Video Synopsis based on a Sequential Distortion Minimization Method

Costas Panagiotakis[1], Nelly Ovsepian[2], and Elena Michael[2]

[1] Dept. of Commerce and Marketing, Technological Educational Institute (TEI) of Crete, 72200 Ierapetra, Greece,
cpanag@staff.teicrete.gr,
[2] Dept. of Computer Science, University of Crete, P.O. Box 2208, Greece,
nelli.ov@hotmail.com, elmich@csd.uoc.gr

Abstract. The main goal of the proposed method is to select from a video the most "significant" frames in order to broadcast, without apparent loss of content by decreasing the potential distortion criterion. Initially, the video is divided into shots and the number of synopsis frames per shot is computed based on a criterion that takes into account the visual content variation. Next, the most "significant" frames are sequentially selected, so that the visual content distortion between the initial video and the synoptic video is minimized. Experimental results and comparisons with other methods on several real-life and animation video sequences illustrate the high performance of the proposed scheme.

Keywords: Video summarization; Key frames; Video synopsis;

1 Introduction

The traditional representation of video files as a sequence of numerous consecutive frames, each of which corresponds to a constant time interval, while being adequate for viewing a file in a movie mode, presents a number of limitations for the new emerging multimedia services such as content-based search, retrieval, navigation and video browsing [1]. Therefore, it is important to segment the video into homogenous segments in content domain and then to describe each segment by a small and sufficient number of frames [2] in order to get a video summarization.

Video summarization algorithms attempt to abstract the main occurrences, scenes, or objects in a clip in order to provide an easily interpreted multimedia synopsis. The videos consist of a sequence of successive images which are called frames and represent scenes in movie-motion. The video summarization algorithms are based on detection of representative frames inside on basic temporal units which are called shots. They are designed to detect the most suitable frames from each shot in order to shorten the video without high distortion. A shot can be defined as a sequence of frames that are or appear to be continuously captured from the same camera. Key frames are the most significant images which are extracted from video footage. They have been used to distinguish videos, summarize them and provide access points [3].

Key frames selection approaches can be classified into basically three categories, namely cluster-based methods, energy minimization-based methods and sequential methods [1,4]. Cluster-based methods take all frames from every shot and classify by content similarity to take key-frame. The disadvantage of these methods is that the temporal information of a video sequence is omitted. The energy minimization based methods extract the key frames by solving a rate-constrained problem. These methods are generally computational expensive by iterative techniques. Sequential methods consider a new key frame when the content difference from the previous key frame exceeds the predefined threshold.

In [5], key frames are computed based on unsupervised learning for video retrieval and video summarization by combination of shot boundary detection, intra-shot-clustering and keyframe “meta-clustering”. It exploits the Color Layout Descriptor (CLD) [6], on consecutive frames and compute differences between them define the bounds of each shot. Recently, dynamic programming techniques have been proposed in the literature, such as the MINMAX approach of [7] to extract the key frames of a video sequence. In this work, the problem is solved optimally in $O(N^2 \cdot K_{max})$, where K_{max} is related to the rate-distortion optimization. In [8], a video is represented as a complete undirected graph and the normalized cut algorithm is carried out to globally and optimally partition the graph into video clusters. The resulting clusters form a directed temporal graph and a shortest path algorithm is proposed for video summarization.

Video summarization has been applied by many researchers with multiple approaches. Most of them are dealing with minimizing content features, defining restrictions on distortion, applying simple clustering-based techniques and ignoring temporal variation. In addition, due to its high computational cost ($O(N^3)$ when the number of key frames is proportional to the number of video frames N), most of the prementioned methods have been used to extract a small percentages of initial frames that represent well the visual content but they have not been used to reproduce a video synopsis. Video synopsis is quite important task for video summarization, since it is another short video representation of visual content and video variation. This paper refers to video summarization by the meaning of video synopsis creation. The resulting video synopsis takes into account temporal content variation, shot detection, and minimizes the content distortion between the initial video and the synoptic video. At the same time the proposed method has low computational cost $O(N^2)$. Another advantage of this work is that it can be used under any visual content description.

