A mixed classification approach of shots for constructing scene structure for movie films [1] [2]

R. Hammoud and D. G. Kouam
MOVI-INRIA Rhône-Alpes and GRAVIR-CNRS
655 avenue de l'Europe, 38330 Montbonnot Saint Martin, FRANCE
riad.hammoud@inrialpes.fr

I*rish* **M***achine* **V***ision and* **I***mage* **P***rocessing Conference, pages 223-230*
September, 2000

Abstract In order to facilitate the user's access to a movie film the scene structure should be constructed. A general framework for constructing scenes consists in clustering the shots into groups and then merging overlapped groups to extract scenes. The clustering of shots represents the most challenging task and should be done correctly. This paper presents a mixed approach to cluster shots using both features computed on key-frames and the corresponding localized moving objects. Two shots are matched on the basis of color-metric histograms, color-metric autocorrelograms and the number of similar objects localized into them. Using the hierarchical classification technique, a partition of shots is identified for each feature separately. From all these partitions a unified partition is deduced based on a proposed distance between their components (clusters). The components of the resulted partition are then linked together using the temporal relations of Allen in order to construct a temporal graph of clusters from which the scenes will be extracted. The experimental results of the proposed approach on three real-world video sequences demonstrate its performance. A comparative study between the original approach which uses only one descriptor and the extended one proposed here is analyzed and reported.

1 Introduction

In the context where the digital video data are available on the web in great quantity, the ability to understand and structure them becomes of greater importance, for the purpose of facilitating user's access to the video content. The automatic video structuring represents the fundamental task for many recently emerging multimedia applications: video browsing and navigation , content-based indexing and video summarization and trailers [17][7].

Commonly, two basis tasks for identifying the low-level structure of a video document are performed first: *shot boundary detection* and *key-frames extraction*. A *shot* is defined as an unbroken sequence of frames recorded from a single camera, which forms the building block of a video. Beyond the shot level an abstraction level can be constructed by mapping the entire shot to a small number of representative frames, called *key-frames* [17].

The drawbacks of low-level structures, shots and key-frames, are that they (1) contain too many entries to be efficiently presented to the user - for example there are 3225 shots

[1] This work is supported by Alcatel CRC Grant Alcatel-Inria No. 198G098.
[2] Demos of this work are available at `http://www.inrialpes.fr/movi/people/Hammoud/`

in *October* of S.M. Eisenstein [12] to be presented to the user; and (2) do not capture the underlying semantic structure of the video based on which the user might wish to browse/retrieve [14] [16] [4].

In recent years the research in this area focuses on the construction of high level structures (*groups* and *scenes*) for movie films [14] [8]. The clustering of shots (or segmented objects in shots) creates links in the video ([6], [8]). In this case the access to the video is non-linear where the user can navigate in the content of a constructed group of non continuous shots (objects) in time. However, the *groups* level does not reflect the semantics of the video. A *video scene* is defined as a collection of semantically related and temporally adjacent shots, depicting and conveying a high-level concept or story. The people watch the video by its semantic scenes not the physical shots or key-frames. The rest of this paper focuses on the construction of scenes for movie films.

Many approaches have been proposed in the literature for constructing scene structure [16] [15][4]. They basically perform two fundamental steps: "clustering of shots" into groups and "merging of overlapped groups" into scenes. The color histogram is commonly extracted from the representative key-frames of a shot. Based on this descriptor the visual similarity between shots is detected.

The clustering of shots represents the most challenge task of this work. It should be done correctly since the scene level is constructed on the top of the cluster one. It is obvious that two similar shots of the same scene do not have necessary the same global color distributions (histograms). Two different shots with the same global color distribution will be matched similar. On the other hand, the color histogram is very sensible to partial changes in the decor of a scene. In contrast, the content of shots, like localized objects, may be considered as important clues for detecting robustly similar shots.

