SPATIOTEMPORAL MODELING AND MATCHING OF VIDEO SHOTS

Eric Galmar, Benoit Huet

Institut Eurécom
Département multimédia
Sophia-Antipolis, France

ABSTRACT

In this paper, we propose a framework to model video sequences using spatiotemporal description of video shots. Spatiotemporal volumes are extracted thanks to an efficient segmentation algorithm. Video shots are described by building an adjacency graph which models the visual properties of the volumes and the spatiotemporal relationships between them. The cost of extracting visual descriptors for the whole shot is reduced by efficiently propagating and merging region descriptors on spatiotemporal volumes. For the comparison of video shots, we propose a similarity measure which tolerates variability in the spatiotemporal representation. Promising experimental results are observed on different visual video shot categories.

***Index Terms*—** Spatiotemporal representation, video shot matching, region-based video indexing and retrieval.

1. INTRODUCTION

The increasing access to video databases has impelled the development of video content analysis area. In order to browse or search particular scenes in large databases, one important aspect is to extract important information for shots in an efficient and reliable way. Hence good representation of video shots and indexing techniques appear both as crucial. Common CBVR systems have mostly relied on either entirely image-based or region-based representation. Spatial segmentation is performed on key-frames to bring out the different shot visual elements, so that the shots are compared by region matching techniques [11, 3]. However, the segmentation is prone to important variations within a video sequence, as the scene changes due to occlusion, shadowing and camera motion.

To make these systems more robust, temporal aspects have also been investigated. One group of techniques extract directly motion descriptors for the indexation of video shots [12]. Motion feature has also been also widely used to segment video shots by tracking regions in consecutive frames, such as in the VideoQ system [1]. The efficiency of these methods is closely linked to the quality of the motion estimation process, which can be degraded in case of complex motion or poorly textured regions. In addition complex moving objects cannot be extracted easily by the low-level features. To overcome the problem, Lee et al. [8] have proposed a graph-based region matching technique using spatial relationships between object regions.

The latest approach to depict video shots is to consider spatial and temporal video data simultaneously. In [2], video shots are modelled by a set of elementary moving color patches extracted from the 3D pixel data. In [6], probabilistic video blobs are considered instead. However, the computational cost of these methods remains high for long sequences.

The proposed method is both related to the spatiotemporal and region-based approaches. We consider that spatiotemporal volumes describe meaningful visual elements and that spatiotemporal relationships between these volumes can underline the visual shot structure. This paradigm can be instantiated by depicting video shots with an Attributed Relation Graph (ARG). The ARG structure is composed of a volume adjacency graph (VAG) representing the relationships among the extracted volumes and of spatiotemporal features as vertex attributes. We further propose an efficient technique to extract volume descriptors that benefits from segmentation properties and a method to build similarity measure between shots adapted to the spatiotemporal representation.

The article is organized as follows. In section 2, we give an overview of the proposed framework. Then in section 3, we explain how we construct the spatiotemporal representation from the video data and introduce how we create the visual volume descriptors. A matching and similarity measure between video shots is presented in section 4. Finally, we illustrate the potential of the framework with a retrieval experiment on different visual shot categories.

2. FRAMEWORK

A global view of the framework is depicted fig.1. The workflow is defined as follows. Video data is assumed to be temporally segmented into video shots from camera changes. The spatiotemporal segmentation module extracts coherent volumes from a given video shot. Then, the corresponding VAG and the segmentation maps are used to produce the ARG attributes by the technique described in section 3.2. Graph and volume descriptors are stored in a database. After this stage, search and retrieval of video shots can be performed by matching of ARGs.

3. SPATIOTEMPORAL MODELING

In this section, we explain the different steps needed to reduce video shot content by the proposed spatiotemporal modeling. Whereas regions extracted from sampled key-frames take into account only local spatial information of the shot, spatiotemporal volumes emphasize shot subparts that remain temporally consistent. Thus more confidence on the relevance of the extracted regions can be obtained from considering shot volumes instead of frame regions.

3.1. Spatiotemporal segmentation

Generally speaking, spatiotemporal segmentation extracts continuous volumes from a video sequence with respect to a certain set of vi-

This research was supported by the European Commission under Contract FP6027026-K-Space.

