CLUSTERING OF SCENE REPEATS FOR ESSENTIAL RUSHES PREVIEW

E. Rossi [1], *S. Benini* [1], *R. Leonardi* [1], *B. Mansencal* [2], *J. Benois-Pineau* [2]

[1] DEA-SCL, Università di Brescia, Via Branze 38, 25123 Brescia, Italy
[2] LaBRI, Université Bordeaux 1/Bordeaux 2/CNRS/ENSEIRB, 33405 Talence Cedex, France

ABSTRACT

This paper focuses on a specific type of unedited video content, called *rushes*, which are used for movie editing and usually present a high-level of redundancy. Our goal is to automatically extract a summarized preview, where redundant material is diminished without discarding any important event. To achieve this, rushes content has been first analysed and modeled. Then different clustering techniques on shot key-frames are presented and compared in order to choose the best representative segments to enter the preview. Experiments performed on TRECVID data are evaluated by computing the mutual information between the obtained results and a manually annotated ground-truth.

Index Terms— rushes, summarization, TRECVID

1. INTRODUCTION

The amount of multimedia content in digital form is ever growing, thanks to the always faster progress of technology and the decreasing prices of multimedia devices. In this scenario, the need to access video information for effective retrieval or browsing affect both home and professional environments.

This work focuses on a specific type of professional video, called *rushes*, that are raw audio-visual footage edited to build the final version of a feature movie. Our aim is to find an efficient way to present a preview of rushes in the form of a video summary. Some parts of the proposed method have been already exploited in the COST292 approach [1] submitted to the TRECVID 2008 campaign on rushes summarization [2].

According to TRECVID, the final summary should contain only the relevant parts, where undesiderable content has been removed and only one take of each scene is shown. Summary length should not exceed 2% of the duration of the original video and it should contain only relevant objects, events and camera events (*i.e.,* pan, zoom, etc.).

The evaluation of the summary takes into account various subjective and objective parameters such as the fraction of important segments included in the final summary, the easiness to find and understand the desired content, the redundancy of the summary and the system effort spent to produce it. The challenge is the construction of a system which gets the best results for every criterion considered for the evaluation.

Several techniques have been proposed to deal with rushes summarization ([3], [4], [5], and [1]). Some approaches compute the informativeness of each segment and accelerate the playback if the information is low [5]. Other approaches extract "ad-hoc" features to weight the importance of each shot and decide whether to include it in the final summary [6].

This work aims to the maximization of included events and the minimization of the redundancy factor. First we investigate an efficient approach for the key-frame selection. A tailored clustering approach based on visual features and an heuristic method to select most representative segments are then presented. Experiments on TRECVID material compare the clustering results against a manually annotated ground-truth, with a procedure derived from information theory.

This paper is organized as follows: Section 2 describes the available data set and the model of the structure built for the analysis. Section 3 presents the proposed method, while in Section 4 the evaluation procedure is introduced and results are discussed. Finally, in Section 5 conclusions are drawn.

2. RUSHES DATA DESCRIPTION

Different type of rushes footage are used by broadcasters to build documentaries or news programs, or by production companies to edit movies. This paper focuses only on *film rushes*, that is footage that is shot to produce a movie. The main characteristic of this content is the high level of redundancy. As a matter of fact, rushes present many takes of the same scene, due to actor errors for example, where a *scene* is as a set of contiguous frames depicting a part of an action in a single location and in a brief period of time.

A scene is considered a *repeat* of another one if the action depicted and the point of view are the same, even if there can be some slight differences of duration or in the lines spoken by the actors. Repetitions of a same scene are contiguous in time. As shown in Figure 1 a collection of repeated scenes is called *set*. At physical level, repeats of the same scene can be contained into subsequent shots or one shot only.

Rushes material may also contain some content not really related to the storytelling, such as scene preparation by assistants, clap boards, talks between actors and director, scenes with fixed camera, and undesirable content such as colour bars, frames whose colour is uniform or blurred (*junk frames*).