The rest of the paper is organized as follows: Section 2 gives the problem formulation. Section 3 presents the proposed methodology of the video synopsis creation. segmentation of periodic human motion. The experimental results are given in Section 4. Finally, conclusions and discussion are provided in Section 5.

2 Problem Formulation

The problem of video synopsis belong to video summarization problems. Its goal is to create a new video, shorter than the initial video according to a given parameter α, without significant loss of content between the two videos (the distortion between the original video and the video synopsis is minimized). The ratio between the temporal duration of the video synopsis and the initial video is equal to $\alpha \in [0, 1]$. Let N denote the number frames of original video. Then, the video synopsis consists of $\alpha \cdot N$ frames. Therefore, we have to select the $\alpha \cdot N$ representative key frames. The broadcasting of the video synopsis is done with the original frame ratio meaning that the real speed of the new video has been increased by the factor of $\frac{1}{\alpha}$ on average. For example, we have a video with 5 sec duration with 25 frames/sec, so the whole video is consisted of $5 \times 25 = 125$ frames and the given parameter $\alpha = 0.2$ the final video will have $125 \times 0.2 = 25$ frames. In other words, the final duration will be one sec which is 20% of initial video.

Let C_i, $i \in \{1, ..., N\}$ denote the visual descriptor of i-frame of original video. Let $S \subset \{1, ..., N\}$ denote the frames of video synopsis. According to the problem definition, it holds that the number of frames of video synopsis ($|S|$) is equal to $\alpha \cdot N$. Then, the distortion $D(\{1, ..., N\}, S)$ between the original video and video synopsis is given by the following equation:

$$D(\{1, ..., N\}, S) = \sum_{i=1}^{S(1)} d(i, S(1)) + \sum_{i=S(|S|)+1}^{N} d(i, S(|S|)) \tag{1}$$
$$+ \sum_{i=S(1)+1}^{S(|S|)} min_{S(j) \leq i \leq S(j+1)}(d(i, S(j)), d(i, S(j+1)))$$

where $d(i, S(j))$ denotes the distance between the visual descriptor of i-frame and $S(j)$-frame. $S(j)$ and $S(j+1)$ are two successive frames of video synopsis so that $S(j) \leq i \leq S(j+1)$, this means that $S(j)$ is determined by the index i. The first and the second parts of this sum concern the cases that the frame i is located before the first key frame $S(1)$ or after the last key frame $S(|S|)$, respectively. Therefore, the used distortion that is defined by the sum of visual distances between the frame of original video and the “closest” corresponding frame of video synopsis, can be considered as an extension of the definition of Iso-Content Distortion Principle [1] in the domain of shots.

3 Methodology

Fig. 1 illustrates a scheme of the proposed system architecture. The proposed method can be divided into several steps. Initially, we estimate the CLD for each frame of the original video. Next, we performed shot detection (see Section 3.1). Based on the shot detection results and to the given parameter α we estimate the number of frames per shot that the video synopsis (see Section 3.1). Finally, the

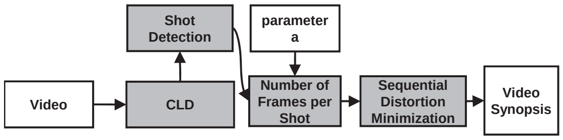

Fig. 1: Scheme of the proposed system architecture.

video distortion is sequentially minimized according to the proposed methods resulting to the video synopsis (see Section 3.2).

The proposed method can be executed under any choice or combination of audio/visual content descriptors. Descriptors based on image segmentation results or on camera motion estimation techniques are computational expensive. Moreover, there is not any guarantee that their results will be accurate for any video content variation [1]. To overcome these problems, we adopt the MPEG-7 visual descriptors [6, 9] as appropriate features, such as the Color Layout Descriptor (CLD), a low cost and compact descriptor, which suffices to describe smoothly the changes in visual content of a shot. These descriptors have been successfully used on our prementioned work on key frames extraction problems [1, 4]. The CLD is a compact descriptor that uses representative colors on a grid followed by a DCT and encoding of the resulting coefficients. We used the following semimetric function D to measure the content distance of two CLDs, $\{DY, DCb, DCr\}$ and $\{DY', DCb', DCr'\}$,
$D = \sqrt{\sum_i (DY_i - DY'_i)^2} + \sqrt{\sum_i (DCb_i - DCb'_i)^2} + \sqrt{\sum_i (DCr_i - DCr'_i)^2}$, where, (DY, DCb, DCr) represent the ith DCT coefficients of the respective color components [1, 4].