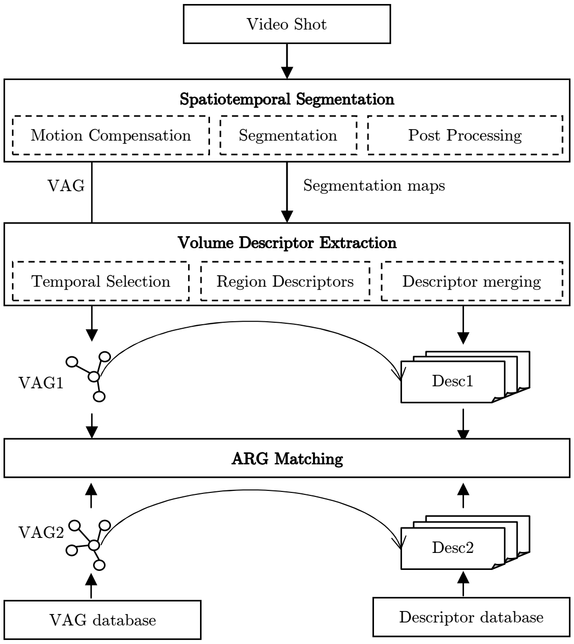


Fig. 1. The spatiotemporal framework for ARG matching.

sual features. Several methods have been proposed, where the most impressive ones are based on graph-cuts [13], but at the expense of an important computational cost.

We consider here the method proposed by [5] which has good trade-off between efficiency and accuracy of the extracted volumes. It is based on the use of graph merging algorithms for grouping pixel and regions into homogeneous volumes.

Before applying the segmentation algorithm, camera motion is compensated by the robust method suggested in [9] using an affine motion model (6 parameters). As a post-processing stage, the volumes can be locally re-segmented when the projected frame regions do not have relevant size.

Figure 2 shows an example of ARG representation for video shots. Each circle represents a volume. The radius is function of the volume size and its color is the mean color of the volume. In the examples, main volumes are related to the head, the jacket of the character and the background which is splitted in several parts. Similar node structure can be found inside the person whereas the background structure changes.

3.2. Volume Descriptors

Visual descriptors have been intensively investigated in content-based image and video retrieval. A standardization of these descriptors is proposed by MPEG-7 and has been proved to work reasonably well for different domain of applications [4]. To extend existing spatial region/image descriptors to volume descriptors, two main approaches can be considered:

- Extraction from the whole volume at once.
- Aggregation of frame region descriptors.

The first approach is straightforward for color-based features which do not depend on the volume mask. In the MPEG-7 standard [7], these include Dominant Color and Scalable Color descriptors. Other spatial descriptors consider spatial distribution of one feature (Edge Histogram, Color Structure) or are based on 2D image transforms (Color Layout, Homogeneous Texture, Region Shape). In the first case, the extension to the spatiotemporal domain requires the redefinition of the descriptors. In the second case, fast implementations of 3D image transforms have been proposed.

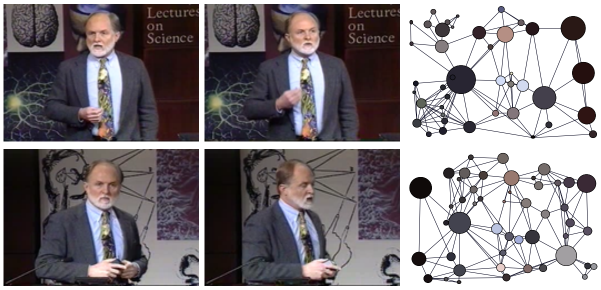


Fig. 2. Examples of video shots and their corresponding ARG.

In the second approach, a volume is considered as a sequence of frame regions. For histogram-based descriptors, MPEG-7 has proposed the Gof/Gop color for joining multiple image frames of a video segment by computing the *mean*, *median* or *intersection* of histograms bins. When adapting to multiple frame regions, the choice of one of these methods is governed by the expected volume properties. Short-length volumes are likely to be homogeneous, so that the descriptors can be averaged. When the volume duration is more important, median could be preferred as the region features have more variability. The intersection is quite pessimistic on the accuracy of the extracted volume, as it represents the least common characteristics between frame regions. The descriptors concerned are typically Edge Histogram, Color Structure, Color Layout, Region Shape. Scalable Color must be first reconstructed in the HSV domain. For Homogeneous Texture, volume descriptor can be obtained by computing the average intensities and energies inside the volume, along with the standard deviations. Aggregation can be preferred in practice as the volume and region descriptors are the same. Moreover, it enables to reuse existing implementations (MPEG-7 XM) and to communicate with MPEG-7 based systems.

Besides spatial descriptors, we also store the sizes and location of the volumes. This last feature can be used to qualify spatiotemporal relationships. More precisely, each volume is located by its center and its bounding box. To further consider temporal evolution of the volumes, we store the locations of the frame regions in the same way.