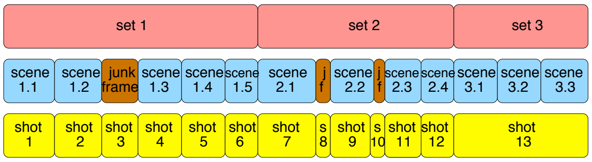

Fig. 1: Rushes videos can be divided into *sets* containing different *repeats* of the same *scene*.

3. PROPOSED FRAMEWORK

The framework proposed to create an essential preview of rushes is shown in Figure 2. The shot detection and the removal of junk frames have been automatically obtained as described in [1]. This work focuses on the key-frame extraction, the clustering algorithm and the selection of the most representative shots (*i.e.,* the blocks highlighted in Figure 2).

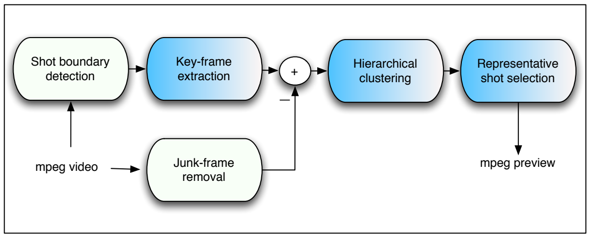

Fig. 2: Framework for generating essential rushes preview.

3.1. Key-frame selection

Once the video has been decomposed into separate shots, a representative frame (*i.e.,* a key-frame) per shot is selected. In order to effectively position key-frames, we first analysed the nature of video scenes. In general there is no rule on the scene length: they can be either very long, or very short. Moreover, at the beginning of a scene there can be a random setup time in which the assistants arrange the scene. However repeats of the same scene usually begin in the same way, at least for what concerns chromatic composition and lighting.

Therefore our first choice is to extract key-frames in fixed positions at the beginning of the shot, in detail at the 5^{th}, 30^{th}, 55^{th}, 80^{th}, 100^{th} and 200^{th} frame (for shots shorter than 200 frames, the last frame is chosen). We also compare with key-frames extracted at different percentage of the shot length (10%, 30%, 50%, 70%, 90% and 99%).

3.2. Low-level representation

Two MPEG-7 descriptors, *Colour Layout* and *Edge Histogram* [7], are then computed on key-frames. These descriptors, though computationally inexpensive, are able to capture the scene chromatic composition and background texture. As described in [1] other features can be extracted from shots (such as the presence of human faces or motion patterns) in order to identify the most important parts to keep in the preview. However, for the minimization of the redundancy factor, the two proposed features are effective enough to detect repeats.

3.3. Hierarchical clustering

In order to group similar shots into sets, a *hierarchical agglomerative clustering* has been tailored to the case of rushes data. Similarity between shots S_i and S_j is given by the euclidean distance $||\mathbf{x}_i - \mathbf{x}_j||$ between the feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ extracted from the key-frames of shot S_i and S_j.

Regarding distances between clusters, *single*, *complete* and *average-link* criterions (*i.e.*, *minimum*, *maximum*, and *average distance* between items, respectively) have been compared. Two elements are novel here: a time-constraint on shots is introduced to deal with the rushes structure, and a criterion to stop the clustering is proposed.

3.3.1. Time-constrained analysis

Usually hierarchical algorithms do not put a constraint on which items can be clustered. To deal with the rushes structure a time-constraint has been introduced: since repeats of the same scene are contiguous, only adjacent shots are allowed to merge. Considering a sequence of N shots: S_1, S_2, $\ldots S_N$, two clusters $\mathcal{X}_k$ and $\mathcal{X}_l$ can be merged only if they contain contiguous shots, that is:

$$\exists i \in [1, N-1] : S_i \in \mathcal{X}_k \text{ and } S_{i+1} \in \mathcal{X}_l \, . \qquad (1)$$

With this constraint, we restrict errors to two cases: two subsequent shots can be wrongly added to the same cluster, or two subsequent shots of the same set are splitted in two clusters. Low-level feature characterisation tries to reduce the first type of error. The second one is acceptable considering that one of our goals was the maximization of included events.