3.1 Shot Detection

This section presents the shot detection method. Shot detection is optional and it is used in order, to ensure that the video synopsis contains frames for each shot and to decrease the computational cost of the proposed sequential algorithms. We perform detection for sharp shot changes only. This is done by using the chi-squared distance between of the lightness histogram of each frame with the next one. This histogram distance has been also successfully used for texture and object categories classification, near duplicate image identification, local descriptors matching, shape classification and boundary detection [10].

Hereafter, we present the method that we have used to compute the number of frames of each detected shot for video synopsis. So, the goal of this section is to find out the percentage of the frames of each shot that are capable enough to represent the whole shot. We have used the metric L_k that is defined with the sum CLD distance between all successive frames in shot k: $L_k = \sum_{i \in SH_k} d(i, i+1)$, where SH_k denotes the set of frames of shot k. L_k shows the sum of sequential visual changes of shot k. The higher L_k, the higher number of frames have to be selected from shot k. The selected frames (frames of video synopsis) are also called key frames. The number of frames (b_k) in shot k, that is proportional to L_k, is defined by the following equation: $b_k = \frac{\alpha \cdot N \cdot L_k}{\sum_{k=1}^{|SH|} L_k}$, where $|SH|$ denotes the number of shots. This definition of b_k also satisfies the constraint that the video synopsis should contain $\alpha \cdot N$: $\sum_{k=1}^{|SH|} b_k = \alpha \cdot N$. In the special case of $b_k \leq 1$ which means that all frames of the shot have the same content, we set $b_k = 2$ so that the video synopsis summarizes all of the shots of the video.

3.2 Sequential Distortion Minimization

This section presents the proposed Sequential Distortion Minimization algorithm (SeDiM) for video synopsis creation. This method selects b_k frames for the k shot, so that the distortion between the original video and the video synopsis is sequentially minimized. The ordering of key frames selection corresponds to their significance on content description. The SeDiM method is described hereafter:

Let CAN_k denote the set of candidate frames of shot k for video synopsis. Initially, we set $CAN_k = SH_k$. Let S_k be the frames of video synopsis of shot k. Initially, we set $S_k = \emptyset$. For each shot k, we iteratively select the frame f from CAN_k so that if we include it in set S_k the current video distortion of shot k is minimized (see Equation 2). Next, we remove it from set CAN_k and we add it on set S_k:

$$f = argmin_{u \in CAN_k} \sum_{i \in SH_k} D(SH_k, S_k \cup u) \tag{2}$$
$$CAN_k = CAN_k - \{f\}, \quad S_k = S_k \cup f$$

When the number of key frames of shot k become b_k, CAN_k is being the empty set ($CAN_k = \emptyset$), since we can not select more frames from this shot. The process continues until the number of key frames of video synopsis become $\alpha \cdot N$.

Concerning the computational cost, this procedure can be implemented in $O(N^2)$. The worst case is appeared when the video consists of one shot. In this case it holds that ($N = |SH_1|$). In the start (fist step), the finding of global minima of $D(\{1, ..., N\}, \emptyset)$ needs $O(N^2)$ (see Equation 1). In the n-step of the method, we have to compute $D(\{1, ..., N\}, S \cup u)$ only when the previous or the next key frame of u is the last key frame that have been added in S in previous step ($n-1$). Otherwise, it holds that $D(\{1, ..., N\}, S \cup u) = D(\{1, ..., N\}, S)$. This needs $O(\frac{N^2}{n^2})$, since the video content changes "smoothly" in the sense that the selected frames are about equally distributed during the time. Let $T(.)$ denote the computation cost of the algorithm. It holds that $T(1) = O(N^2)$. In the n-step, we have to find the minimum of $D(.,.)$ that can be given in $O(N)$ and to update the specific values of $D(.,.)$ in $O(\frac{N^2}{n^2})$. So, the total computational cost is $O(N^2)$.