Based on this short discussion, the similarity between shots is based upon many features. In this paper, a mixed approach to cluster shots using both features computed on key-frames and the corresponding localized moving objects. Two shots are matched on the basis of color-metric histograms, color-metric auto-correlograms and the number of similar objects localized into them. Using a hierarchical classification technique, a partition of shots is identified for each feature separately. From all these partitions a unified partition is deduced based on a proposed distance between their components (clusters). Following [4], the components of the resulted partition are then linked together using the temporal relations of Allen [1] in order to construct a temporal graph of clusters from which the scenes will be extracted.

The organization of this paper is as follows. The next section describes the recent works in this area. Section 3 details the proposed approach. Experimental results are given in section 4. A comparative study between the original approach which uses only one descriptor and the extended one proposed here is analyzed and reported in section 5. These experiments on three real-world video sequences demonstrate the performance of the proposed approach. Concluding remarks are given in section 6.

2 Related work

To construct the scene level beyond the shot one, Yeung et al. [16] identify first groups of shots using the hierarchical clustering algorithm, and then extract scenes form the *Scene Transition Graph* by merging overlapped clusters. Hammoud et al. [4] use an adaptive clustering algorithm to group shots. Then, they use the temporal relations of Allen (meets, before, overlaps and during) to link the formed clusters together. The linked clusters by "overlaps" and "during" relations are merged together. The "meets" relation between two non-merged clusters defines the boundaries of a scene. The "before" relation is used to describe the boundaries of a "narrative sequence". A "narrative sequence" is a set of scenes of the film and it is detected when a gradual transition is identified. Both Yeung and Hammoud use the global color histogram to represent the visual content of a video shot (abstracted by a key-frame). Two shots are considered similar when the visual similarity between them is less than a predefined threshold and if they belong to the same predefined "temporal window" [16] or to the same "narrative sequence" [4].

Recently, Rui et al. [15] proposed an intelligent unsupervised classification technique to identify clusters of shots. At the shot level, they characterize the temporal information by extracting cumulated difference of histograms over all frames of the shot (called "shot activity"). The spatial information is summarized by the color histogram computed on the first and the last frames of shots. The distance similarities between shots, based on these descriptors, are normalized firstly, and then linearly combined using automatically determined thresholds. Our approach is close to this work by the use of multiple descriptors to characterize the video shot. But, the distance similarities based on these descriptors are used independently by the clustering process of shots. Then a unified partition of shots is deduced from all obtained partitions using a distance between their components (see the next section).

3 A mixed approach for constructing scene structure

The construction of the scene level requires a set of tasks to be done robustly: *basic segmentation of the video*, *characterizing of shots*, *clustering of shots* and *extraction of scenes*. In the following a description of these tasks is given.

3.1 Basic segmentation

To analyze a movie film a temporal segmentation into shots is done firstly. In this work we use the method of [2]; It relies on a robust, multi-resolution and incremental estimation of a 2D affine motion model between successive frames, accounting for the global dominant image motion. This method is also used to detect gradual transitions between shots (dissolves, black transitions, ...). These gradual transitions define the "narrative sequence" layer in the video.

Within a shot, the entities like moving objects can be detected and used later in the identification process of similar shots. Many research effort has been given in this area [10] [3]. The motion approach is widely used. In our approach, and following [3], the

estimated motions during the *cut detection* process are used to localize and track mobile objects within shots.

3.2 Characterizing of shots

As mentioned in the introduction, a shot forms the building block of a video, and a cluster of similar shots forms the basic unit of a video scene. In this work, each shot is represented by only one key frame (the middle frame). However, more sophisticated approaches can be implemented [5], where multiple frames would be selected based on the density variation of the content of a shot.

The similarity between shots should not be necessary limited to the global color distribution. Two shots of the same scene may have very close contents like detected persons and very far global color distributions due to partial changes/occlusions in the decor of the scene. Figure 1 illustrates an example of two successive shots (represented by their middle images), of the same video scene, which have very close tracked objects (the yellow car) and totally different image backgrounds.

Figure 1: Two shots with similar tracked objects (yellow car) and different backgrounds

However, the global descriptors like color histograms and auto-correlograms [9] are useful for the recognition process when there is no detected objects in the framework of shots or when the appearances of the same object in different shots are very variable (partial occlusions, ...).