3.3.2. Stop criterion

Another typical problem of unsupervised clustering is determining how many clusters are present. Traditional approaches impose an *a-priori* criterion to stop the clustering, for example by setting in advance the final number of clusters or their maximum dimension. Anyway, these global criteria often fail in preserving the visual coherence of clusters. Here we propose a criterion based on the *inter-* and *intra-cluster* distances.

Inter-cluster distance is defined as the distance between centroids and it indicates if clusters are detached or not:

$$D_{inter}(\mathcal{C}) = \sum_{j=1}^{K} \sum_{i=1, i \neq j}^{K} |\mathcal{X}_j||\mathcal{X}_i| d(\mathbf{c}_i, \mathbf{c}_j) \qquad (2)$$

where $\mathcal{C}$ is the resulting set of K cluster (also called *cluster configuration*), $\mathcal{X}_i$ and $\mathcal{X}_j$ are two clusters belonging to configuration $\mathcal{C}$, $|\mathcal{X}_i|$ is the cardinality of cluster $\mathcal{X}_i$, that is the

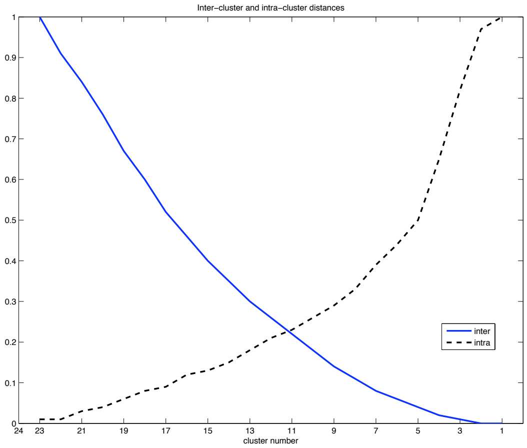

Fig. 3: Normalized inter and intra-cluster distances. The intersection point determines the final number of clusters.

number of objects in the cluster and $d(\mathbf{c}_i, \mathbf{c}_j)$ is the Euclidean distance between the two centroids of $\mathcal{X}_i$ and $\mathcal{X}_j$, respectively.

On the other hand, intra-cluster distance is the distance between objects of the same cluster and it is useful to understand whether a cluster is compact or not. It is defined as

$$D_{intra}(\mathcal{C}) = \sum_{i=1}^{K} \sum_{j=1}^{|\mathcal{X}_i|} d(\mathbf{x}_j^{(i)}, \mathbf{c}_i) \tag{3}$$

where $\mathbf{x}_j^{(i)}$ is the j^{th} key-frame belonging to cluster $\mathcal{X}_i$.

In order to stop the clustering process, the two normalized curves of intra and inter-distance are first computed. Then the final number of clusters is found at the intersection of the two curves, and it represents the optimal compromise between intra-cluster compactness and inter-cluster separation. An example of normalized intra-cluster and inter-cluster distance trends is shown in Figure 3 for a TRECVID test video.

3.4. Selection of representative shot

Since we expect that most repeats collapse in the same cluster, only one shot per cluster is selected. By analysing rushes content it has been noticed that the last repeat of a scene is usually the one which contains less errors and which is the most likely to be selected by the director for the final editing. Therefore our choice is to select, as a representative, the last shot (in temporal order) belonging to the cluster.

4. EXPERIMENTAL RESULTS

The proposed method was tested on the manually annotated TRECVID data set presented in Table 1, and experiments were conducted by varying all the possible parameters. Keyframes were extracted first at fixed positions and then at different percentage of the shot length. Distances between keyframes were computed first relying only on colour layout (**l**) and then by using both colour layout and edge histogram (**e**). Hierarchical clustering was run by using all link criterions, that are single (**s**), complete (**c**) and average (**a**). Results obtained by adopting the time-constraint (**t**) on shots were compared to those obtained without the constraint (**n**).

