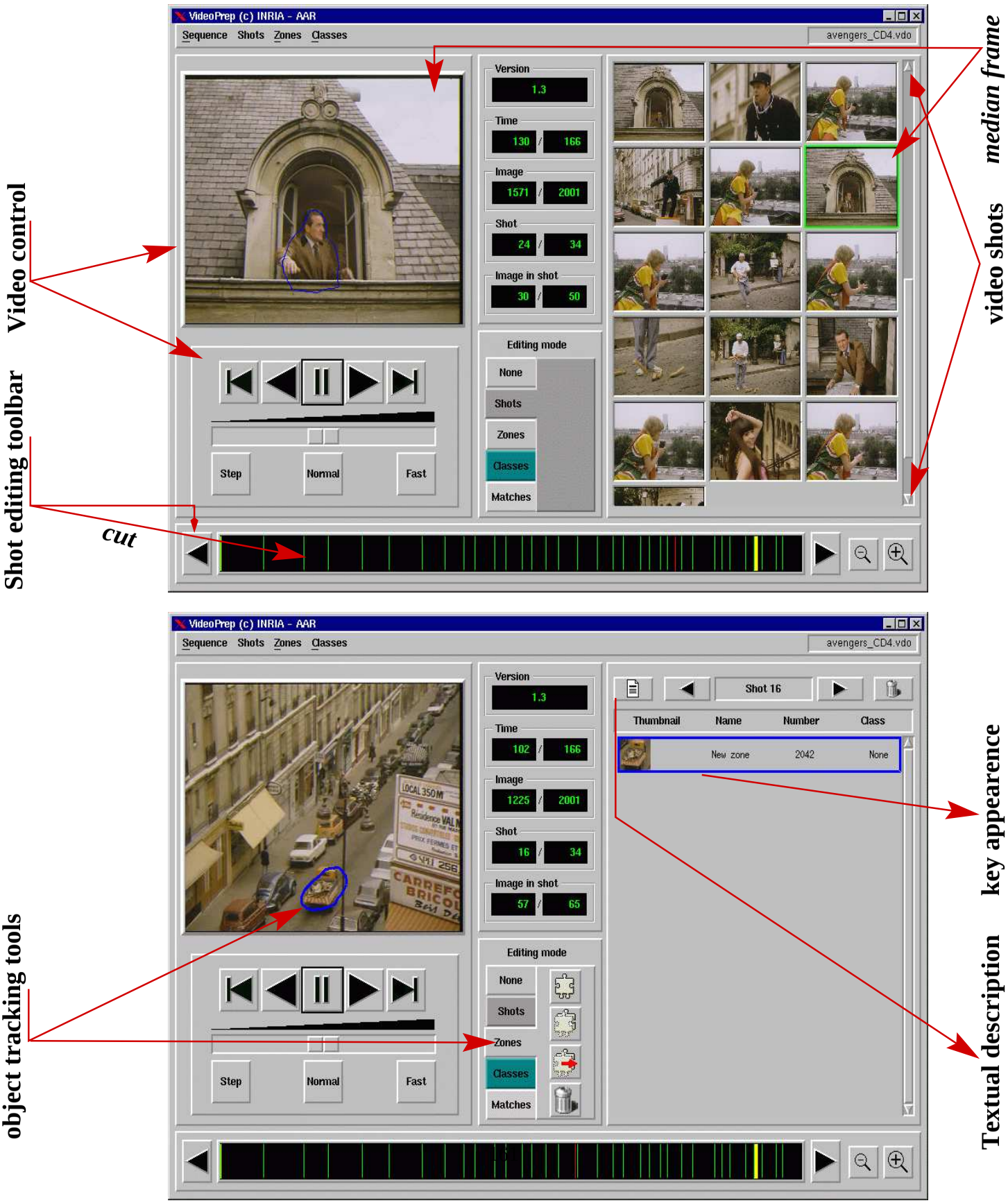

Figure 1: System for constructing video structures : partitioning into shots (top), tracking of objects (bottom)

(2) To avoid a manual selection of "object models", the Ascendant Hierarchical Classification algorithm is used to automatically identify clusters of objects based on different implemented descriptors. The unsupervised classification based on estimated Gaussian mixtures for tracked objects gives good results [5]. The module of this method is not yet integrated in the system.

