

# Video Summarization Using a Self-Growing and Self-Organized Neural Gas Network

Dim P. Papadopoulos, Savvas A. Chatzichristofis, and Nikos Papamarkos

Department of Electrical and Computer Engineering  
Democritus University of Thrace  
Xanthi 67100, Greece  
{dimipapa4, schatzic, papamark}@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 video summarization problem is considered as a single query image retrieval problem. According to the proposed method, each frame is considered as a separate image and is described by the recently proposed Compact Composite Descriptors (CCDs) and a visual word histogram. In order to classify the frames into clusters, the method utilizes a powerful Self-Growing and Self-Organized Neural Gas (SGONG) network. Its main advantage is that it adjusts the number of created neurons and their topology in an automatic way. Thus, after training, the SGONG give us the appropriate number of output classes and their centers. The extraction of a representative key frame from every cluster leads to the generation of the video abstract. A significant characteristic of the proposed method is its ability to calculate dynamically the appropriate number of clusters. Experimental results are presented to indicate the effectiveness of the proposed approach.

## 1 Introduction

In the last decades, observing the increasingly use of multimedia data, it is realized that they have penetrated 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<sup>1</sup> while in 2010, more than 2 billion videos watched per day on-line<sup>2</sup>. 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/she is interested in a video without the need to watch the entire video but only the essential content of it.

Over the last years a noteworthy amount of work in the field of video summarization has been observed (e.g. [22,29,21,18,4]). In the literature a lot of significant approaches

---

<sup>1</sup> <http://tinyurl.com/yz5wb8x>

<sup>2</sup> <http://www.focus.com/images/view/48564/>

of this issue are demonstrated. Nevertheless, in a recent survey [12] the authors conclude that “video abstraction is still largely in the research phase”. In [27] the authors conclude also that “practical applications are still limited in both complexity of method and scale of deployment”. The main idea behind video summarization is to take the most representative and most interesting segments of a long video in order to concatenate it to a new, smaller, sequence.

Truong et al.[27] proposed two basic forms of video summaries: key frames and video skims. Key Frames, also called representative frames or R-frames is a collection of salient images extracted from the underlying video source. Video skims, also called a moving-image abstract, moving storyboard, or summary sequence consists of a collection of video segments (and corresponding audio) extracted from the original video. One popular kind of video skim in practice is the movie trailer. Both 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 [15]. It is a fact that the majority of techniques, in which the summarization of a video is aimed, are focused on the extraction of key frames instead of the preview of the video.

Video summarization methods can also be separated by the low-level features which are used for content analysis[20]. In general, video summarization is either performed by low level image features (e.g. [6]), audio features (e.g. [28]), textual elements (e.g. [10]), or a fusion of several features (multimedia/multimodal methods, e.g. [21]). Regarding low level image features, authors in [20,19] created a key frame selection tool, which implements summarization of video clips by key frame extraction based on several global and local image features.

In this paper, we propose a new key frame extraction approach using low level features from the visual content of the image that 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 global low level features for video summarization has been illustrated in [19]. Additionally, the proposed method utilizes a visual words (VW) histogram [11]. VWs are inspired directly by the bag-of-words model (BOVW), a well-known and widely used method in text retrieval, where a document is represented by a set of distinct keywords. The same concept governs the BOVW model, in which an image is represented by a set of distinct visual words derived from local features. In [20] the authors conclude that histogram of visual words produces more stable results than the ones based on global image features. CCDs are described in Section 2, whereas BOVW and visual-word histograms are described in details in Section 3.

According to the proposed method, video is considered as a sequence of frames. Each frame is considered as a separate image and is described by CCDs and from a histogram of visual-words. Additionally, the whole video is described by an artificially generated image, which is generated dynamically from the video. Afterwards, the distance of the low level features of each frame with the low level features of the artificially generated image is calculated, in order to extract the video summary. These distances are inserted as input in a powerful Self-Growing and Self-Organized Neural Gas (SGONG) network[3]. The SGONG network has the ability to calculate the optimal number of

output neurons and finally to classify each frame of the video in the appropriate cluster. More details about the SGONG 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 its ability to calculate dynamically the appropriate number of clusters. Consequently, a 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:

1. the Color and Edge Directivity Descriptor (CEDD) [24]
2. the Fuzzy Color and Texture Histogram (FCTH) [24],
3. the Brightness and Texture Directionality Histogram (BTDH) descriptor [8] and
4. the Spatial Color Distribution Descriptor (SpCD) [9]

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) [7].