3.3. Temporal selection

Extracting descriptors in all frames and all volumes of the shot is a tedious task, especially when considering complex ones. Therefore, it could be desirable reduce substantially the cost of this task in function of the available time or the desired accuracy of the descriptors.

For the aforementioned spatial descriptors, we propose to use the segmentation to select temporally a set of frames, extract the region descriptors and finally aggregate the region descriptors to their corresponding volumes. For this purpose, we consider a selection of frames F_T at times T. Given a fixed size for T, we choose the set T_{sel} that maximizes the span of the labeled volumes:

$$T_{sel} = \underset{T}{\operatorname{argmax}} \sum_{V \cap F_T \neq \phi} |V| \tag{1}$$

The first advantage of this criterion is its independence to the

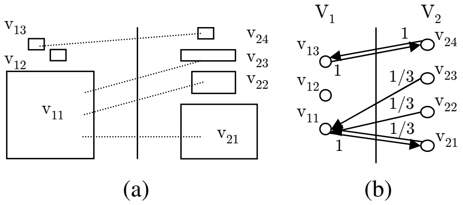


Fig. 3. (a)Example of matching between video volumes. (b) The corresponding bipartite matching graph.

descriptor type. Compared with fixed sampling, a second benefit is that it offers scalability for the extracted descriptors in function of the desired total volume span for the shot. Indeed the span increases with the number of frames selected.

4. SHOT COMPARISON

To compare efficiently video shots, we define a similarity measure between two ARGs. One important point is that a volume in one shot can appeared as split into several volumes in another shot due to scene changes, occlusion and several other factors. In order to address this problem and limit the complexity of the matching, we constrain that each volume in the two ARGs to have at most one match, but we tolerate that two volumes have the same match. This constraint is illustrated fig.3(a).

We define an attributed ARG as a tuple (V, E, ν). A node $v \in V$ is a volume, an edge $e \in E$ a spatiotemporal relationship between two volumes and ν is a fonction that generates node attributes in the set of volume descriptors $\mathcal{D}$. Let $G_1(V_1, E_1, \nu_1)$ and $G_2(V_2, E_2, \nu_2)$ two attributed ARGs. We consider a directed bipartite graph $L(V, E, W)$ with $V = V_1 \cup V_2$ which represents the matching between V_1 and V_2 (fig.3(b)). The arcs in E represents either a match from V_1 to V_2 or V_2 to V_1. We denote respectively by $w_{v_1 \to v_2}$ and $w_{v_1 \leftarrow v_2}$ the weight of the matches from V_1 to V_2 and from V_2 to V_1. For two vertex sets $Q_1 \subset V_1$ and $Q_2 \subset V_2$, we extend the notation $w_{Q_1 \to Q_2}$ as the sum of the weights from Q_1 to Q_2.

The indegree of a node $v_i \in V_i \subset L$ is denoted $deg^+(v_i)$, its outdegree $deg^-(v_i)$. Given the defined constraints on the matching, we have $deg^+(v_i) \in [0, |L/V_i|]$ and $deg^-(v_i) \in [0, 1]$. We impose that the weights of the matches incoming to a node are distributed uniformly. This is defined as follows :

$$w_{v_1 \to v_2} = \begin{cases} 0 & \text{if } deg^+(v_2) = 0 \\ \frac{1}{deg^+(v_2)} & \text{else} \end{cases} \tag{2}$$

$w_{v_2 \to v_1}$ is defined symmetrically. The idea is that when one node has been matched, we do not consider furthermore its properties, but only its relationships in these graphs. To establish the matches we focus on the visual attributes of the ARG. Based on the definition of the ARG and matching graph, we define a similarity measure which takes into account both the structural and the visual properties. The basic similarity between two nodes $(v_1, v_2) \in V_1 \times V_2$, $s_n(v_1, v_2)$ is defined as:

$$s_n(v_1, v_2) = \alpha s_v(v_1, v_2) + \beta s_s(v_1, v_2) \tag{3}$$

s_v is the overall similarity between the volume visual descriptors. Usually, this can be computed by linear combination of the descriptors in $\mathcal{D}$:

$$s_v(v_1, v_2) = \sum_{d \in \mathcal{D}} \alpha_d s_d(v_1, v_2) \tag{4}$$

where s_d the similarity measure defined for the feature $d \in \mathcal{D}$. The structural similarity $s_s(v_1, v_2)$ is based on the matches between their respective neighborhoods $\mathcal{N}_1(v_1)$ and $\mathcal{N}_2(v_2)$ in G_1 and G_2. The principle is inspired by the normalized cuts [10] which is a dissociation measure between two subgraphs.