▶ **User in the loop.** Practically, it is very difficult to perform a perfect clustering of this kind of noisy data (occlusions, illumination changes, etc). The system provides some interactive tools to correct the results of the automatic clustering: (1) *select/browse* clusters at different levels of the hierarchy, (2) *Drag* a badly classified object and *Drop* it into another cluster or a new one.

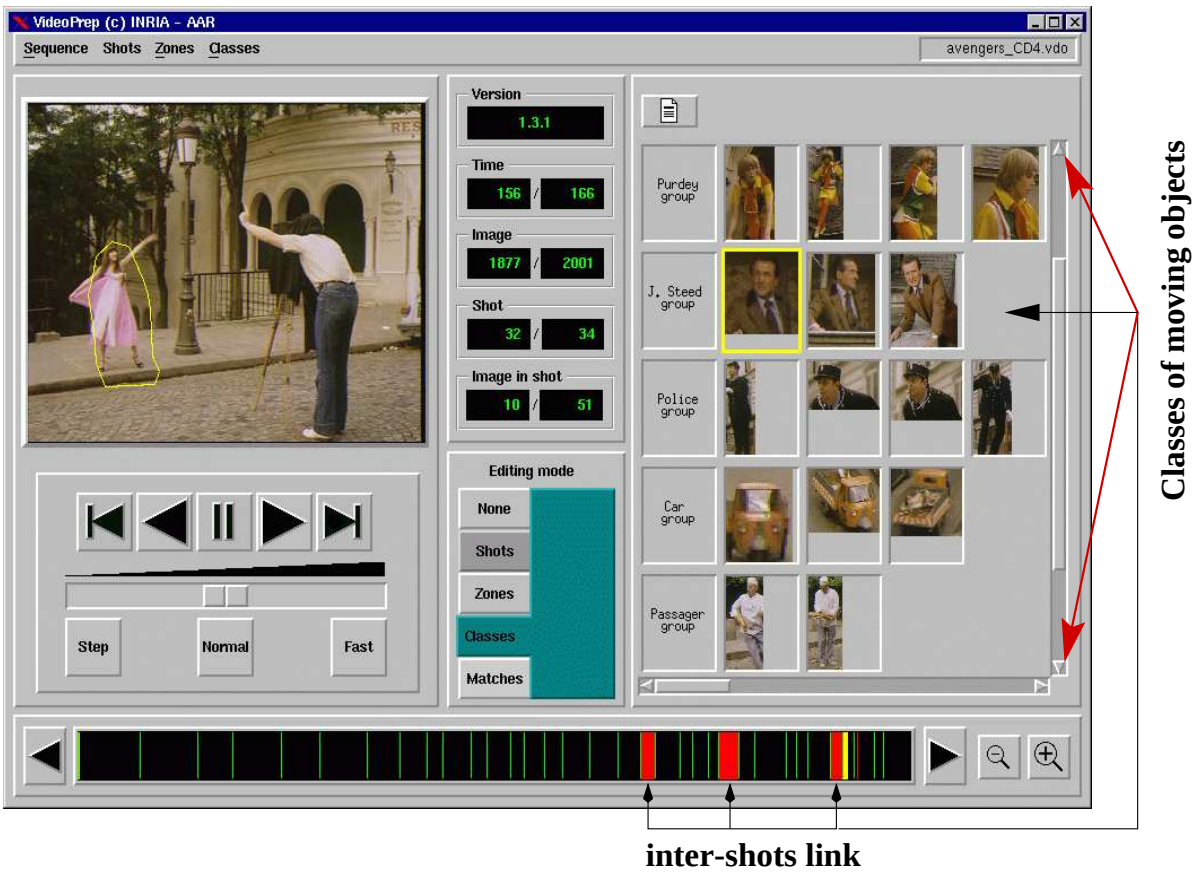

Figure 2: System for constructing video structures : grouping tracked objects into clusters

▶ **key-frames extraction.** A Key-frame is an existing frame which can represent the whole set or a subset of frames of the shot. Usually each shot is represented by only the first frame. In general shots are dynamic, so a single key-frame is not sufficient to represent effectively the content. The modeling of appearances of a tracked object consists in grouping similar views together. An efficient technique is to select from each group of similar views the median image as a key-frame (see [5]).

