Multimedia Summarization in Law Courts: An Environment for Browsing and Consulting

E. Fersini[1], G. Arosio[1], E. Messina[1], F. Archetti[1,2], D. Toscani[2]

[1] DISCo, Università degli Studi di Milano-Bicocca,
Viale Sarca, 336 - 20126 Milano, Italy
{archetti, arosio, fersini, messina}@disco.unimib.it
[2] Consorzio Milano Ricerche,
Via Cicognara 7 - 20129 Milano, Italy
{archetti, toscani}@milanoricerche.it

Abstract. Digital videos represent a fundamental informative source of those events that occur during a penal proceedings, which thanks to the technologies available nowadays, can be stored, organized and retrieved in short time and with low cost. Considering the dimension that a video source can assume with respect to a courtroom recording, various necessities have been highlighted by the main judicial actors: fast navigation of the stream, efficient access to data inside and effective representation of relevant contents. One of the possible solutions to these requirements is represented by multimedia summarization aimed at deriving a synthetic representation of audio/video contents, characterized by a limited loss of meaningful information. In this paper a multimedia summarization environment is proposed for defining a storyboard for proceedings celebrated into courtrooms.

1 Introduction

Multimedia summarization techniques analyze several informative sources comprises into a multimedia document, with the aim at extracting a semantic abstract. Multimedia summarization techniques available in literature can be divided in three main categories: (1) internal techniques, which exploit low level features of audio, video and text; (2) external techniques, which refer to the information typically associated with a viewing activity and interaction with the user; (3) hybrid techniques, which combine internal and external information. These techniques are focused on different types of features: (a) domain specific, i.e. typical characteristics of a given domain known a priori and (b) non-domain specific, i.e. non-generic features associated with a particular context.

With respect to internal techniques the main goal is to analyze low-level features derived from text, images and audio content within a multimedia document. In [2] the performance related to the identification of special events are increased by combining scene recognition techniques with OCR-based approaches for subtitles recognition in baseball video documents. In [3] entire scenes containing text are recognized using OCR techniques for a subsequent identification

of key events through audio and video features in football matches. In [1] the semantics of objects and events occurring within news video are extracted from subtitles and used to specialize / improve the systems of automatic speech recognition.

In order to reduce the semantic gap between low level features and semantic concept for producing a meaningful summary, research is moving towards the inclusion of external information that usually include knowledge of the context in which evolves a multimedia document and user-based information. The techniques able to generate a video summary on the basis of external information are limited to three case studies [4] [5] [6] focused on using domain specific features. In [4] a summarization technique is proposed in order to gather context information from the acquisition/registration phase and collected by monitoring the movement of citizens around their houses. Cameras at a specific position and pressure sensors are used to track users. Since users are not required to provide any kind of information, the summary produced by analyzing data concerning the movement (such as the distance between steps and direction changes). In [5] and [6] semantic annotations collected during the production phase of the video and described by the standard MPEG-7 are analyzed. In particular in [6] a sequence of audio-video segments is produced on the basis of annotations from video sports (baseball matches), such as players' names or specific events occurring during the match. In [5] a video, characterized by a set of MPEG-7 macro-semantic annotations collected during the acquisition phase, is further annotated by users in order to indicate their level of interest in each video segment. The associations between preferences and the macro-annotations are then modelled by using supervised learning approaches to enable the generation of automatic summary of new multimedia documents.

An attempt that tries to combine the peculiarities of the previous techniques is represented by Hybrid Techniques. Hybrid summarisation techniques combine the advantages provided by internal and external approaches by analyzing a combination of internal and external information. As overviewed for the previous techniques, the hybrid ones can be distinguished in domain specific and non-domain specific. Examples of domain-specific hybrid techniques are related to music videos [7], broadcast news [8] and movies [9]. In non-domain specific approaches we can find:
- in [10] the summarization approach could be described by two stages: (1) frames are grouped by a clustering approach, using colour image features; (2) in editing phase, manual annotations of representative frame for each cluster, with subsequent spread to frames of the same cluster, are required. Summary for a new video is then generated by using representative elements of each cluster that generates a matching with user query.
- in [11] an annotation tool is used during the editing phase in order to propagate semantic descriptors to non-labelled contents. During the summary generation phase, the user profile is considered in order to create a customized synthetic representation.

By analyzing the state of the art related to multimedia summarization, no evidences about summary over courtroom proceedings are given. In this paper we are mainly addressing the problem of deriving a storyboard of a multimedia document coming from penal proceedings recordings, by proposing an external summarization technique based on the unsupervised clustering algorithm named Induced Bisecting K-Means. The main outline of this paper is the following. In section 2 the proposed multimedia summarization environment is presented. In section 2.1 the workflow for deriving a storyboard for the judicial actors is described. In section 3 details about the exploited clustering algorithm are given. Finally, in section 4 conclusions are derived.