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 images pixels into clusters in order to extract the brightness information using Gustafson Kessel [14] fuzzy classifier and the other one is used to extract texture information suggested by the Directionality histogram in [26].

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 Bag of Visual Words

Content based image retrieval with global features is notoriously noisy for image queries of low generality, i.e. the fraction of relevant images in a collection [2]. On the other hand, local-feature approaches provide a slightly better retrieval effectiveness than global features [1] but are more expensive computationally [23].

In order to surpass the aforementioned difficulty the Bag-of-visual-words (BOVW) approach is adopted. When employing this approach, the extracted local features are

clustered using k-means classifier and the computed cluster centers are called visual words. The set of visual words forms a visual vocabulary also known as codebook. For every new image added in the collection its local features must be extracted and assigned to the best fitting visual word from the existing codebook. By the end of that process a local feature histogram is composed for each image in the collection. Multiple approaches to enhance the bag of words approach have been proposed in literature [16].

In this paper, a visual word histogram with a universal vocabulary/codebooks is used. SURF [5] local features are extracted from all the 237434 images from the ImageCLEF 2010 Wikipedia test collection. Then, we randomly select 100000 descriptors to create the visual words that will form our codebook. SURF descriptors are clustered in 256 clusters and the mean vector (using k-means) is used as a visual word.

For every frame the SURF features are extracted. The local features are assigned to the best fitting visual word using the nearest neighbor method from the earlier created codebooks.

## 4 Self-Growing and Self-Organized Neural Gas Network

The Self-Growing and Self-Organized Neural Gas (SGONG) Network[3] is an unsupervised neural classifier. SGONG network combines the advantages both of the Kohonen Self-Organized Feature Map (SOFM) [17] and the Growing Neural Gas (GNG) [13] neural classifiers according to which, the learning rate and the radius of the neighboring domain of neurons is monotonically decreased during the training procedure. The SGONG network has been used in [3] in order to reduce the colors of an image. It has also been utilized by a new method for hand gesture recognition[25]. It has the ability to cluster the input data, so as the distance of the data items within the same class (intra-cluster variance) is small and the distance of the data items stemming from different classes (inter-cluster variance) is large. A significant characteristic of this classifier is that it adjusts the number of created neurons and their topology in an automatic way. To achieve this, at the end of each epoch of the SGONG classifier, three criteria are introduced. These criteria are able to improve the growing and the convergence of the network. A main advantage of the SGONG classifier is its ability to determine the final number of clusters.

The SGONG consists of two layers, the input and the output layer. It has the following main characteristics:

- Is faster than the Kohonen SOFM in its convergence.
- In contrast with GNG classifier, a local counter is defined for each neuron that influences the learning rate of this neuron and the strength of its connections. This local counter depends only on the number of the training vectors that are classified in this neuron.
- The dimensions of the input space and the output lattice of neurons are always identical.
- Criteria are used to ensure fast convergence of the neural network. Also, these criteria permit the detection of isolated classes.

The coordinates of the classes' centers are defined by the corresponding coordinates of the output neurons. Each output neuron is described by two local parameters. The first

parameter is related to the training ratio and the second one refers to the influence by the nearby neurons. At the beginning of the training, the SGONG network consists of only two neurons. As the training procedure progresses, the network inserts new neurons in order to achieve better data clustering. Its growth is based on the following criteria:

- A neuron is inserted near the one with the greatest contribution to the total classification error, only if the average length of its connections with the neighboring neurons is relatively large.
- The connections of the neurons are created dynamically by using the Competitive Hebbian Learning method.

The main characteristic of the SGONG is that both neurons and their connections approximate effectively the topology of the input data.

