A Novel Video Summarization Method Based on Compact Composite Descriptors and Fuzzy Classifier

Vicky S. Kalogeiton*, Dim P. Papadopoulos*, Savvas A. Chatzichristofis* and Yiannis S. Boutalis*

Department of Electrical and Computer Engineering, Democritus University of Thrace, Xanthi, Greece
vasikalo@ee.duth.gr, dimipapa4@ee.duth.gr, schatzic@ee.duth.gr, ybout@ee.duth.gr

Abstract

In this paper, a novel method to generate video summaries is proposed, which is allocated mainly for being applied to on-line videos. The novelty of this approach lies in the fact that the authors of this paper transfer the video summarization problem to a single query image retrieval problem. This approach utilizes the recently proposed Compact Composite Descriptors (CCDs) and a fuzzy classifier. In particular, all the video frames are initially sorted according to the distance between an artificially generated, video depended, image. Then the ranking list is classified into a preset number of clusters using the Gustafson Kessel fuzzy classifier. The video abstract is calculated by extracting a representative key frame from every cluster. A significant characteristic of the proposed method is its ability to classify the frames of the video into one or more clusters. Experimental results are presented to indicate the effectiveness of the proposed approach.

Keywords: Video Summarization, Compact Composite Descriptors, Fuzzy Classifier

1. Introduction

In last decades, observing the increasingly use of multimedia data, it is realized that they have penetrated for keeps in our everyday life. A characteristic example of multimedia data is the digital video, whose on-line use, especially the last years, has been increased dramatically. This fact automatically entails that video web sites have become overcrowded and the amount of data has reached to an uncontrollable point. It is no coincidence that in August 2008 YouTube was considered to be the world's second search engine[1]. Consequently, the situation necessitates the generation of a representative video abstraction with a view to facilitating the user to decide rapidly and easily whether or not he is interested in a video without the need to watch the entire video but only the essential content of it.

*Authors of this paper are members of the DUTh EECE Robotic Team
1. http://tinyurl.com/yz5wb8x

Over the last years a noteworthy amount of work in the field of video summarization has been observed. In the literature a lot of significant approaches of this issue are demonstrated.

In particular, in the recent past, two basic forms of video summaries have been proposed [Truong et Al., (2007)]: key frames and video skims. The term of key frame refers to a representative stationary image while video skims refer to a moving-image abstract. Both of the two pre-mentioned forms of generating a video summary are presented in a method that is based on clustering all the frames of a video and extracting the key frames of the most optimal clusters and then the preview is formed using the video shots that the key frames belong to [Hanjalic et Al., (1999)]. It is a fact that the majority of techniques, in which the summarization of a video is aimed, focus on the extraction of key frames instead of the preview of a video. The technique of key frame extraction contributed to the creation of a tool to benchmark different low level features for video summarization [Lux et Al., (2009)] whereas another technique is based on extracting multiple key frames and then using k-medoid algorithms the frames are clustered and the best representative one is chosen [Hadi et Al., (2006)]. A different approach of this issue is the generation of a video index for summary using an automatic tool, based on MPEG-7 visual descriptors [Lee et Al., (2003)]. Many studies, surveys, and research papers on video summarization have been published during the last decade (e.g. [Yeung et Al., (1997)], [Komlodi et Al., (1998)] , [Parshin et Al., (2000)], [Zhang et Al., (2002)], [Ciocca et Al., (2006)], [Matos et Al., (2008)]).

Given the pre-mentioned techniques the authors of this paper have developed a novel approach, which expands the problem of video summarization to a problem of single query image retrieval.

More particularly, the method utilizes the recently proposed Compact Composite Descriptors (CCDs). The effectiveness of CCDs against to several local or global low level features for video summarization has been illustrated in [Lux et Al., (2009)]. CCDs are described in Section 2. According to the proposed method, video is dealt as a sequence of frames. Each frame is dealt as a separate image and is described by CCDs. Additionally, the whole video is described by an artificially generated image, which is generated dynamically from the video. In order to be calculated the video summary, is calculated the distance of CCDs, which describes every frame, with the CCDs of the artificially generated image. Given that in the procedure more than one descriptor participate, late fusion techniques, which are described in Section 3, are used. By terminating the procedure a ranking list is created, which includes all the frames sorted based on their distance from the artificial image. Afterwards, the Gustafson Kessel fuzzy classifier [Gustafson et Al., (1978)] divides the ranking list into a preset number of clusters. More details about this classifier are given in Section

4. The total of the clusters sets the video summary. The frame that corresponds to the center of each cluster is considered as the frame that is able to describe the cluster. A significant characteristic of the proposed method is that whichever frame of the video may participate to one or more clusters. Consequently, a fuzzy video summary is generated. The entire procedure is given in details in Section 5 while the experimental results are shown in Section 6. Finally the conclusions are drawn in Section 7.

