# Scalable Video Summarization of Cultural Video Documents in Cross-Media Space based on Data Cube Approach

Karina R. Perez-Daniel and Mariko Nakano Miyatake
SEPI, ESIME Culhuacan
National Polytechnic Institute, IPN
Mexico City, Mexico
Email: krperezd@gmail.com, mnakano@ipn.mx

Jenny Benois-Pineau, Sofian Maabout and Gabriel Sargent
LaBRI
University of Bordeaux I
Bordeaux, France
Email: benois-p,sofian.maabout,gabriel.sargen@labri.fr

***Abstract*—Video summarization has been a core problem to manage the growing amount of content in multimedia databases. An efficient video summary should display an overview of the video content and most of existing approaches fulfil this goal. However the information does not allow user to get all details of interest selectively and progressively. This paper proposes a scalable video summarization approach which provides multiple views and levels of details. Our method relies on the usage of cross media space and consensus clustering method. A video document is modelled as a data cube where the level of details is refined over nonconsensual features of the space. The method is designed for weakly structured content such as cultural documentaries and was tested on the INA corpus of cultural archives.**

**Keywords—Video summarization, Scalability, Cross-media space, Consensus clustering, Data cube.**

## I. INTRODUCTION

Video documentaries are one way to capture the cultural heritage of a country and they can be used for the preservation and dissemination of the culture. However, large volumes of such content as well as their large duration enhance the necessity of developing a fast, easy and multidimensional access to them. Video summarization is a compact representation of video content, which provides access to most relevant information based on similarities.

Due to the importance of the problem of browsing and retrieving information in large data, several video summarization approaches have been proposed [1,2] and a specific task of TRECVID campaign such as rushes summarization [3] was run. Summarization is usually done by grouping similar video segments on the basis of continuous audio channel [4]. Generally speaking video summarization approach can be inspired by data analysis techniques such as clustering or supervised learning, etc.

It is crucial to incorporate the video summarization approaches into large scale multimedia applications. However, the strong requirements of those applications in terms of scale, time response and high dimensional information make the *scalability* a very challenging problem.

The scalability can be seen as the ability of proposed approach to generalize on a large scale of data. Another interpretation comes from multi-scale data representation and means that the data can be described in a coarse-to-fine manner, this is how we understand the scalable video summarization. A *scalable* video summary allows navigating in abstracted video content in a progressive manner according to the user request.

The main contribution of this work is to provide a scalable video summary in terms of media content. This approach is inspired by *data cube* On Line Analytical Processing (OLAP) operations [5]. The idea is borrowed from hierarchical information retrieval frameworks, which have become particularly popular in Multimedia archives [6]. The data cube concept has been proposed to facilitate user's navigation through multidimensional space where each move corresponds to a query using some combination of the dimensions. In this work we consider different descriptors and embed them into consensus clustering framework which allows data cube partitioning in multimodal audio-visual description space. To test this approach a sample of National Audiovisual Institute (INA) cultural video corpus [7] was used. To evaluate the performance of this method, *precision* and *recall* measures are considered. The evaluation consists in the comparison between human detection of segment boundaries (manual annotation) and the automatic visual summary obtained by the proposed method.

The rest of the paper is organized as follows. Section II presents the overall approach. Section III shows more details in the consensus clustering implementation in our multidimensional space. In Section IV are presented the results and discussion and in Section V the conclusion and perspectives of the paper are given.

## II. VISUAL SUMMARIES IN A CROSS-MEDIA SPACE

Summarization of documentaries and other cultural programs is a challenging issue because of the absence of shot production rules. Besides, the duration, the content as well as the presentation vary a lot over each documentary. Another

978-1-4799-3990-9/14/$31.00 ©2014 IEEE

important aspect to consider is that the nature of a documentary does not imply several repeats of the same scene. However, frequently a documentary contains similar video content along the time and this characteristic has to be considered to build the video summary. For this reason in this we propose to take advantage of the multiple times that similar video content can be found in a documentary to propose a scalable video summarization approach, where each scale (level of detail) shares some attributes, with the selected *keyframe* of the video summary displayed allowing user's navigation in the video document.