In the matching graph L, we compare the flow incoming to $\mathcal{N}_2(v_2)$ from $\mathcal{N}_1(v_1)$ to the total flow incoming to $\mathcal{N}_2(v_2)$. This gives the strength of the match from $\mathcal{N}_1(v_1)$ to $\mathcal{N}_2(v_2)$. Moreover, when v_2 and one of its neighbors $n_2 \in \mathcal{N}_2(v_2)$ matches both v_1, v_2 is excluded from $\mathcal{N}_2(v_2)$. Indeed in this case v_2 and n_2 are likely to correspond to subparts of v_1, i.e. they could be merged in a single node. The reasoning is the same for the matches from $\mathcal{N}_2(v_2)$ to $\mathcal{N}_1(v_1)$. We note $M_2(v_1) = \{n_2 \in \mathcal{N}_2(v_2) | w_{v_1 \leftarrow v_2} \neq 0\}$ and $M_1(v_2) = \{n_1 \in \mathcal{N}_1(v_1) | w_{v_1 \to v_2} \neq 0\}$ these excluded vertex sets in the neighborhood of v_2 and v_1, respectively. Thus we consider restricted neighborhood of v_2 to $\mathcal{N}_2^*(v_2) = \mathcal{N}_2(v_2)/M_2(v_1)$ and $\mathcal{N}_1^*(v_1) = \mathcal{N}_1(v_1)/M_1(v_2)$. Formally, the similarity is defined as :

$$s_s(v_1, v_2) = \frac{1}{2} \left(\frac{w_{\mathcal{N}_1(v_1) \to \mathcal{N}_2^*(v_2)}}{w_{V_1 \to \mathcal{N}_2^*(v_2)}} + \frac{w_{\mathcal{N}_1^*(v_1) \leftarrow \mathcal{N}_2(v_2)}}{w_{\mathcal{N}_1^*(v_1) \leftarrow V_2}} \right) \tag{5}$$

Finally, if there are no matched nodes in $\mathcal{N}_1(v_1)$ or $\mathcal{N}_2(v_2)$, the similarity is set to zero, as there are no common matches between the neighborhoods.

Now, we consider the complete ARGs. The total similarity is computed on a set of matched pairs $S \in V_1 \times V_2$:

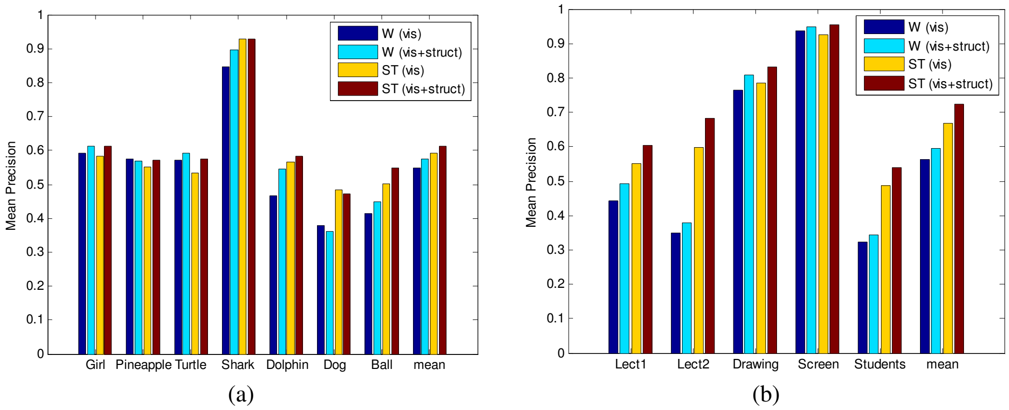
$$s(G_1, G_2) = \frac{1}{|S|} \sum_{(v_1, v_2) \in S} s_n(v_1, v_2) \tag{6}$$

The selection of the matched pairs is based on the visual similarity. First we compute a similarity matrix between V_1 and V_2 and find the best matches S_1 from V_1 to V_2 and S_2 from V_2 to V_1. When the best match for a volume v_1 is not reliable, we further compare $\mathcal{N}_1(v_1)$ to the neighborhood of the possible candidates. For each node in $\mathcal{N}_1(v_1)$, we find the best match in the candidate neighborhood. Then we compute the average distance on the k-best matches, where k is the minimum cardinality of the neighborhoods. In this way more visual information is considered to select the match.

When all matches have been established, the next step consists in pruning the matches which are the most visually different. First, they are not likely to represent the same element, and secondly computing the structural similarity will be not relevant. One method is to consider the distribution of similarities and choose the number of matches $|S|$ from the x-percentile of the distribution. Given a fixed percentile, $|S|$ is low when a few volume matches clearly distinguish from the other, and high if all the distribution of matches is uniform. Finally, the algorithm to compute the similarity measure is summarized below.