▶ **Scenes extraction.** A video scene is defined as a collection of semantically related and temporally adjacent shots, depicting and conveying a high-level concept or story. Our approach to extract scenes is an extension of the method of [3]. The method consists firstly in grouping similar shots, of the same predefined temporal window and the same "narrative sequence", into clusters, then exploring the temporal graph of clusters to extract scenes. The temporal relations of Allen (meets, before, ...) are used to connect the nodes of the graph. A scene is formed by merging nodes (clusters) of a sub-graph, which does not contain a temporal relation of type "meets" that can disconnect it into two other sub-graphs. The extension of this method is done at the clustering stage. Three descriptors are used to match similar shots represented by key-frames: histograms, correlograms and the number of similar objects in two compared shots. On each descriptor the hierarchical classification algorithm is performed where the number of clusters is determined using a predefined threshold. Each one of these descriptors summarizes differently the content of a shot. So, the obtained clusters by different descriptors are not necessarily similar. A distance that measures the intersection between two clusters of two different descriptors is computed. Here the goal is to deduce from the three sets of clusters only one set. Based on this, we construct the temporal graph of clusters from which the scenes are extracted as explained previously. The experimental results depict in [4] shown the performance of this extended method against its original form. The related modules to this functionality are in the course of being integrated into this system.

2 Browsing and using structures

Once the constructing structures for a movie film is achieved, the "end-user" has the ability to explore the content of the film in a new way. The cluster structure defines in the movie film links between objects. At this level, the end-user clicks an object of interest (actor, car, ...), jumps to its next or previous occurrence in the film, plays the corresponding shot, plays the corresponding action, discovers a related WWW link, etc. The scene structure allows the end user to access the video document as a book (with a table of contents). Each scene describes a story or action of the film.

Figure 3 illustrates the end-user interface for browsing and navigation in the different structure level of the movie film. This interface is developed in `Java` in order to be used on different user platforms. The next version of this application will be accessible on the World Wide Web.