Video	Duration	Scenes	Shots
MRS025913.mpg	25:42	8	38
MRS035132.mpg	05:27	2	6
MRS042543.mpg	32:16	8	28
MRS042548.mpg	26:28	7	53
MRS043400.mpg	13:02	4	19
MRS144760.mpg	27:10	7	46
MRS150148.mpg	27:43	11	59
MRS157469.mpg	35:35	36	82
MRS336905.mpg	03:32	9	10
MS210470.mpg	15:49	13	83

Table 1: TRECVID rushes test set.

4.1. Evaluation

Recall and precision parameters can be used for evaluation only when the number of the ground-truth clusters and of the obtained ones is the same. Unfortunately this is not the case.

To deal with this problem the *Normalized Mutual Information* (NMI) [8] can be used. It derives from Information Theory [9] and is a symmetric measure to quantify the statistical information shared between two distributions. The idea is that the optimal clustering should share the most information with the ground-truth clustering [10].

Let then X and Y be random variables representing the ground-truth clusters and the automatically obtained ones, respectively. In case of Maximum Likelihood Estimation, mutual information between X and Y can be expressed as:

$$I(X,Y) = \sum_i \sum_j \frac{|x_i \cap y_j|}{N} \log \left(\frac{N|x_i \cap y_j|}{|x_i||y_j|} \right) \tag{4}$$

where $|x_i|$ is the cardinality of the ground-truth cluster x_i, $|x_i \cap y_j|$ is the number of items of obtained cluster y_j shared with ground-truth cluster x_i and N is the shot total number.

Since this value is not bounded by the same constant for all data sets, it has been normalized in the range $[0, 1]$:

$$NMI(X,Y) = 2 \cdot \frac{I(X,Y)}{H(X) + H(Y)} \tag{5}$$

with $H(X)$ and $H(Y)$ the entropy of X and Y:

$$H(X) = -\sum_i \frac{|x_i|}{N} \log \frac{|x_i|}{N} \; . \tag{6}$$

NMI is maximum in case of one-to-one mapping between ground-truth clusters and the obtained clustering results.

4.2. Results

The values of Normalized Mutual Information averaged on the ten analyzed videos are shown in Figure 4 and 5 for different choices of key-frame position.

The best average score is obtained positioning the key-frame at the **30%** of the shot length, with a "**aet**" configuration, that is with an average link clustering (**a**), using both colour layout and edge histogram (**e**) and by introducing the time-constrain on shots (**t**). The second best average score is obtained with the key-frame positioned at **30%** of the shot length, with a "**set**" configuration (*i.e.,* single-link, colour layout and edge histogram, and time-constrained clustering).

The good performances obtained by the 30% "aet" configuration were also confirmed by the visual inspection of the resulting clusters. Repeated scenes were mostly grouped in the same cluster, even if sometimes, as expected, it happens that a set is split between two clusters. Regarding the choice of the last shot as a cluster representative, as a matter of fact only in the 2.86% of cases the last repeat did not contain all the important events listed in the TRECVID ground-truth.

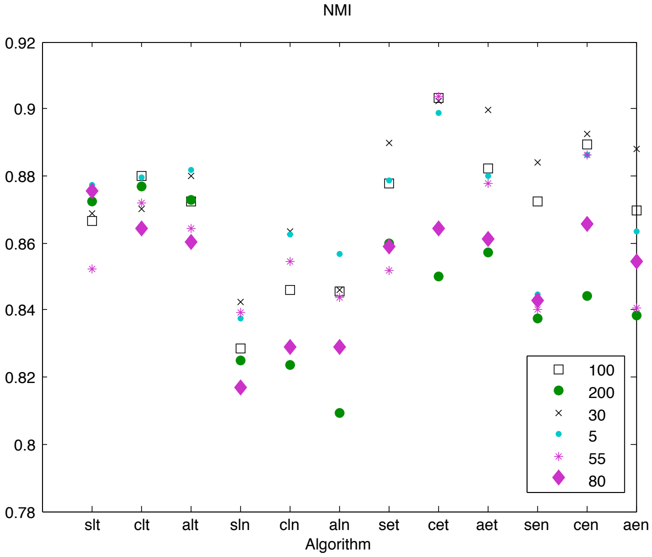

Fig. 4: NMI values with fixed positioning of key-frames.