2. Compact Composite Descriptors

The family of Compact Composite Descriptors (CCDs) includes the following four descriptors:

i) the Color and Edge Directivity Descriptor (CEDD) [Chatzichristofis et Al., (2008a)],
ii) the Fuzzy Color and Texture Histogram (FCTH) [Chatzichristofis et Al., (2008b)],
iii) the Brightness and Texture Directionality Histogram (BTDH) descriptor [Chatzichristofis et Al., (2010b)] and
iv) the Spatial Color Distribution Descriptor (SpCD) [Chatzichristofis et Al., (2010c)].

The Color and Edge Directivity Descriptor (CEDD) and the Fuzzy Color and Texture Histogram (FCTH) are used to describe natural color images. CEDD and FCTH use the same color information, since two fuzzy systems are applied to them, resulting in reducing the scale of the colors of the image to 24. These 2 descriptors demand a small size for indexing images. The CEDD length is 54 bytes per image while FCTH length is 72 bytes per image. The early fusion of CEDD and FCTH leads to a new descriptor, called Joint Composite Descriptor (JCD**)** [Chatzichristofis et Al., (2009)].

The Brightness and Texture Directionality Histogram (BTDH) descriptor combines brightness and texture characteristics in order to describe grayscale images. A two unit fuzzy system is used to extract the BTDH descriptor; the first fuzzy unit classifies the brightness value of the image's pixels into clusters in order to extract the brightness information using Gustafson Kessel [Gustafson et Al., (1978)] fuzzy classifier and the other one is used to extract texture information suggested by the Directionality histogram in [Tamura et Al., (1978)].

The Spatial Color Distribution Descriptor (SpCD) is used for artificially generated images combining color and spatial color distribution information. This descriptor uses a fuzzy linking system that reduces the scale of the image to 8 colors. SpCD captures the spatial distribution of the color by dividing the image into sub-images not to mention the fact that its length does not exceed 48 bytes per image.

3. Late Fusion of CCDs

The combination of the CCDs is enabled by the use of late fusion techniques. In literature there is a majority of linear late fusion methods such as, Comb SUM**,** Borda Count, IRP and Z-score. In [Chatzichristofis et Al., (2010a)], all the previous techniques were applied in CCDs and it has been proved through experimental process that the more effective method for CCDs is Z-score.
Z-score is a method that involves score distribution (SD) and the aim is to succeed the fusion of ranking lists into one. In the beginning, the scores are normalized to the number of standard deviations depending on if they are higher or lower than the mean score.

The function of this method is described as follows:

Assuming there are three ranking lists for a query, each one for each descriptor (JCD, BTDH, SpCD). Thus, for each image i is calculated the Z-score for each one of the ranking lists according to the function:

$$s'(i) = \frac{s(i) - \mu}{\delta} \tag{1}$$

Where $s(l)$ is the distance of the image i with a query image, μ the mean (average value of distances) and δ the typical deviation. Finally, $s'(l)$ values of each image are sorted in a new ranking list.

4. Gustafson Kessel Fuzzy Classifier

Gustafson Kessel (GK) [Gustafson et Al., (1978)] classifier is an expansion of the fuzzy C-Means classifier. By replacing the Euclidean distance by the Mahalanobis distance, ellipsoidal clusters could also be recognized instead of only spherical ones. In this paper, GK classifier classifies the frames into a preset number of clusters. The centroids of the clusters are equivalent to the key frames of the video.

The Gustafson Kessel algorithm:

Let the total of prototypes $X = \{x_1, x_2, ..., x_n\}$ with $X_i \in R^p$,
Definition:
a) L : the number of the clusters
b) a : the preset maximum number of repetitions
c) v_i : the vector of the cluster center

d) C_i : the covariance array of the cluster

The U^0 table of participatory functions is started, either at random or based on a particular approach. The centers of V^0 clusters and the covariance matrixes C^0 are calculated. Then, the tables A^0 of the clusters are calculated. Next, the U^0 tables are recalculated. A value is set for m . Indicator $a = 0$.