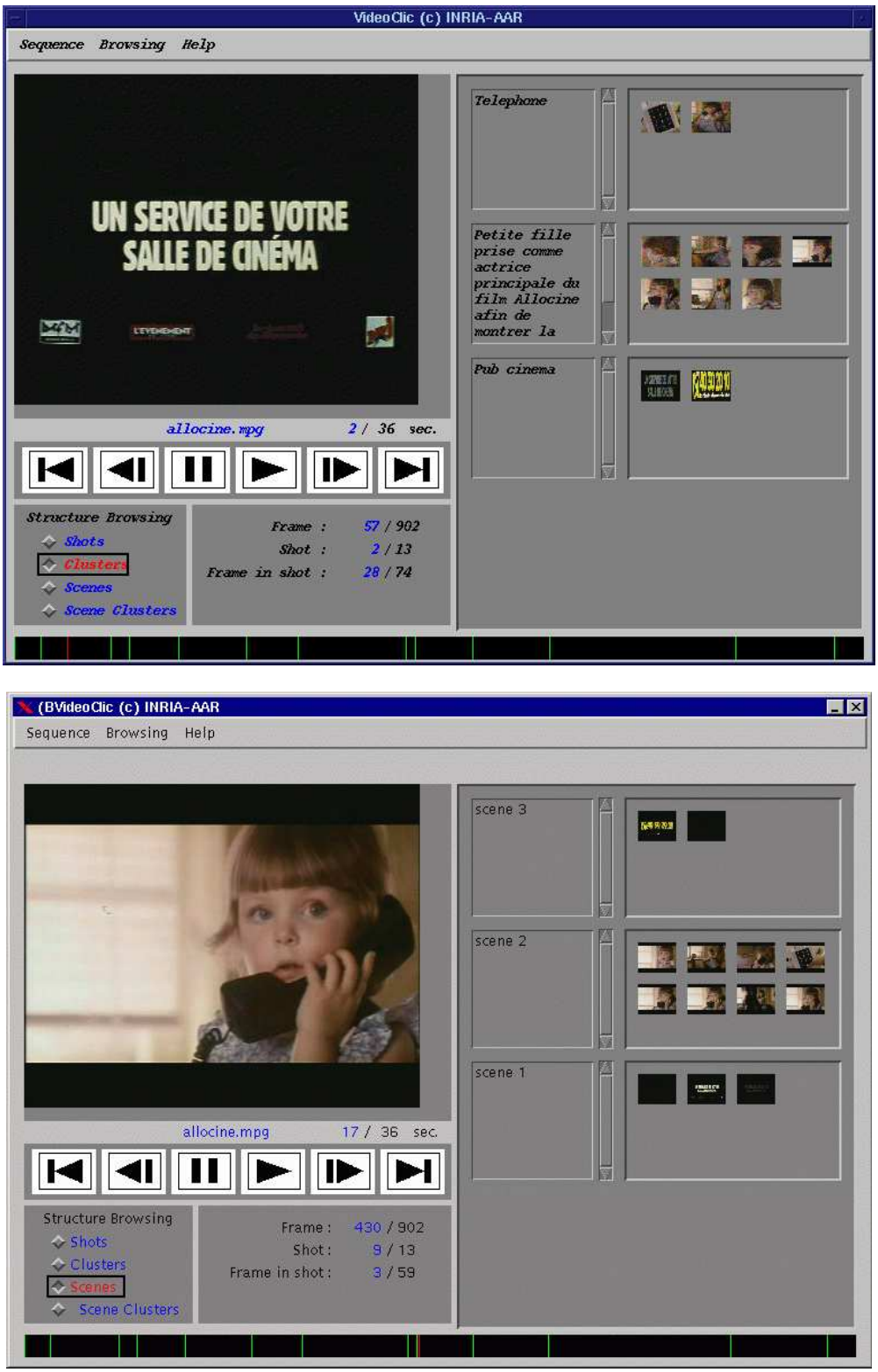

Figure 3: End-user system for browsing video structures : (top) browsing clusters, (bottom) browsing scenes.

References

[1] S. Benayoun, H. Bernard, P. Bertolino, M. Gelgon, C. Schmid, and F. Spindler. Structuration de vidéos pour des interfaces de consultation avancées. In *CORESA 98 – Journées d'études et d'échanges COmpression et REprésentation des Signaux Audiovisuels.*, June 1998.

[2] M. Gelgon and P. Bouthemy. A region-level graph labeling approch to motion-based segmentation. In *CVPR*, pages 514–519, Puerto Rico, June 17-19 1997.

[3] R. Hammoud, L. Chen, and F. Fontaine. An extensible spatial-temporal model for semantic video segmentation. In *First Int. Forum on Multimedia and Image Processing, Anchorage, Alaska*, 1998.

[4] R. Hammoud and D. G. Kouam. A mixed classification approach of shots for constructing scene structure for movie films. In *Irish Machine Vision and Image Precessing Conference*, pages 223–230, The Queen's University of Belfast, Northern Ireland, 31 August-2 Septembre 2000.

[5] R. Hammoud and R. Mohr. Building and browsing hyper-videos: a content variation modeling solution. *Pattern Analysis and Applications*, 2000. Special Issue on Image Indexation, Submitted.

[6] R. Hammoud and R. Mohr. Mixture densities for video objects recognition. In *International Conference on Pattern Recognition*, volume 2, pages 71–75, Barcelona, Spain, 3-8 September 2000.