## 5 Implementation- Method Overview

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 of the proposed method includes the dynamic construction of an artificial image. In order to determine the value of each pixel of the artificially generated image, it is executed a uniform color quantization in the frames of the video with 216 unique colors. Thus, every pixel of the artificially generated image is the corresponding most frequent used pixel of all the color quantized frames. In other words, as artificially generated image is defined an image whose each pixel is described by the following equation:

$$F(R, G, B)_{x,y} = \sum_{F=1}^N p_F(R, G, B)_{x,y} \quad (1)$$

$$p(R, G, B)_{x,y} = p_{Max(F(R,G,B)_{x,y})}(R, G, B)_{x,y} \quad (2)$$

Where  $F(R, G, B)_{x,y}$  is the number of pixels that can be found in the position  $x, y$  and their values is  $p_F(R, G, B)_{x,y}$ . The  $(R, G, B)$  value of the 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  $128 \times 128$  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) and the visual words histogram are calculated. Note that, the descriptors are calculated from the original frames and not from the color quantized and 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).



**Fig. 1.** (A) Original Video, (B) Key Frames and Timeline

**Table 1.** Evaluation Videos

Title	Type	URL
Waka-Waka	Video Clip	<a href="http://www.youtube.com/watch?v=pRpeEdMmmQ0">http://www.youtube.com/watch?v=pRpeEdMmmQ0</a>
Al Tsantiri News	Tv Show	<a href="http://www.youtube.com/watch?v=KjbA3L6kQa8">http://www.youtube.com/watch?v=KjbA3L6kQa8</a>
Mickey Mouse	Animation	<a href="http://www.youtube.com/watch?v=jOvFIoBoxag">http://www.youtube.com/watch?v=jOvFIoBoxag</a>
Gummy Bear	Animation	<a href="http://www.youtube.com/watch?v=astISOtCQ0">http://www.youtube.com/watch?v=astISOtCQ0</a>
Radio Arvila	Tv Show	<a href="http://www.youtube.com/watch?v=UHbjy9k53cU">http://www.youtube.com/watch?v=UHbjy9k53cU</a>

As it has already mentioned, the problem of video summarization is expanded 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:

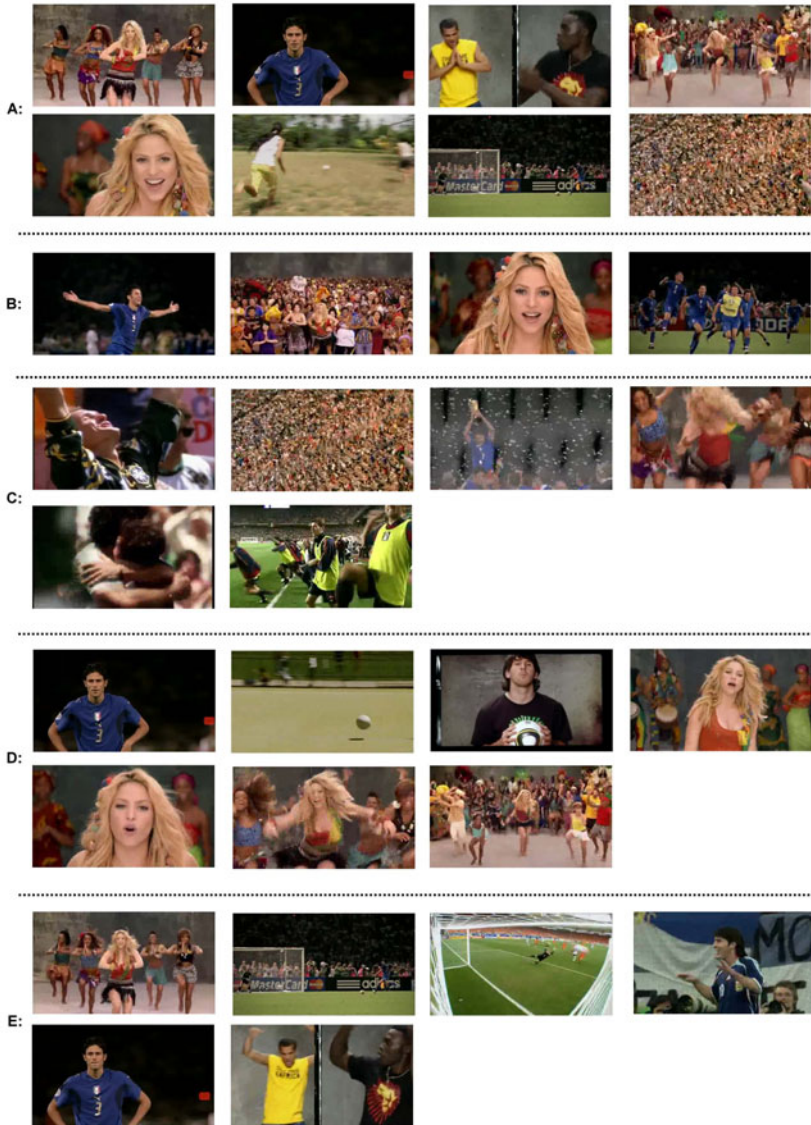
$$D(i, j) = T_{ij} = t(x_i, x_j) = \frac{x_i^T x_j}{x_i^T x_i + x_j^T x_j - x_i^T x_j} \quad (3)$$