2 Multimedia Summarization Environment

In order to address the problem of defining a short and meaningful representation of a debate that is celebrated within a law courtroom, we propose a multimedia summarization environment based on unsupervised learning. The main goal of this environment is to create a storyboard of either a hearing or an entire proceedings, by taking into account the semantic information embedded into a courtroom recording.

In particular, the main information sources used for producing a multimedia summary are represented by:

- automatic speech transcriptions that correspond to what is uttered by the actors involved into hearings/proceedings
- automatic audio annotations coming from emotional states recognition (for example fear, neutral, anger)
- automatic video annotations that correspond to what happen during a debate (for instance change of witness posture, new witness, behavior of a given actor)

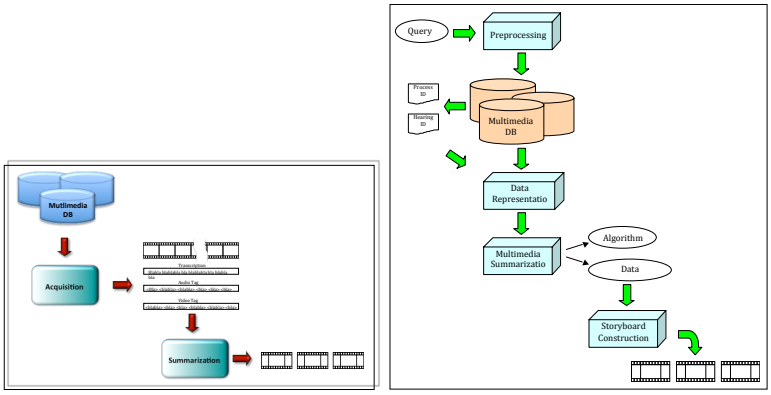
The Multimedia Summarization Environment includes two different modules: the acquisition module and the summarization module.

- The *acquisition module*, given a query specified by the end user, retrieves multimedia information from the Multimedia Database in terms of audio-video track(s), speech transcription and semantic annotations.
- The *summarization module* is aimed at producing on-demand storyboard by exploiting the information retrieved by the acquisition module. The summary is created by focusing on maximally query-relevant passages and reducing cross-document redundancy.

A simple overview of the modules involved into the multimedia summarization environment is depicted in figure 1 (a).

2.1 Multimedia Summarization Workflow

In order to summarize a multimedia document according to the user needs, a query statement is defined for acquiring requirements in terms of query and to start the entire workflow (see figure 1 (b)).



(a) Overview of the multimedia summarization macro-modules

(b) Overview of the multimedia summarization workflow

Fig. 1. Multimedia Summarizaion Environment

The user query is specified at the graphic interface level, where a list of trials are available, in terms of keywords in which we are interested (whatever is uttered by the involved speaker, the emotional state of actors, etc...).
Once the query has been specified, it is submitted to the pre-processing module. The aim of this module is to optimize the user query by eliminating noise and by reducing the size of vocabulary, i.e. stop words removal and stemming are performed to enhance retrieval performance on transcription and annotations.

After the preprocessing activity, the query will be ready to be submitted to the retrieval module, which is aimed at accessing to the multimedia database in order to retrieve all the information that match the user query in terms of transcription of the debate, audio and videos annotations (audio and video tag). At this level, two possibilities are given to the end user: to retrieve and summarize an entire trial that matches the query or to summarize only those sub-parts of the proceedings that match the query. In the first case the user query is used to retrieve the multimedia documents related to a trial by executing a high-level skimming of the overall database. After this initial step all the clips of the retrieved hearing are considered for producing the summary. In the second case the query is used to scan the database in a more exhaustive way so that, within a given trial, only the audio, video and textual clips that completely match the user query are retrieved.

In both cases we refer to a (audio, video and textual) clip as a consecutive portion of a debate in which there is one speaker whom is active, i.e. there exist a sequence of words uttered by the same speaker without breaking due to other speakers. In this way we have one clip of audio/video tag and transcription for each speaker period.

The next step in the multimedia summarization workflow relates to data representation module. The aim of this module is to combine information coming from different sources in order to create a unified representation. This activity

is performed through a feature vector representation, where all the information able to characterize the audio, video and textual clip of interest are managed as features and weights. Examples of features exploited by this representation are given by the textual transcription, the audio and video tag, the start and end time of the relevant sub-parts of the debate.