To achieve the scalable video summarization, we follow the methodology illustrated in Figure 1, which mainly consists of two stages: *Video summarization* stage and *scalability*.

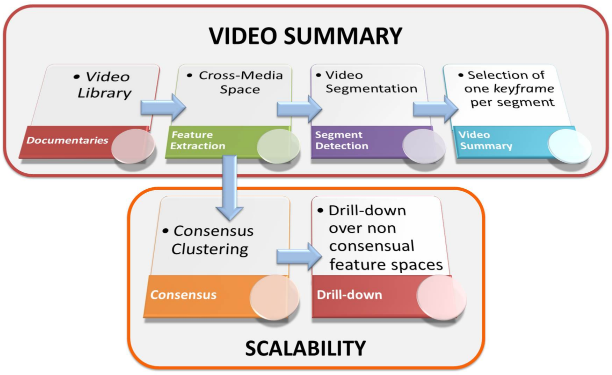

Fig. 1. Proposed architecture for Scalable Video Summarization

Video summarization of a documentary is a compact representation of a video document where it is desirable that each segment be represented by a *keyframe*. In this approach we seek for representing a documentary as a set of contiguous audio-visual segments which are not obviously video shots but can be longer or shorter than them. They are homogeneous in a cross-media space which is composed by the so-called low level visual and audio features. In order to build such partition of the video into segments we use the consensus clustering approach which allows using different dimensions of the description space in the clustering paradigm, see the *consensus* block in the diagram of Figure 1.

Our goal is to ensure a scalable navigation in a video summary. Hence, to model a video document in a cross-media description space we use the OLAP data cube model [5]. It allows navigating into the clusters obtained by consensus clustering according to the preferences of the user. We materialize this by *drill-down* block in Figure 1.

In the following of this section we will present all blocks of our method such as feature extraction, segment detection and video summary construction.

## *A. Cross-Media Space*

The data cube approach allows different combinations of features in the famous early *fusion paradigm* in multimedia. Hence, the synchronization of different visual and audio features is crucial when describing video content.

*1) Visual Features:* To describe visual information we limit ourselves to global frame descriptors. For the color, we use the well-known MPEG7 features such as Color Structure Descriptor (CSD) [8] and Scalable Color Descriptor (SCD) [9]. CSD has presented effectiveness in image retrieval based on color [9] and in shot boundary detection [10], while SCD is more sensible to color variations. CSD [8] captures the distribution of colors in the image as well as the local spatial structure of colors. Here we have considered a 64 quantization levels in HMMD color space to get a 64-dimensional vector. The second color descriptor SCD, is a histogram computed in HSV color space and then encoded by Haar transform coefficients. In this work we use 128 bins.

To describe the shape and texture in video frames we use the Pyramid of Histogram of Oriented Gradients (PHOG) [10] which has been proven to be efficient in recognition of global shape of objects and human actions. It represents a pyramid of blocks with the histogram of oriented gradients; we use two levels of it [11]. Hence the dimension of this feature is 100. All visual features are computed at the frame rate of 1fps.

*2) Audio Features:* We consider descriptions of the audio stream through Mel-Frequency Cepstral Coefficients (MFCC) and Chroma vectors. A MFCCs is a set of coefficients which encodes the rough shape of the spectral envelope of the signal [12]. A Chroma vector is a set of coefficients which quantizes the energy associated to the twelve semi-tones of the spectrum in western music theory.

In this work, 13 MFCCs (including 0th order) and Chroma vectors of dimension 12 are regularly extracted from the audio using Yaafe [13], with hop size of 2048 and 8192 points respectively, and analysis window sizes of 4096 and 32768 points respectively (other MFCC and Chroma vectors extraction parameters are set as the default ones in Yaafe). The set of values $x^d$ of each feature in time is then normalized by linear mapping.

Audio features are computed at higher frequency than video frame rate, therefore in this paper we use the sampling rate of MFCC descriptor to generate the audio-visual description of the video by the linear interpolation of both audio and visual features.