The calculation process starts as:

i) Calculation of v_i :

$$v_i = \frac{1}{\sum_{k=1}^{n} u_{ik}^m} \sum_{k=1}^{n} u_{ik}^m x_k \quad i = 1, 2, ..., n \quad k = 1, 2, ...n \tag{2}$$

ii) Calculation of C_i :

$$C_i = \sum_{k=1}^{n} u_{ik}^m (x_k - u_i) \times (x_k - u_i)^T \quad i = 1, 2, ..., L \quad k = 1, 2, ...n \tag{3}$$

iii) Calculation of A_i :

$$A_i = C_i^{-1} = \sqrt[p]{\det(C_i)} (C_i)^{-1} \quad i = 1, 2, ..., L \tag{4}$$

iv) The Mahalanobis distance of every prototype x_k of the cluster (v_i, A_i) is calculated:

$$d_{ik}^2 = (x_k - v_i)^T \times A_i \times (x_k - v_i) \quad i = 1, 2, ..., L \quad k = 1, 2, ...n \tag{5}$$

v) Then, the new participatory function U^a of every prototype in every cluster is calculated:

$$u_{ik} = \frac{\left[\frac{1}{d_{ik}^2} \right]^{\frac{1}{m-1}}}{\sum_{j=1}^{L} \left[\frac{1}{d_{ik}^2} \right]^{\frac{1}{m-1}}} \quad i = 1, 2, ..., L \quad k = 1, 2, ...n \tag{6}$$

The calculation process is repeating, increasing every time the number of the repetitions a ($a = a + 1$), until $\left| u_{ik}^{(a)} - u_{ik}^{(a-1)} \right| \le e$.

5. Implementation- Method Overview

This method is designed to generate automatically video summaries, which is allocated mainly for being applied to on-line videos. A detailed description of the method is demonstrated in the following steps:

To begin with, the video is decomposed into its frames. Each frame corresponds to independent image. The first step includes the dynamic construction of an artificial image. Every pixel of the artificially generated image is the corresponding most frequent used pixel of all the frames. In other words as artificially generated image is defined an image, whose each pixel is described by the following equation:

$$F(R,G,B)_{x,y} = \sum_{F=1}^{N} p_{Frame}(R,G,B)_{x,y} \tag{7}$$

$$p(R,G,B)_{x,y} = p_{Max(F(R,G,B)_{x,y})}(R,G,B)_{x,y} \tag{8}$$

Where $F(R,G,B)_{x,y}$ is the number of pixels that can be found in the position x,y and their value is $p_{Frame}(R,G,B)_{x,y}$. The (R,G,B) value of pixel of the artificial image in a position x,y equals to the value (R,G,B) of the pixels that have the higher $F(R,G,B)_{x,y}$.

In order to avoid out of memory problems and to make the algorithm more efficient and quicker, all the frames of the video are resized into a smaller size. This procedure is taking place using tiles for each frame, and not the entire frame. For the calculation of the tiles of each frame is used the bicubic method and the final size of each tile is set to be 64X64 pixels. This number is chosen as a compromise between the image detail and the computational demand.

In the next step for each frame of the video the Compact Composite Descriptors (CCDs) are calculated. Note that, the descriptors are calculated from the original frames and not from the resized ones. The CCDs descriptors that are extracted are the Joint Composite Descriptor (JCD), the Brightness and Texture Directionality Histogram (BTDH) descriptor and the Spatial Color Distribution Descriptor (SpCD).

As it has already mentioned, authors expand the problem of video summarization to a single query image retrieval problem. The artificial image is used as the query image in order to retrieve and sort the frames of the video to ranking lists. This sorting is accomplished by calculating the distance between the descriptors of the artificial

image and the descriptors of each frame. The distance is calculated by using the Tanimoto coefficient .The procedure is repeating for every descriptor (JCD, BTDH and SpCD) and in the end three ranking lists are constructed.

Afterwards, given that in the procedure participate three descriptors and hence three ranking lists are created, it is necessary the fusion of these three ranking lists into one. Thus, the late fusion method Z-score is used, which calculates a score (Z-score), which is the distance between each frame and the artificially generated image, and then fuses the three lists into one according to this score. The late fusion method results in the creation of a ranking list, that includes all the frames sorted based on their Z-score.

The final ranking list is actually an array, which illustrates for every frame its distance from the artificially generated image. Those distances are used as input into the Gustafson Kessel fuzzy classifier and are separated into a preset number of clusters. Each cluster corresponds to a “scene”. The total of the “scenes” describes the whole video.

For each cluster there is a representative key frame, which describes the cluster. This key frame is the nearest one of all the corresponding frames to the center of the cluster as it results from the Gustafson Kessel fuzzy classifier. Thus, for each cluster a key frame is extracted. These key frames are considered as the most significant frames of the cluster. A significant characteristic of the proposed method is that each frame of the video may participate to one or more than one of the generated clusters, which leads to the generation of a fuzzy video summary. The method presented in this paper extracts four key frames, while the method can easily be expanded for a largest number of clusters.