The color histogram captures only the color distribution in an image (or region) where the color auto-correlogram expresses how the spatial correlation of color changes with distance. Thus, the auto-correlogram is one kind of spatial extension of the histogram [9]. According to the type of information dominating in each shot, each descriptor allows a better comparison of a particular aspect of the shots, and consequently to lead to a comparison of the shots according to this aspect. In the presented framework, each video shot is characterized by its content (detected objects) and the color histogram and the color auto-correlogram both computed on the whole key frame.

3.3 Matching of shots

Two shots are matched on the basis of the global color histograms and color auto-correlograms via the L_1 distance measure. This distance is commonly used when comparing two feature vectors, because it is simple and robust.

$$L_1 = \sum_{i \in [1,n], k \in [d]} \frac{\mid h_{c_i 1}^{(k)} - h_{c_i 2}^{(k)} \mid}{1 + h_{c_i 1}^{(k)} + h_{c_i 2}^{(k)}} \quad (1)$$

where n is the total number of colors of an histogram and $h_{c_i j}^{(k)}$, $i = 1..n$, $j = 1, 2$, $k = 0$ or $k \in [1..d]$, represents the frequency of pairs of pixels of color c_i at a distance k in the jth image. For color histograms k is equal to zero.

In our experiments, the number of colors, n, was fixed to 64 and the spatial distance, d, used in the computation of auto-correlograms was fixed to 5 (this value of d is recommended by the author of this descriptor [9]).

Also, two shots are matched on the basis of the number of similar objects localized into them. The similarity between objects is determined using the above distance measured between auto-correlograms where a predefined threshold is determined in an interactive way.

Let S_i and S_j be the two shots of η_i and η_j objects detected into them respectively. Let η_{ij} the number of similar objects between S_i and S_j. The proposed metric distance to measure the number of similar objects between S_i and S_j is as follows:

$$D(S_i, S_j) = 1 - \frac{\eta_{ij}}{Max(\eta_i, \eta_j)} \quad (2)$$

For example, if S_i and S_j have the same number of objects and if these objects are matched similar, $D(S_i, S_j)$ will be close to zero, and so these two shots are considered similar according to their content.

3.4 Clustering of shots

By definition, a scene expresses an action of a short time duration and it belongs to a "narrative sequence" (see section 2). Based on this, a clustering strategy should take into account this temporal dimension and the boundaries of narrative sequences. Such considerations will avoid to classify two similar shots together if they are far in time [16] or/and if they do not belong to the same narrative sequence [4].

As described in the previous section, the matching of shots is done using three similarity distance measures. Our approach to cluster shots integrates these three distance measures and the above temporal criteria as follows.

Firstly, the complete-link hierarchical classification algorithm is performed on each proximity matrix of a similarity distance [11]. The number of clusters is determined using a predefined threshold of similarity; when the minimal distance between groups of shots is greater than the predefined threshold, the clustering process is stopped. Notice that, the threshold of similarity should be very close to zero in order to avoid a miss-classification of shots. This leads to a set of clusters with few number of elements. During the clustering process, the temporal distance between two grouped shots is checked as also if they belong to the same "narrative sequence". The "narrative sequences" are already identified during the temporal partitioning of the video into shots (see section 3.1) and the temporal threshold is determined in an interactive way.

Secondly, a unified partition of shots is deduced from the three partitions formed as described above. The proposed algorithm here to construct the final partition of shots consists in merging clusters of different partitions which have a certain number of common shots. It is based on the fact that the similarity between two shots is not always a linear combination of different descriptors, but this similarity may be reached using only one descriptor. Before to describe the two steps of this algorithm, we propose the following distance, $d_{\cap}$, to measure the intersection between two clusters of shots, ω_1 and ω_2:

$$d_{\cap}(\omega_1, \omega_2) = 2 - card(\omega_1 \bigcap \omega_2)(\frac{1}{card(\omega_1)} + \frac{1}{card(\omega_2)}) \quad (3)$$

The two parts of the mixed classification approach of shots are the following:
Clumping classification:

0- Let $\Omega_i, i = 1, ..., n$ be the n partitions constructed for the n different descriptors with $n \geq 2$. Initialize i to 1.