## *B. Video Summary construction*

Video summarization we propose consists in grouping video frames into video segments and then, selecting a representative *keyframe* per segment. Then, partition of each segment can be fulfilled on user request thus ensuring the scalability property. This approach can be formalized with the data cube OLAP model.

In general terms, data cube structure [5] consists of several *dimensions*, where each dimension represents some attribute in the database, which is represented by a *measure*. In our case *dimensions* are given by the proposed descriptor spaces, while *measures* are given by the clustering of them.

Data cube offers flexibility for navigation into the data by displaying a summary at different granularity level. To reach this goal, data cube considers several OLAP operations such as *drill-down*. This operation implies the data summarization by climbing down hierarchically into the data. Hence in a data cube model, we need to define the clustering in a complete description space and then define it accordingly to particular dimensions. We formalize the high level clustering in a complete space as a *segment detection* and present it below.

A *segment* is a collection of data points close in the description space and contiguous in time. We obtain them by K-means clustering [14] in description space first and by temporal alignment secondly.

To set the number of clusters *K* required by K-Means, we propose to compute the *target number of segments*, K, in terms of the *target summary duration* $d_{summary}$, which is given by (1):

$$d_{summary} = \frac{d_{video} * \rho}{100} \qquad (1)$$

where $d_{video}$ is the duration of the video document in seconds and $\rho$ is the expected percentage of the summary w.r.t. the duration of the overall video document. Thus, the *target number of segments* is calculated as follows (2):

$$K = \frac{d_{summary}}{\widehat{N}} \qquad (2)$$

where $\widehat{N}$ is the average number of frames per segment in the current category of video obtained by manual annotation of the data base.

K-means offers the data partitioning according to feature similarity and often similar frames can be found in different segments along the video, which indeed can be chronologically distanced. To solve this problem we use *density based clustering* as a post-processing stage. Thus, the density connectivity between the members of a set $S = \{s_1, s_2, \dots, s_n\}$ is given by the time stamp of each member.

Let $\{S_1, S_2, \dots, S_K\}$ be a set of clusters obtained by K-means in the complete description space, where $S_k = \{s_{1k}, s_{2k}, \dots, s_{nk}\}$ $\forall k \in \{1, 2, \dots, K\}$ and considering that $s_{jk}$ exists at the time instant $t_{s_{jk}}$ $\forall j \in \{1, 2, \dots, n_k\}$, such that $\overline{S_k} = \{t_{s_1}, t_{s_2}, \dots, t_{s_n}\}$. Then, the member $t_{s_{(j+1)}} \notin \overline{S_k}$ iff $t_{s_{(j+1)k}} > (t_{s_{jk}} + th)$. Thus, in that case, a new subset $\hat{S}$ emerges, where $th$ is a temporal distance threshold (set to the highest sampling rate), otherwise, the member $t_{s_{(j+1)}} \in \overline{S_k}$. Therefore, considering the temporal distance, now the set of partitions is given by $\hat{S} = \{\widehat{S_1}, \widehat{S_2}, \dots, \widehat{S_Q}\}$.

Finally the set of partitions $\{\widehat{S_1}, \widehat{S_2}, \dots, \widehat{S_Q}\}$ is chronologically sorted to represent the chronological occurrence of each segment as $\widehat{S_p}$.

## III. Scalable Video Summary Navigation Based on Data Cube and Consensus Clustering

Scalable video summary in terms of content description refers to the multidimensional access to different feature spaces. Scalability makes it possible to navigate in the video summary to get detailed information about the selected segment, over the selected feature space.

A consensus clustering is a process of merging agreements over several clustering on the same data set but with different parameters[15]. In our case of multimodal description space, we propose to build several partitions of the same data set, which is one video document. Hence the process presented in section 2 is repeated on different subspaces of a complete description space yielding a pre-computed set of partitions of the same data set. We propose the scalable navigation in the data set using less consensual partitions of clusters. We introduce this formally in the following.

Let us define a video $V$ as a finite set of cardinality $|V| = m$, which has been described by $\eta$ feature spaces. From the clustering of each feature space we can get a set of $\eta$ set partitions $\varphi$ such that $v = \{\varphi_1, \varphi_2, \dots, \varphi_\eta\}$, where $\varphi_i$ is a set partition of the *i-th* feature space.