where  $x^T$  is the transpose vector of the descriptor  $x$ .

In the absolute congruence of the vectors, the Tanimoto coefficient takes the value 1, while in the maximum deviation the coefficient tends to zero.

The procedure is repeating for every descriptor (JCD, BTDH, SpCD, visual words histogram) and in the end four ranking lists are constructed.

The next step includes the classification of the frames, which is implemented by the Self-Growing and Self-Organized Neural Gas (SGONG) network. The SGONG is fed by the distances of all frames from the artificially generated image. At this point it is worth mentioning that it is required the setting of some important parameters. The setting of these parameters are significant for the correct operation of SGONG network and regard the adding and the removing of the neurons. Moreover, another parameter that should be considered is the maximum number of neurons. This number should be

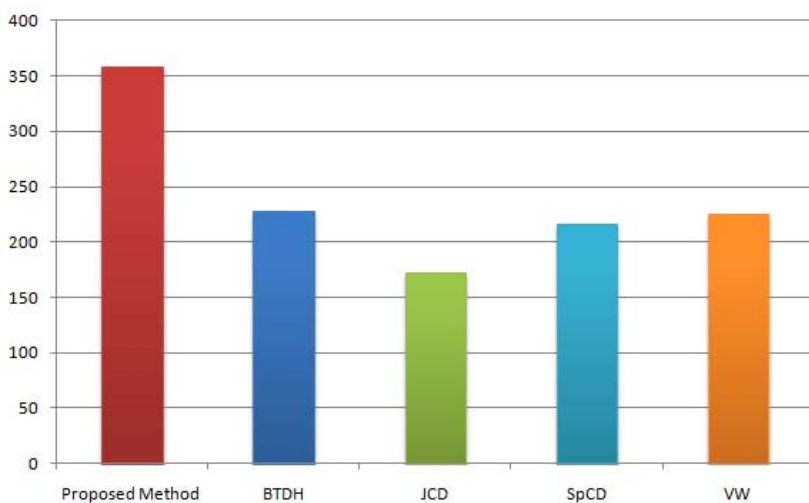


**Fig. 2.** Key Frames Extracted Per Method. (A) Proposed Method Produced 8 Key Frames, (B) BTDH Descriptor Produced 4 Key Frames, (C) JCD Descriptor Produced 6 Key Frames, (D) SpCD Descriptor Produced 7 Key Frames and (E) Visual Words Approach Produced 6 Key Frames.

chosen appropriately in order to the successful convergence of the SGONG network into a number of neurons less than this threshold. Also, the right choice of the learning rate of the network is an absolute necessity.

After training, the weights of the output neurons define the centers of the 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 SGONG 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 its ability to calculate dynamically the appropriate number of clusters, which is based in the main advantage of the SGONG to adjust the number of created neurons and their topology in an automatic way.

In order to be illustrated the participation of every frame in every scene/cluster visually, is used a timeline. For every key frame, which has been calculated according to the proposed method, there is a timeline. The green color (see Fig.1) corresponds to the parts of the video that participate in this scene.