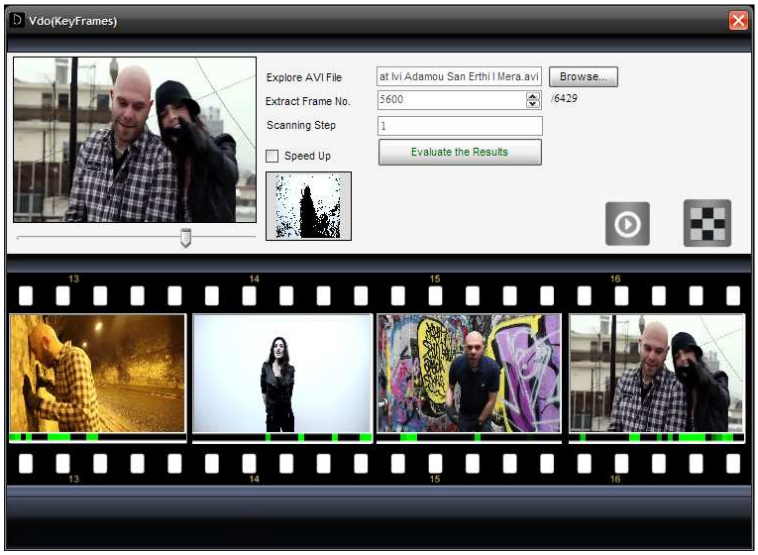

Figure 1. Systems’ Screenshot

In order to be illustrated the participation of every frame in every scene/cluster visually, is used a fuzzy timeline as illustrated in Figure 1. Below every key frame, which has been calculated according to the proposed method, there is a timeline. The green color corresponds to the parts of the video that participate in this scene. The hue of the color corresponds to the degree of the participation value of every frame, according to the following equation:

$$Color = (R, G, B) = (0, m, 0) \quad (9)$$

Where m is the participation value in this class. In case of $m = 1$ (full participation) the color is green, while in case of $m = 0$ the color is black.

6. Experimental Results and Evaluation

In order to evaluate the proposed method, the evaluation method that was proposed in [Yahiaoui et Al., (2001)] is used.

Video URL	Proposed Method	CEDD with K-Means	FCTH with K-Means	AutoCorrelograms with K-Means
http://tinyurl.com/2dk3nsn	93.75	94.20	91.20	**95.60**
http://tinyurl.com/y8rn76q	**78.48**	65.45	72.32	61.23
http://tinyurl.com/336bcxd	**76.92**	72.65	73.20	75.20
http://tinyurl.com/349ge3v	**91.57**	85.20	87.25	90.24
http://tinyurl.com/3y8xlgv	97.26	**98.20**	95.36	**98.20**
http://tinyurl.com/36uoajk	82.24	**86.20**	80.50	80.20
http://tinyurl.com/2vqtchr	**82.83**	70.98	51.40	78.56
http://tinyurl.com/yqr5h7	87.58	72.56	73.88	**89.00**
http://tinyurl.com/qygm56	93.92	**94.50**	92.47	90.24
http://tinyurl.com/24lg6yu	**87.40**	86.75	80.99	83.53
Overall	**87.192**	82.669	79.857	84.200

Table 1. Experimental Results

According to this method, the descriptors of the key frames are extracted and this time, these images are used as query images. That means, that the descriptors of the key frames are compared to the corresponding-descriptors of all the frames and based on the distance between each frame of the video with the key frames, ranking lists are constructed. Thus, at the end of the procedure, for every key frame, three ranking lists are constructed, given that three descriptors participate in the procedure. The late fusion method Z-score is used once more in order to fuse these three raking lists into one. Afterwards, the number of the frames of the video, whose distance from at least one key frame is smaller than a threshold (selected experimentally) according to the pre-mentioned ranking lists, is counted. Thus, the percentage of the frames of the video that correspond to the key frames, out of the total frames of the video is known.

Experiments were done in 10 videos selected from YouTube. The average length of these videos is 4.197 minutes. According to experimental results the average of evaluation results comes up to **87.192%**.
In order to compare the proposed method to other methods, the method that was proposed in [Lux et Al., (2009)] was modified, so that four clusters are generated. As illustrated in Table 1, the proposed method achieves an average score of **87.192%** that appears to be much higher than the score of the other accomplished methods.

