Journal of Computer Science 8 (6): 853-858, 2012
ISSN 1549-3636

Video Retrieval using Histogram and Sift Combined with Graph-based Image Segmentation

Tran Quang Anh, Pham Bao, Tran Thuong Khanh,
Bui Ngo Da Thao, Tran Anh Tuan and Nguyen Thanh Nhut
Faculty Mathematics and Computer Science,
University of Science Ho Chi Minh City, Vietnam

Abstract: Problem statement: Content-Based Video Retrieval (CBVR) is still an open hard problem because of the semantic gap between low-level features and high-level features, largeness of database, keyframe's content, choosing feature.In this study we introduce a new approach for this problem based on Scale-Invariant Feature Transform (SIFT) feature, a new metric and an object retrieval method. **Conclusion/Recommendations:** Our algorithm is built on a Content-Based Image Retrieval (CBIR) method in which the keyframe database includes keyframes detected from video database by using our shot detection method. Experiments show that the approach of our algorithmhas fairly high accuracy.

Key words: Content-Based Video Retrieval (CBVR), Content-Based Image Retrieval (CBIR), Scale-Invariant Feature Transform (SIFT), natural important problem, various properties

INTRODUCTION

Finding and retrieving relevant videos from video collections is a natural important problem. It is more and more necessary when videos are generated at increasing rate nowadays. Motivated by this demand, a lot of video retrieval researches have been made to find more effective methods which can be applied in real applications such as video-on-demand systems, digital libraries, Nowadays most of current digital systems support retrieval using low-level features, such as color, texture and motion (Zhu *et al*., 2005) (example: Google's search engine, Yahoo's search engine…). But, generally these features don't reflect users' demands clearly because they only express little content of videos, while the users often care about high-level semantics or concepts. It's a reason why many content-based video retrieval methods have been developed.

Considered as a conceptual extension of CBIR into the video domain (TRECVID, 2006) CBVR problem can be traced back to early 1980s with the introduction of CBIR. Although being a young field, there are many different approaches in CBVR proposed, such asousing visual information methods, retrieval based on textual information presented in the video, relevance feedback algorithms (Geetha and Narayanan, 2008) A framework of these methods often includes breaking videos into shots, keyframes and retrieve suitable keyframes for input data based on some chosen features extracted from these shots or key frames (Flickner *et al*., 1995) There are many different approaches which focus on various properties of frames and videos (such as visual effects, motion, sound,) used to solve each sub-problem.

A common first step for most content-based retrieval techniques is shot segmentation. Even if there are some approaches do not use histogram, histogram difference is still the most widely used method (Geetha and Narayanan, 2008) Many shot detection techniques use it as a feature, such as a feature optimal choice method based on rough-fuzzy set of (Han *et al*., 2005) hidden Markov model method of (Boreczky and Lynn, 1998) sliding window method of (Li and Lee, 2005) and some other directly bases on histogram, such as the method of (O'Toole *et al*., 1999) and our method, which is presented.

Keyframe feature extraction is always one of main study in video retrieval problem, especially when video retrieval techniques are mostly extended directly or indirectly from image retrieval techniques nowadays. Although this approach does not use the spatial-temporal relationship among video frames effectively, this extension also gains some success (Geetha and Narayanan, 2008) in our model, SIFT feature is chosen due to its ability of being almost unchanging under variations of recording frames (light intensity, rate and geometric transformations). Moreover, SIFT detection algorithm runs fast and SIFT matching algorithm has high precision and recall.

For a large video database, clustering is always chosen to abbreviate and organize the content of videos. In most case, it is used to create a useful indexing scheme for video retrieval by grouping similar shots. There are mainly two types of clustering: partition

Corresponding Author: Pham Bao, Faculty Mathematics and Computer Science, University of Science, Ho Chi Minh City, Vietnam

clustering where similar data is arranged into separate clusters (example: shot clustering techniques of (Cao *et al*., 2003) K-means, ISODATA,) and hierarchical clustering which generates a hierarchical classification tree and considers groups as nodes of the tree (Geetha and Narayanan, 2008) That means hierarchical clustering methods tell us relationship (in tree structure) of different groups at different levels. Therefore, in our scheme, we choose a hierarchical clustering method for clustering process. Moreover, we apply a new metric to "increase the difference" between feature vectors (in compare to Euclidean metric).