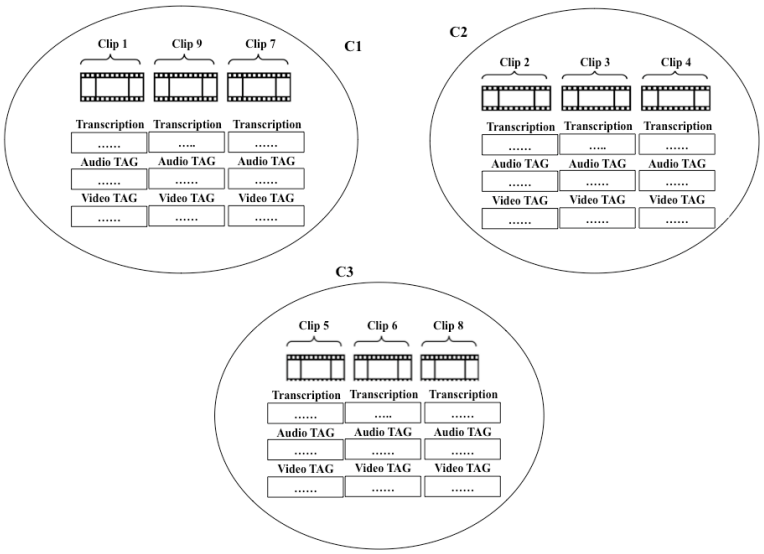


Fig. 2. Clustering output

Given the data structure that has been created, the multimedia summarization module may start the summary generation. The core component is based on a clustering algorithm named Induced Bisecting K-means [13]. The algorithm creates a hierarchical organization of (audio, video and textual) clips, by grouping in several clusters hearings or sub-parts of them according to a given similarity metric. This algorithm is able to build a dichotomic tree in which coherent concepts are grouped together, i.e. each cluster created by the algorithm contain a set of audio, video and textual clips containing similar concepts that are also coherent with the query submitted by user.
The last step related to the storyboard construction, where the final storyboard is derived from the dichotomic tree structure coming from the previous multimedia summarization activity. Given the dichotomic tree, a pruning step is performed in order to choose only those clusters that respect a given intra-cluster similarity threshold. Suppose that the pruning activity after the Induced Bisecting K-means returns a set of clusters as reported in figure 2 where C1, C2 and C3 are the resulting clusters and the clips named $1, \ldots, 9$ represent the sub-parts of the debate. The storyboard construction activity considers the representative elements of each cluster (centroids) as the relevant audio, video and textual clips

for the summary. The storyboard is generated starting from the centroids by presenting to the end user the first frame, together with the references of the given audio, video and textual clip, references of the trial/hearing, start and end time of the segments and so on. By referencing figure 2, only the first frames related to segments 2, 5 and 7 (representative of the obtained 3 clusters) are presented to the end user as pictures that could be clicked to start the corresponding audio-video portion.
In the following subsection details about the core component of the multimedia summarization environment, i.e. the Induced Bisecting K-Means clustering algorithm, are given.

3 The hierarchical clustering algorithm

The approaches proposed in the literature for hierarchical clustering were mostly statistical with a high computational complexity . A novel approach, Bisecting k-Means was proposed in [12], has a linear complexity and is relatively efficient and scalable. It starts with a single cluster of multimedia clips and works in the following way:

Algorithm 1 Bisecting K-Means

1: Pick a cluster S of clips m_l to split
2: Select two random seeds which are the initial representative clips (centroids)
3: Find 2 sub-clusters S1 and S2 using the basic k-Means algorithm
4: Repeat step 2 and 3 for ITER times and take the split that produces the clustering with the highest Intra Cluster Similarity (ICS)
5: $ICS(S_k) = \frac{1}{|S_k|^2} \sum_{m_i, m_j \in S_k} cosine(m_i, m_j)$
6: Repeat steps 1, 2 and 3 until the desired number of clusters is reached.

The major disadvantage of this algorithm is that it requires the a priori specification of the number of clusters K and the parameter ITER for creating several splits of the same group in order to choose the best one. An incorrect estimation of K and ITER may lead to poor clustering accuracy. Moreover, the algorithm is sensitive to the noise which may affect the computation of cluster centroids. For any given cluster let N be the number of clips belonging to that cluster and R the set of their indices. In fact, the j^{th} element of a cluster centroid used by the k-Means algorithm during step 3 is computed as $c_j = \frac{1}{N} \sum_{r \in R} m_{ij}(r)$ where N represents the number of clips belonging to the cluster and $m_{rj}(r)$ is the vectorial representation of j features of clip i belonging to cluster r. The centroid c may contain also the contribution of noisy features contained into the clip representation which the pre-processing phase have not been able to remove. To overcome these two problems we exploit an extended version of the Standard

Bisecting k-Means, named Induced Bisecting k-Means [13], whose main steps are described as follows:

Algorithm 2 Induced Bisecting K-Means

1: Set the Intra Cluster Similarity (ICS) threshold parameter τ
2: Build a distance matrix A whose elements are given by the Euclidean distance between clips $a_{ij} = \sqrt{\sum_k (m_{ik} - m_{jk})^2}$ where i and j are clips
3: Select, as centroids, the two clips i and j s.t. $a_{ij} = \max_{l,m} A_{lm}$
4: Find 2 sub-clusters S1 and S2 using the basic k-Means algorithm
5: Check the ICS of S_1 and S_2 as
6: If the ICS value of a cluster is smaller than τ, then reapply the divisive process to this set, starting form step 2
7: If the ICS value of a cluster is over a given threshold, then stop. 6. The entire process will finish when there are no sub-clusters to divide.

The main differences of this algorithm with respect to the Standard Bisecting k- Means consist in: (1) how the initial centroids are chosen: as centroids of the two child clusters we select the clips of the parent cluster having the greatest distance between them; (2) the cluster splitting rule: a cluster is split in two if its Intra Cluster Similarity is smaller than a threshold parameter τ. Therefore, the optimal number of cluster K is controlled by the parameter τ and therefore no input parameters K and ITER must be specified by the user. Our algorithm outputs a binary tree of clips, where each node represents a clip collection which elements are similar. This structure is processed according to [7], in order to obtain a flat representation of clusters.

4 Conclusion and Future Work

In this paper a multimedia summarization environment has been presented in order to allow judicial actors to browse and navigate multimedia documents related to penal hearings/proceedings. The main component of this environment is represented by the summarization module, which create a storyboard for the end user by exploiting several semantic information embedded into a courtroom recording. In particular, automatic speech transcriptions joint with automatic audio and video annotations have been used for deriving a compressed and meaningful representation of what happens into a law courtroom. Our work is now focused on creating a testing environment for a quality assessment of the storyboard produced by our environment.

Acknowledgment

This work has been supported by the European Community FP-7 under the JUMAS Project (ref.: 214306).

References

1. J. Kim, H. Chang, K. Kang, M. Kim, H. Kim, Summarization of news video and its description for content-based access, International Journal of Imaging Systems and Technology, 267-274, 2004.
2. C. Liang, J. Kuo, W. Chu, J. Wu, Semantic units detection and summarization of baseball videos, in: Proc. of the 47th Midwest Symposium on Circuits and Systems, pp. 297-300, 2004.
3. D.W. Tjondronegoro, Y. Chen, B. Pham, Classification of selfconsumable highlights for soccer video summaries, in Proc. of the IEEE International Conference on Multimedia and Expo, pp. 579-582, 2003.
4. G. de Silva, T. Yamasaki, K. Aizawa, Evaluation of video summarization for a large number of cameras in ubiquitous home, in: Proc. of the 13th Annual ACM International Conference on Multimedia, pp. 820-828, 2005.
5. A. Jaimes, T. Echigo, M. Teraguchi, F. Satoh, Learning personalized video highlights from detailed MPEG-7 metadata, in: Proc. of the IEEE International Conference on Image Processing, pp. 133-136, 2002.
6. Y. Takahashi, N. Nitta, N. Babaguchi, Video Summarization for Large Sports Video Archives, in Proc. of the IEEE International Conference on Multimedia and Expo, pp. 1170-1173, 2005.
7. L. Agnihotri, N. Dimitrova, J.R. Kender, Design and evaluation of a music video summarization system, in Proc. of the IEEE International Conference on Multimedia and Expo, pp. 1943-1946, 2004.
8. H. Yang, L. Chaisorn, Y. Zhao, S. Neo, T. Chua, VideoQA: question answering on news video, in: Proc. of the 11th Annual ACM International Conference on Multimedia, pp. 632-641, 2003.
9. T. Moriyama, M. Sakauchi, Video summarization based on the psychological unfolding of drama, Systems and Computers in Japan, pp 1122-1131, 2002.
10. Y. Rui, S.X. Zhou, T.S. Huang, Efficient access to video content in a unified framework, in: Proc. of the IEEE International Conference on Multimedia Computing and Systems, pp. 735-740, 1999.
11. B.L. Tseng, C.-Y.L.J.R. Smith, Using MPEG-7 and MPEG-21 for personalizing video, IEEE Transactions on Multimedia, pp42-52, 2004.
12. M. Steinbach, G. Karypis, V. Kumar, A comparison of Document Clustering Techniques, In KDD Workshop on Text Mining, 2000
13. F. Archetti, E. Fersini, P. Campanelli, E. Messina, A Hierarchical Document Clustering Environment Based on the Induced Bisecting k-Means, in Proc. of Flexible Query Answering Systems, 4027/2006