*Agreements* and *disagreements* between two partitions $\varphi_i$ and $\varphi_j$ of $v$, can be calculated by considering all pairs of elements in $\varphi_i$ and $\varphi_j$ , where an *agreement* is given by $a_{ij_1} = \{pairs\ that\ are\ co-clustered\ in\ \varphi_i\ and\ \varphi_j\}$ and $a_{ij_2} = \{pairs\ that\ are\ not\ co-clustered\ in\ \varphi_i\ neither\ in\ \varphi_j\}$ , while a *disagreement* is given by $b_{ij_1} = \{pairs\ co-clustered\ in\ \varphi_i\ but\ not\ in\ \varphi_j\}$ and $b_{ij_1} = \{pairs\ co-clustered\ in\ \varphi_j\ but\ not\ in\ \varphi_i\}$.

Symmetric distance difference (sdd) can be used to measure the distance $d(,)$ between two partitions according to (3)

$$d(\varphi_i, \varphi_j) = b_{ij_1} + b_{ij_2} = \binom{m}{2} - (a_{ij_1} + a_{ij_2}) \qquad (3)$$

where $\binom{m}{2}$ are the pairs of members of clusters. To compare the distance between partitions in the set of partitions $\{\varphi_1, \varphi_2, \dots, \varphi_\eta\}$, let us consider the sum of distances as (4):

$$SD = \sum_{i=1}^{\eta} d\left(\varphi_i, \varphi_j\right) \qquad (4)$$

where η is the number of set partitions, $\varphi_i$ is the current set partition and $\varphi_j$ is a specified partition, where $\varphi_j \in$ v. Thus the most consensual set partition is given by the one with the minimum value of SD, while the least consensual one is defined by the set partition with the maximum value of SD.

Feature spaces with lower consensus are indeed, different views of the video summary but still they have some similar visual content. Hence, they can be used to *drill-down* on the data cube. It means a selection of a pre-computed partition of a cluster of interest in the less consensual subspace. An example is shown in Figure 2. Here, the three less consensual subspaces are considered. As illustrated in Figure 2, the *dimensions* of the data cube are *A*, *B* and *C*, where *A, B, C* are the three feature spaces with less consensus and the *measures* are clusters over dimensions, where the number of clusters is given by (2). The *base cuboid* is yielded by clustering over *ABC* and the *cuboids* of the first level are the clusters between *AB*, *AC* and *BC*, while the *cuboids* of the second level are the clusters in each single feature.

In a scalable video summary paradigm the user selects a *keyframe* in a summary on selected dimension and navigates in the keyframe cluster as shown in Figure 2. This cluster is *re-clustered* in the selected subspace according to the process of section II-B. Then the median frame of each cluster is considered to present the new video summary. This video summary presents partition of a cluster related to the selected *keyframe* and, thus it gives a refined view of a video summary.

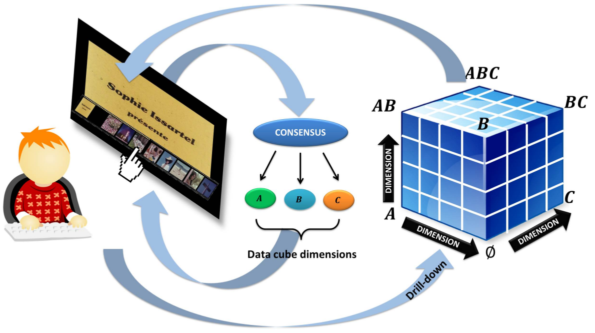

Fig. 2. Data cube and user interaction

## IV. Results and Discussion

This method was tested in a sample of the INA corpus data base. This database consists of 1249 videos of 14 categories of cultural content overall duration of 2000 hours of video. The content of the videos comprises documentaries, TV shows, musical comedies, dancing performances, among others. We consider one video per each category. In average the duration of each video is 41:39 minutes. The number of segments annotated is variable as well as the duration of each segment. *Precision* and *Recall* measures were used to assess the accuracy of the video summary obtained by the proposed method. The ground truth is considered as the human annotation of segments by using the software ELAN [16], where a segment has been considered as a set of contiguous frames with similar audio-visual content. Human annotator establishes the segment boundaries as well as the corresponding label for each segment.