1- For each pair of clusters (ω_k, ω_l) that $\omega_k \in \Omega_i$ and $\omega_l \in \Omega_{i+1}$, if $d_{\cap}(\omega_k, \omega_l) \leq$ *Threshold*, then merge ω_l into ω_k and remove ω_l from Ω_{i+1}.

2- $\Omega_i = \Omega_i \bigcup \Omega_{i+1}$, $i = i + 1$. If $i < n$, goto 1, else goto 3.

Horizontal merging:

3- For each pair of clusters (ψ_k, ψ_l) that ψ_k and $\psi_l \in \Psi$ (the new constructed set of clusters), if $d_{\cap}(\psi_k, \psi_l) \leq$ *Threshold*, then merge ψ_l into ψ_k and remove ψ_l.

The *clumping classification* procedure produces a unique set of clusters of shots Ψ. However, some of these clusters are not disjoint. The *Horizontal merging* process consists in merging overlapping clusters together. The result is a partition of distinct clusters of shots.

3.5 Extraction of scenes

Generally, a scene/action of a movie film is projected as a continuous flow of shots in time. An intuitive way to construct the scenes is to adopt a strategy of merging clusters of shots as in [4] and [16]. In this section, an overview of the method of [4] to extract the scenes is given. This method will be used in our approach.

The shots are grouped into homogeneous clusters with respect to their temporal locality. The components of a cluster are not always successive in time and they are interleaved by the components of another clusters. Figure 2 (left) illustrates a representation of clusters on the time axis. The components of each clusters are displayed in their ascending order of the time code (frame number). Two shots may be linked by a simple "cut" or a gradual transition (GT). A gradual transition between two shots defines the boundaries of a narrative sequence. Notice that the searching of scenes will be done in each narrative sequence separately. Based on this representation of clusters versus the time, the sequential and parallel relations of Allen, Meets, Before, Overlaps and During, are generated between

clusters for the purpose to link them [4]. The Overlaps and During relations are generated between two intersected clusters in time while the Meets relation links two successive clusters. When two clusters are successive in time but they belong to different narrative sequences a Before relation is generated.

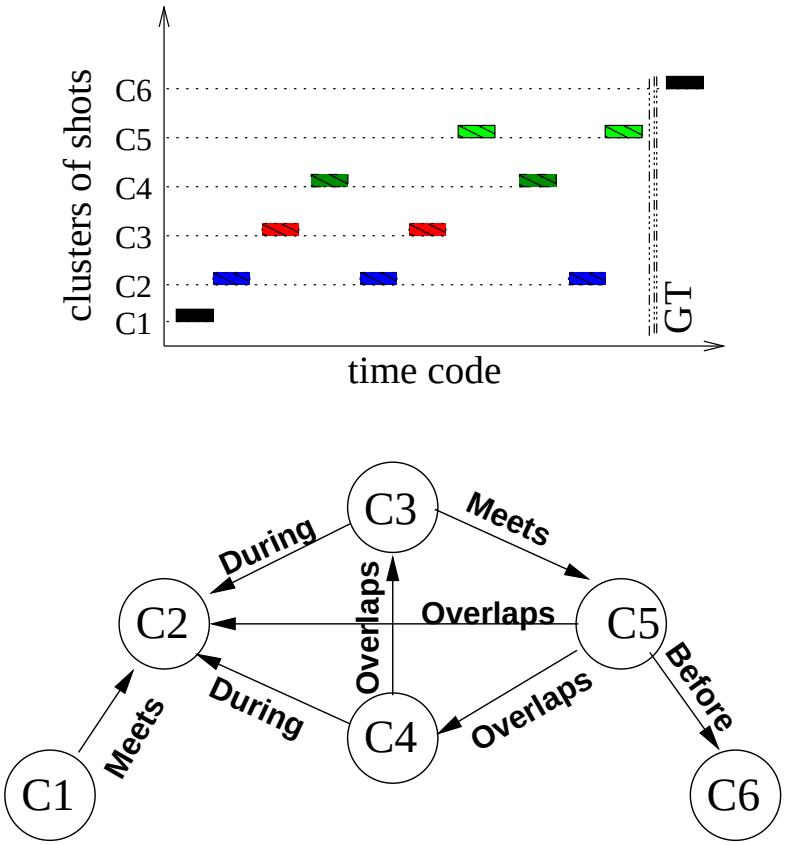

Figure 2: Representation of clusters of shots versus the time code (top) and the corresponding temporal graph of clusters (buttom).