7. Conclusions

In this paper, a novel technique to summarize a video, based on the Compact Composite Descriptors and a fuzzy classifier is proposed. The proposed algorithm appears to have better results than the methods used in the literature and is characterized by its ability to classify in a fuzzy way the frames of the video in the produced clusters.

The proposed method supports basic techniques for future expansion. A future plan constitutes the generation of a system, which does not include a preset number of clusters but dynamically calculates their appropriate one.

8. References

Chatzichristofis S., Arampatzis A. (2010), *Late Fusion of Compact Composite Descriptors for Retrieval from Heterogeneous Image Databases,* in Proc.: 33th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland. (To Appear)

Chatzichristofis, S., Boutalis Y. (2008), *Cedd: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval*, in Proc. ICVS08: Vol. 5008. Springer, p. 312.

Chatzichristofis, S., Boutalis Y. (2008), *Fcth: Fuzzy color and texture histogram-a low level feature for accurate image retrieval. In: Image Analysis for Multimedia Interactive Services*, in Proc. WIAMIS'08: Ninth International Workshop on. pp. 191-196.

Chatzichristofis, S., Boutalis Y. (2010), *Content based radiology image retrieval using a fuzzy rule based scalable composite descriptor,* Multimedia Tools and Applications 46: 493-519.

Chatzichristofis, S., Boutalis Y., Lux, M. (2009), *Selection of the proper compact composite descriptor for improving content based image retrieval,* in Proc: 6th IASTED International Conference, Vol. 134643. pp. 064-070.

Chatzichristofis S., Boutalis Y., Lux M. (2010), *Spcd – spatial color distribution descriptor. a fuzzy rule based compact composite descriptor appropriate for hand drawn color sketches retrieval,* in Proc: 2nd International Conference on Agents and Artificial Intelligence (ICAART). pp. 58-63.

Ciocca G., Schettini S. (2006), *An innovative algorithm for key frame extraction in video summarization,* in Journal of Real-Time Image Processing, 1(1) pp. 69-88, 2006.

Gustafson D., Kessel W. (1978), *Fuzzy clustering with a fuzzy covariance matrix,* in Proc: 1978 IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes. Vol. 17.

Hadi Y.; Essannouni F. & Thami, R. O. H. (2006), *Video summarization by k- medoid clustering*, in Proc. SAC 06: 2006 ACM symposium on Applied computing', ACM, New York, NY, USA, pp. 1400-1401.

Hanjalic A. & Zhang H. (1999), *An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis*, in Proc: IEEE Transactions on Circuits and Systems for Video Technology, 1999 9(8), 1280-1289.

Komlodi A., Marchionini G. (1998), *Key frame preview techniques for video browsing,* In DL Proc. : of the 3rd ACM Conference on Digital Libraries. ACM Press, New York. 118-125.

Lee J., Lee G., Kim W. (2003), Automatic *Video Summarizing Tool using MPEG-7 Descriptors for Personal Video Recorder*, IEEE Transactions on Consumer Electronics, 742-749.

Lux M., Schöffmann K, Marques O., Böszörmenyi L. (2009), *A Novel Tool for Quick Video Summarization using Keyframe Extraction Techniques*, in Proc: 9th Workshop on Multimedia Metadata(WMM'09), CEURWorkship Proceedings, vol. 441.

Matos N., Pereira F. (2008), *Using MPEG-7 for Generic Audiovisual Content Automatic Summarization,* in Proc: In Image Analysis for Multimedia Interactive Services, 2008. WIAMIS'08. Ninth International Workshop on, pp. 41-45.

Parshin , Chen L. (2000), *Implementation and analysis of several keyframe-based browsing interfaces to digital video,* Lecture Notes on Computer Science, 2000, vol. 1923, pp. 206.

Tamura H., Mori S., Yamawaki T., (1978), *Textural features corresponding to visual perception,* IEEE Transactions on Systems, Man and Cybernetics 8 (6), 460-473.

Truong B. T. & Venkatesh S. (2007), *Video Abstraction: A Systematic Review and Classification, in ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)* , Vol. 3, No. 1, Article 3.

Yahiaoui I., Merialdo B., Huet B. (2001), *Generating summaries of multi-episode video*, in Proc: IEEE International Conference on Multidedia and Expo 2002.

Yeung M., Leo B. (1997), *Video visualization for compact representation and fast browsing of pictorial conten,*. in: IEEE Trans. Circ. Syst. Video Technol.1997, 7, 5.

Zhang D., Chang F. (2002), *Event detection in baseball video using superimposed caption recognition,* in Proc: tenth ACM international conference on Multimedia 2002. ACM New York, NY, USA, pp. 315-318.