**Fig. 3.** User ratings

**Table 2.** Number of Key Frames Per Method

Title	Length	Proposed Method	BTDH	JCD	SpCD	Visual Words
Waka-Waka	211 s	8	4	6	7	6
Lazopoulos	196 s	8	5	6	4	6
Mickey Mouse	84 s	11	5	6	3	4
Gummy Bear	164 s	8	5	5	5	5
Radio Arvila	209 s	4	5	3	4	6

## 6 Experimental Results

In order to indicate the effectiveness of the proposed method, a user study was held. The proposed method that utilizes the combination of four descriptors was compared with the single utilization of each descriptor. So, users had to choose their favourite summary between five different summaries for each video. In this study five videos are analysed. Each participating user had to mark the five summaries for each video with a degree (5 points for the best down to 1 point for the worst). Sixteen users were participated in the study. Figure 2 shows the five video summaries for the Waka-Waka video extracted from the proposed method and from the four pre-mentioned techniques.

According to the results of the study illustrated in Figure 3, the proposed method reached the highest rating comparing with the other four techniques. More particularly, the score of the proposed method was 358 points with a clear difference from the other four approaches. The method that utilizes the BTDH descriptor came second with 228 points, while the method based on the visual word histogram and the method that utilizes the SpCD descriptor followed with 226 and 216 points respectively. The method that uses the JCD descriptor was found in the last place with 172 points.

The number of key frames extracted by all methods for each video is depicted in Table 2. It can easily be understood, the proposed method generated much more key frames than the other methods. A striking example is the generation of eleven key frames by the proposed approach in the Micky Mouse video, while the average number of the extracted key frames of the other methods is 4.5.

## 7 Conclusions

In this paper, a novel approach to summarize a video, based on a new Self-Growing and Self-Organized Neural Gas network is proposed. The proposed method utilizes the combination of four descriptors in order to describe the frames of the video. Our approach appears to have quite good results according to the user study held for the purposes of this paper. The method seems to be the best out of the other four techniques for each one of which only one descriptor was utilized. In addition, it has the advantage to determine the optimal number of the extracted key frames of a video.

## References

1. Aly, M., Welinder, P., Munich, M.E., Perona, P.: Automatic discovery of image families: Global vs. local features. In: ICIP, pp. 777–780 (2009)
2. Arampatzis, A., Zagoris, K., Chatzichristofis, S.A.: Dynamic two-stage image retrieval from large multimodal databases. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudooh, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 326–337. Springer, Heidelberg (2011)
3. Atsalakis, A., Papamarkos, N.: Color reduction and estimation of the number of dominant colors by using a self-growing and self-organized neural gas. *Eng. Appl. of AI* 19(7), 769–786 (2006)
4. Bailer, W., Dumont, E., Essid, S., Merialdo, B.: A collaborative approach to automatic rushes video summarization. In: 15th IEEE International Conference on Image Processing, 2008. ICIP 2008, pp. 29–32 (2008)