To compare video summaries, we consider the segment intersection between both video segmentations. Thus, *true positives (TP)*, *false negatives (FN)* and *false positives (FP)* are considered. *TP* occurs when a segment is identified by the human operator as well as by the proposed approach. A *FP* refers to over segmentation by the proposed approach: our method gives several segments inside one ground truth segment. Hence, *Precision* and *Recall* are given by (5):

$$\begin{gathered} Precision = \frac{TP}{TP + FP} \\ Recall = \frac{TP}{TP + FN} \end{gathered} \qquad (5)$$

The scalable video summarization model relies on the idea that the refined version of the summary should contain more fine segmentation inside a cluster obtained in the data-cube computation in the complete space.

We propose to include the evaluation of the refined video summary by the following process:

1. Get the label of the selected keyframe.
2. Find the segments with the same label as the selected keyframe.
3. Compare (using *Precision* and *Recall*) the segments considering by the refined video summary w.r.t. The annotated segments.

To assess the scalable video summary, 10 keyframes were randomly selected to do drill-down over them. Thus the average Precision and Recall obtained from the summary displayed when a cross-modal (audio and visual) feature space was chosen is shown in Figure 3. Drill-down operation was done in the early-fused space of the least consensual visual feature and the least consensual audio feature. The number of clusters $K'$ used for the clustering of the segment selected for drill-down is proportional to the length of the segment.

$$K' = \frac{K * l}{L} \qquad (6)$$

Here $K$ is the number of clusters of the base summary, $l$ the length of the targeted segment and $L$ the length of the complete video. We also computed summaries with arbitrary clustering, that is with a constant time step and random clustering for comparison, both computed with the same number of clusters. As illustrated in Figure 3, proposed method gives balanced results w.r.t. Recall-Precision figures, while arbitrary and random abstraction strongly fail on the recall.

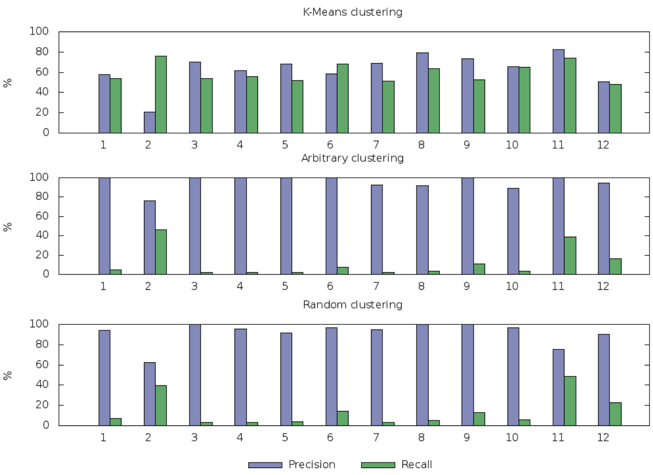

Fig. 3. Precision and recall after a drill down over ten randomly selected segments w.r.t. human segmentation.

Figure 4 illustrates the disagreement over videos by *consensus clustering*. In most of cases there is no agreement between audio features and visual features, while in just 5 cases CSD and PHOG were selected as features with poor agreement w.r.t. the rest of features. In 7 cases SCD was determined as a nonconsensual feature as shown in Figure 4.

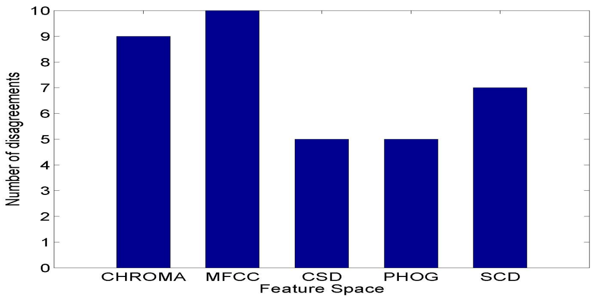

Fig. 4. Disagreements of each feature spaces when one video per category is considered.