In addition, we have proposed a simple variation of SeDiM that is presented hereafter. In this variation, we just assume the first and last frame of each shot as two starting key frames for video synopsis. So, in the case of one-shot, we initialize $S_k = \{SH_k(1), SH_k(|SH_k|)\}$. This algorithm is called SeDiM-IN. The rest of the process is exactly the same with SeDiM. The proposed methods do not guarantee global minima of distortion, since they sequentially minimize the distortion function. SeDiM guarantees global minima of distortion only in the case of $b_k = 1$.

4 Experimental Results

Fig. 2: Snapshots of videos that we have used in the paper.

In this section, the experimental results and comparisons with other algorithms are presented. We have tested the proposed algorithm on a data set containing more than 100 video sequences. We selected 10 videos (eight real-life and two synthetic (animation)) videos of different content in order to evaluate the distortion of each algorithm and take results from videos which have different content. The real-life videos have been recording either in indoor or outdoor environments. The ten used videos consist of 69 shots. The number of shots per video varies from one to 22. In addition, the duration of the videos varies from 300 frames to 1925 frames. Fig. 2 depicts shapshots from these videos. The names of the videos are given in the first column of Table 1.

4.1 Comparison to Other Algorithms

The proposed methods SeDiM and SeDiM-IN have been compared with the content equidistant and time equidistant algorithms in the same data sets and same set of parameters $\alpha = 0.1$ and $\alpha = 0.3$. Hereafter, we present these two algorithms. The content equidistant algorithm (CEA) is inspired by the work [1], where the iso-content principle has been proposed to estimate the key frames that are equidistant in video content. According to this method, the key frames $\{t_1, t_2, ..., t_{b_k}\}$ in shot k are defined by the following equation: $m \simeq \sum_{u=1}^{t_1-1} d(u,u+1) \simeq \sum_{u=t_1}^{t_2-1} d(u,u+1) \simeq ... \simeq \sum_{u=t_{b_k}}^{b_k-1} d(u,u+1)$ where $m = \frac{1}{b_k-1} \sum_{u=1}^{b_k-1} d(u,u+1)$. So, based on the measurement m, first we compute the key frame t_1, next we compute t_2, and so on. Finally we compute t_{b_k}.

The time equidistant algorithm (TEA) is based on equivalent frames in each shot of video by finding key frames as equal intervals in duration of shot. According to this method, the key frame t_i, $i \in \{1, 2, ..., b_k\}$ in shot k is directly defined by the following equation: $t_i = \lfloor \frac{i \cdot |SH_k|}{b_k} \rceil$, where $|SH_k|$ denotes the number of frames of shot k and $\lfloor . \rceil$ denotes the nearest integer function. This is the simplest method for video synopsis creation, since it does not take into account visual changes.

Table 1: The distortion $D(\{1, ..., N\}, S)$ between the original video and video synopsis.

	$\alpha = 0.1$	$\alpha = 0.1$	$\alpha = 0.1$	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.3$	$\alpha = 0.3$	$\alpha = 0.3$
Dataset	**SeDiM**	**SeDiM-IN**	**CEA**	**TEA**	**SeDiM**	**SeDiM-IN**	**CEA**	**TEA**
foreman.avi	19209	**18973**	21814	22069	**6755.1**	6774.1	7738.2	8992.2
coast_guard.avi	**6962.7**	7054.8	7486.6	7079.9	**2521.4**	2562.4	2669.5	4146.4
hall.avi	**3913.8**	3938.8	4309.1	4444.4	**2137**	2141	2228.1	3863
table.avi	**10207**	11578	11529	10928	**4097**	4113.2	4542.4	6046.4
blue.avi	**13826**	14487	14690	16171	**5419**	5494	5736	10631
doconCut.avi	**116420**	122550	142460	148230	**40503**	43602	45412	70521
data.avi	**14635**	15800	17147	15294	**4292**	4303	5058	17260
Wildlife.avi	**27187**	29841	31763	33752	**9052**	9210	10826	12493
MessiVsRonaldo.avi	**74630**	85270	85310	111070	**20971**	22051	23001	40209
FootballHistory.avi	**68434**	79676	80236	95497	**16402**	16842	17323	58503

Table 1 depicts the distortion $D(\{1, ..., N\}, S)$ between the original video and video synopsis of SeDiM, SeDiM-IN, CEA, TEA methods under the ten used video sequences with $\alpha = 0.1$ and $\alpha = 0.3$.