5. CONCLUSIONS

This work analysed in detail script-content rushes. A model of the structure has been proposed together with a method to summarize this type of content. Various aspects such as the analysis of clustering methods, the study of MPEG-7 descriptors, the selection of appropriate key-frames have been taken into account and compared. Future work includes the improvement of the clustering method by using multimodal features and the comparison with other clustering methods.

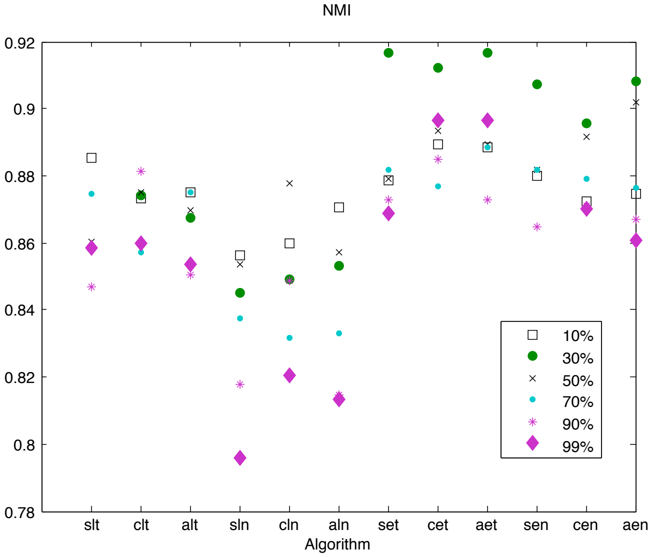

Fig. 5: NMI values with key-frames positioned at different percentages of the shot length.

6. REFERENCES

[1] S. U. Naci et al., "The COST292 Experimental Framework for RUSHES Task in TRECVID2008," in *Proc. of TVS '08, ACM Multimedia*, Vancouver, Canada, 27-31 Oct. 2008.

[2] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *MIR '06: Proc. of the 8th ACM Inter. Workshop on Multimedia Information Retrieval*, New York, NY, USA, 2006, pp. 321–330, ACM Press.

[3] Rushes FP6-045189, "http://www.eitb.com," .

[4] F. Wang and C.-W. Ngo, "Rushes video summarization by object and event understanding," in *TVS '07: Proc. of the inter. workshop on TRECVID video summarization*, NY, USA, 2007, pp. 25–29, ACM.

[5] M. Detyniecki and C. Marsala, "Video rushes summarization by adaptive acceleration and stacking of shots," in *TVS '07: Proc. of the inter. workshop on TRECVID video summarization*, NY, USA, 2007, pp. 65–69, ACM.

[6] J. Kleban et al., "Feature fusion and redundancy pruning for rush video summarization," in *TVS '07: Proc. of the inter. workshop on TRECVID video summarization*, NY, USA, 2007, pp. 84–88, ACM.

[7] P. Salembier and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface*, John Wiley & Sons, Inc., NY, USA, 2002.

[8] S. Basu, M. Bilenko, and R. J. Mooney, "A probabilistic framework for semi-supervised clustering," in *KDD '04: Proc. of the tenth ACM SIGKDD inter. conf. on Knowledge discovery and data mining*, NY, USA, 2004, pp. 59–68, ACM.

[9] T. M. Cover and J. A. Thomas, *Elements of information theory*, Wiley-Interscience, New York, NY, USA, 1991.

[10] A. Strehl and J. Ghosh, "Cluster ensembles: a knowledge reuse framework for combining partitionings," in *Eighteenth national conf. on Artificial intelligence*, Menlo Park, USA, 2002, pp. 93–98, American Association for Artificial Intelligence.