We tested the *drill-down* operation for data cube navigating into the clusters of the less consensual feature spaces according to the selected *keyframe*. Due to pre-computed data cube construction, the data cube navigation by OLAP *drill-down* operation to get the refined version of the summary is less than 1 second in a Intel Core 2 Duo 2.53GHz processor with 4GB of RAM.

## V. CONCLUSION AND PERSPECTIVES

In this paper, we have presented a method for construction of scalable video summaries of audio-visual documents based on data cube architecture. This approach provides a customized access to different versions of different levels of detail of a video summary in cross media space. The proposed video summary relies on nonconsensual feature spaces to achieve scalability. We have performed an evaluation of the proposed method with regard to video summaries obtained by a random selection of clusters and arbitrary abstraction with a constant time step and summaries obtained from humans. The method was applied to generic video content without a clearly defined structure, such as cultural documentaries. At this stage of research it is difficult to assess the completeness of the proposed summary with regard to user requirements. Indeed the user discovers the content via scalable browsing. Hence a very large-scale experiment is needed for such an assessment, which has to be conducted in trials following this research work.

In the perspective of this work we will consider richer description spaces and work out the user navigation interface and large scale experiments.

## ACKNOWLEDGMENT

This research has been supported by joint French-Mexican research grant ANR-CONACYT Blanc Mexculture. We thank master student M. Elie Génard for his efficient help in conducting computational experiments.

## REFERENCES

[1] X. Jin et al, “Visual cube and on-line analytical processing of images,” In Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM), pp. 849–858, 2010.

[2] J. Almeida et al, “Online video summarization in compressed domain,” Journal of Visual Communication and Image Representation, vol. 24, pp. 729–738, 2013.

[3] E. Rossi et al, “Clustering of scene repeats for essential rushes preview,” Workshop on Image Analysis for Multimedia Interactive Services, pp. 234–237, 2009.

[4] I. Kompatsiaris et al, TV Content Analysis: Techniques and Applications, 2011.

[5] J. Gray et al, “Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub totals,” Data Min. Knowl. Discov., vol. 1, no. 1, pp. 29–53, 1997.

[6] I. Bartolini, “The Windsurf Library for the Efficient Retrieval of Multimedia Hierarchical Data”, In Proceedings of SIGMM’2011, pp. 139-148

[7] http://www.ina.fr. Institut National de l’Audiovisuel.

[8] W. Yong-ge et al, “Research on image retrieval based on scalable color descriptor of mpeg7,” Advances in Control and Communications, pp. 91–98, 2012.

[9] D. S. Messing et al, “The mpeg-7 color structure descriptor: Image description using color and local spatial information,” In proceedings of the International Conference on Image processing, pp. 670–673, 2001.

[10] A. B. Abdelali et al, “A study of color structure descriptor for shot boundary detection,” International Journal of Sciencies and Techiques of Automatic control and computer engineering, pp. 956–971, 2009.

[11] J. Wang, “Human action recognition based on pyramid histogram of oriented gradients,” International Conference on Systems, Man, and Cybernetics (SMC), pp. 2449–2454, 2011.

[12] V. Peltonen et al, “Computational auditory scene recognition,” in Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2002, pp. 1941–1944.

[13] M. Benoit et al, “YAAFE, an Easy to Use and Efficient Audio Feature Extraction Software,” in Proceedings of the 11th International Society for Music Information Retrieval (ISMIR), 2010, pp. 441–446.

[14] A. K. Jain, “Data clustering: 50 years beyond k-means,” Pattern Recognition Letters, vol. 31, pp. 651–666, 2010.

[15] A. Goder et al, “Consensus clustering algorithms: Comparison and refinement,” In Proc. 9th Workshop on Algorithm Engineering and Experiments (ALENEX’08), pp. 109–117, 2008.

[16] http://tla.mpi.nl/tools/tla-tools/elan/