1. Compute all the visual similarities between V_1 and V_2.
2. Find the matches S_1 from V_1 to V_2 and S_2 from V_2 to V_1.
3. Build the selection S from S_1 and S_2.
4. Build the matching graph L from S.
5. Compute the structural similarity for the matches in S.
6. Compute the total distance from visual and structural similarities in S.

Table 1. Building of the similarity measure.



(a)

(b)

Fig. 4. Retrieval performance (mean precision). (a) Docon video. (b) Lecture video.

5. EXPERIMENTS

To highlight the advantages of the proposed framework, we conducted experiments for the task of video retrievalWe considered two different videos: a cartoon video (Docon) from the MPEG-7 dataset, and a lecture video from the open-video project. For each video, we defined visual shot categories corresponding to objects appearing along the movie or a certain type of scene. Significant variation is generally observed within each category, due to viewpoint changes, object or camera motion, and interaction between categories.

We compare different approaches for shot representation and matching. First approach is based on key-frame segmentation based on a watersheding technique (*W*) and matching using the visual descriptors only (*vis*). For the second approach, we consider the full similarity measure, including visual and structural parts (*vis*+*struct*). Third and fourth approaches use the spatiotemporal representation (ST) instead of key-frame regions.

Mean average precision results are reported fig.4(a) and fig.4(b) using Color Structure and Edge Histogram as visual descriptors. Globally, the performance observed for each category is function of the variability of the layout and of the extracted descriptors. Best categories retrieved are depicted globally with discriminative visual descriptors(screens, shark), whereas the other ones with more variable descriptors and less common elements are more difficult to retrieve (girl, ball, students).

As regards shot representation, spatiotemporal approach outperforms the key-frame approach, in particular in the lecture video where key-frame regions can be inaccurate and do not reflect well visual elements in the shot. Using the graph structure leverages the results for the categories with more variable descriptors, but where a common structure still remains between shots. The improvement is noticeable for several categories such as lecturer, girl and ball. This effect is also more remarkable on the spatiotemporal representation, as the neighborhood is enlarged and more reliable matches can be found between shots.

6. CONCLUSION

In this paper, we have presented a new approach to construct spatiotemporal modeling of video shots. Volumes are accurately described by a set of visual descriptors, and the structural relations between volumes are represented by an adjacency graph. With an adapted graph matching technique, this description enables to compute shot similarities in an efficient way and can potentially adapt to scene changes. First experiments show that the framework is quite interesting for video indexing and retrieval applications.

7. REFERENCES

[1] S. Chang, W. Chen, W. Meng, H. Sundaram, and D. Zhong. VideoQ: An automatic content-based video search system using visual cues. In *ACM MM*, 1997.

[2] D. DeMenthon and D. Doermann. Video retrieval using spatio-temporal descriptors. In *ACM MM*, pages 508–517, 2003.

[3] Y. Deng and B. Manjunath. Netra-v toward an object-based video representation. *IEEE Trans. CSVT*, 8(5):616–627, 1998.

[4] H.Eidenberger. How good are the visual mpeg-7 features? *Visual Communications and Image Processing*, 5150:476–488, 2003.

[5] E. Galmar and B. Huet. Graph-based spatio-temporal region extraction. In *ICIAR*, pages 236–247, 2006.

[6] H. Greenspan, J. Goldberger, and A. Mayer. Probabilistic space-time video modeling via piecewise gmm. *IEEE Trans. PAMI*, 26(3):384–396, Mar. 2004.

[7] Information Technology - Multimedia Content Description Interface-Part 3: Visual. MPEG ISO/IEC 15938-3, ISO/IEC/JTC1/SC29/WG11/N4358, July, 2001

[8] J. Lee, J. Oh, and S. Hwang. STRG-indexing, spatio-temporal region graph indexing for large video databases. In *ACM SIGMOD*, pages 718–729, 2005.

[9] J. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, 4:348–365, December 1995.

[10] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. PAMI*, 22(2):888–905, Aug. 2000.

[11] J. Smith and S. Chang. Visualseek, a fully automated content-based system. In *ACM Multimedia*, pages 87–98, 1996.

[12] D. W.Zeng, W.Gao. Video indexing by motion activity maps. In *ICIP*, volume 1, pages 912–915, 2002.

[13] Y. Li, J. Sun. H.-Y. Shum Video object cut and paste. In *SIGGRAPH*, 2005.