According to these experiments, SeDiM yields the highest performance results, outperforming the other algorithms, since in 95% of cases (19 out of 20) gives the lowest distortion. SeDiM-IN is the second highest performance method. When $\alpha = 0.3$ is always the second highest performance method. When $\alpha = 0.1$, in 70% of cases is the second highest performance method. In addition, in foreman.avi, SeDiM-IN gives the lowest distortion when $\alpha = 0.1$. SeDiM usually gives less distortion than SeDiM-IN, because the video synopsis of SeDiM-IN contain the first and last frame each shot, without examine if they are appropriate to optimize the summarization of video.

High performance results are also obtained by CEA that is the third highest performance method, especially when $\alpha = 0.3$. We observed that CEA is better method to get video synopsis than TEA because the equal time intervals in the shot don't guarantee that the selected frames from this method have different or same visual content. CEA ensure that the key frames are selected by equal content differences and with this way maintain the distortion of video in low levels. The initial videos and video synopsis results (with $\alpha = 0.1$ and $\alpha = 0.3$) of SeDiM, SeDiM-IN, CEA, TEA methods are given in [3]. It holds that the video synopsis of the proposed schemata describe well the visual content under any type of videos.

5 Conclusion

In this paper, we have proposed a video synopsis creation scheme that can be used in video summarization applications. According to the proposed frame-

[3] https://www.dropbox.com/sh/rpysux4oa746jty/B265lHwpAB

work, the problem of video synopsis creation is reduced to the minimization of the distortion between the initial video and the video synopsis. The proposed method sequentially minimizes this distortion, resulting in high performance results under any value of the parameter α that controls the number of frames of the video synopsis. In addition, the proposed scheme can be used under any type of video content description.

Acknowledgments

This research has been partially co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program “Education and Lifelong Learning” of the National Strategic Reference Framework (NSRF) - Research Funding Programs: ARCHIMEDE III-TEI-Crete-P2PCOORD and THALIS-UOA- ERASITECHNIS.

References

1. Panagiotakis, C., Doulamis, A., Tziritas, G.: Equivalent key frames selection based on iso-content principles. IEEE Transactions on Circuits and Systems for Video Technology **19** (2009) 447–451
2. Hanjalic, A., Zhang, H.: An integrated scheme for automated video abstraction based onunsupervised cluster-validity analysis. IEEE Trans. On Circuits And Systems For Video Tech. **9** (1999) 1280–1289
3. Girgensohn, A., Boreczky, J.S.: Time-constrained keyframe selection technique. Multimedia Tools and Applications **11** (2000) 347–358
4. Panagiotakis, C., Doulamis, A., Tziritas, G.: Equivalent key frames selection based on iso-content distance and iso-distortion principles. In: International Workshop on Image Analysis for Multimedia Interactive Services, IEEE (2007)
5. Hammoud, R., Mohr, R.: A probabilistic framework of selecting effective key frames for video browsing and indexing. In: International workshop on Real-Time Image Sequence Analysis (RISA’00). (2000) 79–88
6. Manjunath, B., Ohm, J., Vasudevan, V., Yamada, A.: Color and texture descriptors. IEEE Trans. On Circuits And Systems For Video Tech. **11** (2001) 703–715
7. Li, Z., Schuster, G., Katsaggelos, A.: Minmax optimal video summarization. IEEE Trans. Circuits Syst. Video Techn. **15** (2005) 1245 – 1256
8. Ngo, C.W., Ma, Y.F., Zhang, H.J.: Video summarization and scene detection by graph modeling. IEEE Trans. Circuits Syst. Video Techn. **15** (2005) 296 – 305
9. Kasutani, E., Yamada, A.: The mpeg-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval. (2001) 674–677
10. Pele, O., Werman, M.: The quadratic-chi histogram distance family. ECCV (2010) 749–762