At this stage, a temporal graph describes the movie film where the nodes are the formed clusters of shots and the edges represent the temporal relations between them. Figure 2 (right) illustrates an example of a temporal graph of clusters. The extraction of scenes is performed by exploring this temporal graph. Each pair of nodes linked by Overlaps or During relations are merged. This produces a new temporal graph of clusters where the edges are only Meets or Before relations and the nodes represent continuous blocks of shots in time. These blocks of shots define the scenes of a movie film. The Meets relations define transition between scenes of the same narrative sequence and the Before relations define transition between scenes of different narrative sequences.

4 Experimental results

The proposed approach has been experimented on three real-world video clips extracted from the *Avengers* and the *Dances with wolves* movie films, given by the "National Institute of Audiovisuel" in France (INA). The segmentation into shots and moving objects are done on the MPEG decompressed format of these videos. The experimental results are shown in table 1 where $\sharp Scenes$ denotes the number of scenes detected by the proposed approach. The number of frames, shots, segmented objects in key-frames, and the number of final clusters are denoted by $\sharp F_r$, $\sharp S_h$, $\sharp O_{bj}$ and $\sharp C_{lu}$ respectively. The evaluation of such a work is still a difficult task because there are no standard norms to define the boundaries of scenes. Following [15], the two measures of the effectiveness of the construction of scenes "false negatives" (false $\ominus$) and "false positives" (false $\oplus$) are also shown. The "false negatives" indicates the number of scenes missed by the algorithm (for example, when two scenes are detected in a single one this measure is increment by one); and "false positives" indicates the number of scenes detected by the algorithm but which are not considered as scenes by human. These measures for scene's boundaries are obtained from subjective tests. Multiple human subjects are invited to watch the video clips and then asked to give their own structures. The structure that most people agreed with is used as the ground truth of the experiments.

Video name	$\sharp$					$\sharp$false	
	F_r	S_h	O_{bj}	C_{lu}	S_{cenes}	$\ominus$	$\oplus$
Dances with wolves	10000	70	89	17	7	0	0
Avengers1	1804	24	65	12	5	0	0
Avengers2	5341	72	144	20	8	1	0

Table 1: Scene results by the mixed approach

The reported experiments here are done on a limited, but variable, video database. Each video is decomposed of a set of scenes of different lengths. The results shown in table 1 demonstrate the performance of the proposed approach.

5 Comparative analysis and Discussion

For a comparative study the approach is evaluated when the similarity between shots is detected using each descriptor separately without applying the mixed strategy of clustering (see section 3.4). The table 2 summarizes the test results on the first video clip of the Avengers TV movie (*Avengers1*). The parameters (thresholds of spatial/temporal similarities between shots) which have been already determined in an interactive way, when the mixed approach was performed (table 1), are used the same in this experiments.

The measure of similarity between shots is based upon many features. This fact is

confirmed by the above experiments. When the construction of scenes is performed using only one descriptor, there is a significant number of missed and false detected scenes. There are many detected scenes decomposed of only one shot. That means the used descriptor for matching is not dominant in these shots. The use of multiple descriptors as proposed by the mixed approach improves the results.

Descriptor	$\sharp C_{lu}$	$\sharp S_{cenes}$	$\sharp$false $\ominus$	$\sharp$false $\oplus$
histogram	15	7	2	2
auto-correlogram	14	11	0	6
$\sharp$ of similar objects	15	8	0	3

Table 2: Scene results on the *Avengers1* movie film; The scenes are constructed using each descriptor separately.