5. Bay, H., Ess, A., Tuytelaars, T., Gool, L.J.V.: Speeded-up robust features (surf). *Computer Vision and Image Understanding* 110(3), 346–359 (2008)
6. Borth, D., Ulges, A., Schulze, C., Breuel, T.M.: Keyframe extraction for video tagging and summarization. In: *Proc. Informatiktage*, pp. 45–48 (2008)
7. Chatzichristofis, S.A., Arampatzis, A., Boutalis, Y.S.: Investigating the behavior of compact composite descriptors in early fusion, late fusion, and distributed image retrieval. *Radioengineering* 19 (4), 725–733 (2010)
8. Chatzichristofis, S.A., Boutalis, Y.S.: Content based radiology image retrieval using a fuzzy rule based scalable composite descriptor. *Multimedia Tools Appl.* 46(2-3), 493–519 (2010)
9. Chatzichristofis, S.A., Boutalis, Y.S., Lux, M.: Spcd - spatial color distribution descriptor - a fuzzy rule based compact composite descriptor appropriate for hand drawn color sketches retrieval. In: Filipe, J., Fred, A.L.N., Sharp, B. (eds.) *ICAART* (1), pp. 58–63. INSTICC Press (2010)
10. Ciocca, G., Schettini, R.: An innovative algorithm for key frame extraction in video summarization. *Journal of Real-Time Image Processing* 1(1), 69–88 (2006)
11. Cula, O.G., Dana, K.J.: Compact representation of bidirectional texture functions. In: *CVPR* (1), pp. 1041–1047 (2001)
12. Dumont, E., Merialdo, B.: Sequence alignment for redundancy removal in video rushes summarization. In: *Proceedings of the 2nd ACM TRECVideo Video Summarization Workshop*, pp. 55–59. ACM, New York (2008)
13. Fritzke, B.: Growing grid - a self-organizing network with constant neighborhood range and adaptation strength. *Neural Processing Letters* 2(5), 9–13 (1995)
14. Gustafson, D.E., Kessel, W.C.: Fuzzy clustering with a fuzzy covariance matrix. In: *IEEE Conference on Decision and Control Including the 17th Symposium on Adaptive Processes*, vol. 17 (1978)
15. Hanjalic, A., Zhang, H.J.: An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis. *IEEE Transactions on Circuits and Systems for Video Technology* 9(8), 1280–1289 (1999)
16. Kogler, M., Lux, M.: Bag of visual words revisited: an exploratory study on robust image retrieval exploiting fuzzy codebooks. In: *Proceedings of the Tenth International Workshop on Multimedia Data Mining, MDMKDD 2010*, pp. 3:1–3:6. ACM, New York (2010)
17. Kohonen, T.: The self-organizing map. *Proceedings of the IEEE* 78(9), 1464–1480 (1990)
18. Lie, W.N., Hsu, K.C.: Video summarization based on semantic feature analysis and user preference. In: *2008 IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing*, pp. 486–491. IEEE, Los Alamitos (2008)
19. Lux, M., Schoffmann, K., Marques, O., Boszormenyi, L.: A novel tool for quick video summarization using keyframe extraction techniques. In: *Proceedings of the 9th Workshop on Multimedia Metadata (WMM 2009)*. *CEUR Workshop Proceedings*, vol. 441, pp. 19–20 (2009)
20. Kogler, M., del Fabro, M., Lux, M., Schoffmann, K., Boszormenyi, L.: Global vs. local feature in video summarization: Experimental results. In: *SeMuDaTe 2009, 10th International Workshop of the Multimedia Metadata Community on Semantic Multimedia Database Technologies, SeMuDaTe 2009* (2009)
21. Matos, N., Pereira, F.: Using mpeg-7 for generic audiovisual content automatic summarization. In: *Ninth International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2008*, pp. 41–45 (2008)
22. Money, A.G., Agius, H.: Video summarisation: A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Representation* 19(2), 121–143 (2008)

23. Popescu, A., Moellic, P.A., Kanellos, I., Landais, R.: Lightweight web image reranking. In: Proceedings of the Seventeen ACM International Conference on Multimedia, pp. 657–660. ACM, New York (2009)
24. Chatzichristofis, S.A., Zagoris, K., Boutalis, Y.S., Papamarkos, N.: Accurate image retrieval based on compact composite descriptors and relevance feedback information. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)* 2, 207–244 (2010)
25. Stergiopoulou, E., Papamarkos, N.: Hand gesture recognition using a neural network shape fitting technique. *Eng. Appl. of AI* 22(8), 1141–1158 (2009)
26. Tamura, H., Mori, S., Yamawaki, T.: Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics* 8(6), 460–473 (1978)
27. Truong, B.T., Venkatesh, S.: Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)* 3(1), 1551–6857 (2007)
28. Xu, M., Maddage, N.C., Xu, C., Kankanhalli, M., Tian, Q.: Creating audio keywords for event detection in soccer video. In: Proc. of ICME, vol. 2, pp. 281–284 (2003)
29. Zhang, D., Chang, S.F.: Event detection in baseball video using superimposed caption recognition. In: Proceedings of the tenth ACM international conference on Multimedia, pp. 315–318. ACM, New York (2002)