6 Conclusion and perspectives

One challenging problem addressed by the **M**oving **P**icture **E**xpert **G**roup, MPEG-7, is the identification of the scene structure for a movie film [13]. The scene structure allows the end user of the interactive video [7] to access to the video document as to a book (with a table of content). Each scene describes a story or an action of the film.

In this work we have presented a method for identifying the scene structure for real-world movie films. The main part of the proposed method, is the mixed classification of shots. The similarity between shots is based upon multiple descriptors. Two shots are matched on the basis of color-metric histograms, color-metric auto-correlograms and the number of similar objects localized into them. In fact the similarity between two shots is not always a linear combination of different descriptors, but it may be reached using only the dominant descriptor in these compared shots. The mixed classification approach consists in identifying clusters of shots using each descriptor separately (the hierarchical classification algorithm was used). Then, the three obtained partitions of clusters are merged together, based on a distance that measures the degree of overlapping between clusters.

The experiments of the proposed approach on three real-world movie films demonstrate its performance. On the other hand, the reported comparative study confirms the powerful of mixing multiple features as considered by our approach. More expensive experiments on different types of movie films, like comedy, romantic and science fiction films are currently in progress.

Future work will focus the characterization of shots. The matching of objects inter-shots using only the features extracted form their appearances in the key-frames gives poorly results [6]. One direct improvement at this stage is to statistically model the appearances of tracked objects, in the feature space, and then to match them based on these models.

Finally, the drawback of such an approach is the set of parameters which should be chosen carefully for each movie film. For the moment, there is no completely satisfactory

method for determining the number of data clusters for the hierarchical classification technique [11] and this point is still a research problem in clustering analysis.

References

[1] J. F. Allen. Maintainning knowledge about temporal intervals. *CACM*, 26:832–843, 1983.

[2] P. Bouthemy and F. Ganansia. Video partitionning and camera motion characterisation for content-based video indexing. *Proc. 3rd IEEE Int. Conf. Image Processing.*, september 1996.

[3] 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.

[4] 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*, May 1998.

[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. Gaussian mixture densities for indexing of localized objects in a video sequence. Technical report, INRIA, March 2000. http://www.inria.fr/RRRT/RR-3905.html.

[7] R. Hammoud and R. Mohr. Interactive tools for constructing and browsing structures for movie films. In *ACM Multimedia*, pages 497–498, Los Angeles, California, USA, October 30 - November 3 2000. (demo session).

[8] R. Hammoud and R. Mohr. Probabilistic hierarchical framework for clustering of tracked objects in video streams. In *Irish Machine Vision and Image Processing Conference*, pages 133–140, The Queen's University of Belfast, Northern Ireland, 31 August - 2 September 2000.

[9] J. Huang. *Color Spatial Indexing and Applications.* PhD thesis, Cornell University, August 1998.

[10] M. Irani and P. Anandan. A unified approach to moving object detection in 2d and 3d scenes. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 20(6):577–589, June 1998.

[11] A.K. Jain and R.C. Dubes. *Algorithms for Clustering Data.* Prentice Hall, 1988.

[12] J.Aumont and M. Marie. L'analyse des films. *Fac. Cinéma, NATHAN Université*, 1988.

[13] F. Nack and A.T. Lindsay. Everything you wanted to know about mpeg-7. *IEEE Multimedia*, pages 65–77, July-September 1999.

[14] Y. Rui, S. Huang, and S. Mehrota. Exploring video structures beyond the shots. In *Proc. of IEEE conf. Multimedia Computing and Systems*, Austin, Texas USA, June 28-July 1 1998.

[15] Y. Rui, T. S. Huang, and S. Mehrotra. Constructing table-of-content for videos. *ACM Multimedia Systems Journal, Special issue Multimedia Systems on video librairies*, 1999. To appear.

[16] M. Yeung and B. Yeo. Segmentation of video by clustering and graph analysis. *Computer Vision and Image Understanding*, 71(1):94–109, July 1998.

[17] H. J. Zhang, C. Y. Low, S.W. Smoliar, and J.H. Wu. Video parsing, retrieval and browsing: An integrated and content-based solution. *ACM Multimedia*, pages 15–24, 1995.