The object of this study is to retrieve from video database frames which are similar in terms of vision with an input image or object. We describe this process as follow: In section 2, we present the framework of our algorithm. We provide a shot detection method in section 3. Then the next section describes a process of clustering keyframes and builds an index file.

Section 5 mentions three techniques: graph-based segmentation, finding representative vector of each object by using SIFT feature and clustering these vectors. Our new metric is also described in this section. We present results of our experiment in section 6. And section 7mentions some conclusions and extensions.

Video retrieval framework: We change video database to feature vectors to compare with feature vectors extracted from a query image. So the goal here is to extract SIFT feature (Lowe, 1999). In this study we create a video retrieval system by combining some available techniques such as shot detection (Anh *et al*., 2011) graph-based segmentation (Felzenszwalb and Huttenlocher, 2004) SIFT detection algorithm (Lowe, 1999) Model of our system is shown in Fig. 1.

Pre-processing:

- Segmenting each video in the database into shots
- Extracting keyframes from shots. Then we cluster them to get a database of representative keyframesand create an index file to link between them and corresponding videos
- Segmenting and extracting SIFT features from representative keyframes. Calculating feature vector for each object
- Reducing database one more time by clustering objects. Each group of objects is represented by a feature vector

Retrieval: Querying image is proceeded simultaneously according to two stages. At stage 1, we segment the image into objects and calculate SIFT feature vectors of these objects. At state 2, matching state, representative objects which is the most similar to input objects are chosen and keyframes containing them are shows as results.

Our system consists of retrieving based on entire input image or on an object in an image. We use a new metric to match feature vectors of objects in query image with feature vectors in database to determine results.

Shot detection: As we mention above, the popular first step in CBVR schemes is segmenting video into shots. A shot is a group of consecutive frames from the start to the end of recording in a camera which is used to describe a context of a video such as a continuous action, an event, (Geetha and Narayanan, 2008). In our study, we use a novel method combining between image subtraction and histogram comparison method of a research group in University of Science, Vietnam (Anh *et al*., 2011) The algorithm is fast in processing, has acceptable accuracy and study well on cut shot.

The method contains two steps: image subtraction and histogram comparison. The first step built based on an idea: two frames in a same shot are very similar. Therefore, authors measure difference between frame A and its successive frame B at pixel (x_i, y_i) by using gray level of two frames (A(i, j) and B (i, j)) as following Eq. 1:

$$X(i, j) = |A(i, j) - B(i, j)| \quad (1)$$

where, A, $B \in_{M+N} (R)$ After getting the matrix X as the subtraction between A and B, the authors use two thresholds δ_1 and δ_2 to determine if the two frames belong to a shot or not by considering the number of elements of X which is larger than δ_1 (called α (A, B)): A and B are set to belong to a same shot if α (A,B) is smaller than the threshold δ_2.

This step can identify cut shot quickly and accurately. However, the movement of objects in a shot causes much difference in subtraction matrix, that lets to surplus detection. To overcome this problem, authors use histogram comparing. Assuming that two frames A and B are not set to be in a same shot in the first step, authors compute histogram difference between them by Eq. 2:

$$\beta(A, B) = \sum_{0 \le i \le 255} \left| p_i(A) - p_i(B) \right| \quad (2)$$

where, p_i (A) and p_i (B) are values of histogram of A, B at gray level i correspondingly. If β (A, B)>δ_3 (for a chosen thresholdδ_3) then authors conclude that they are frames from two different shots, otherwise they are considered as frames from one shot.

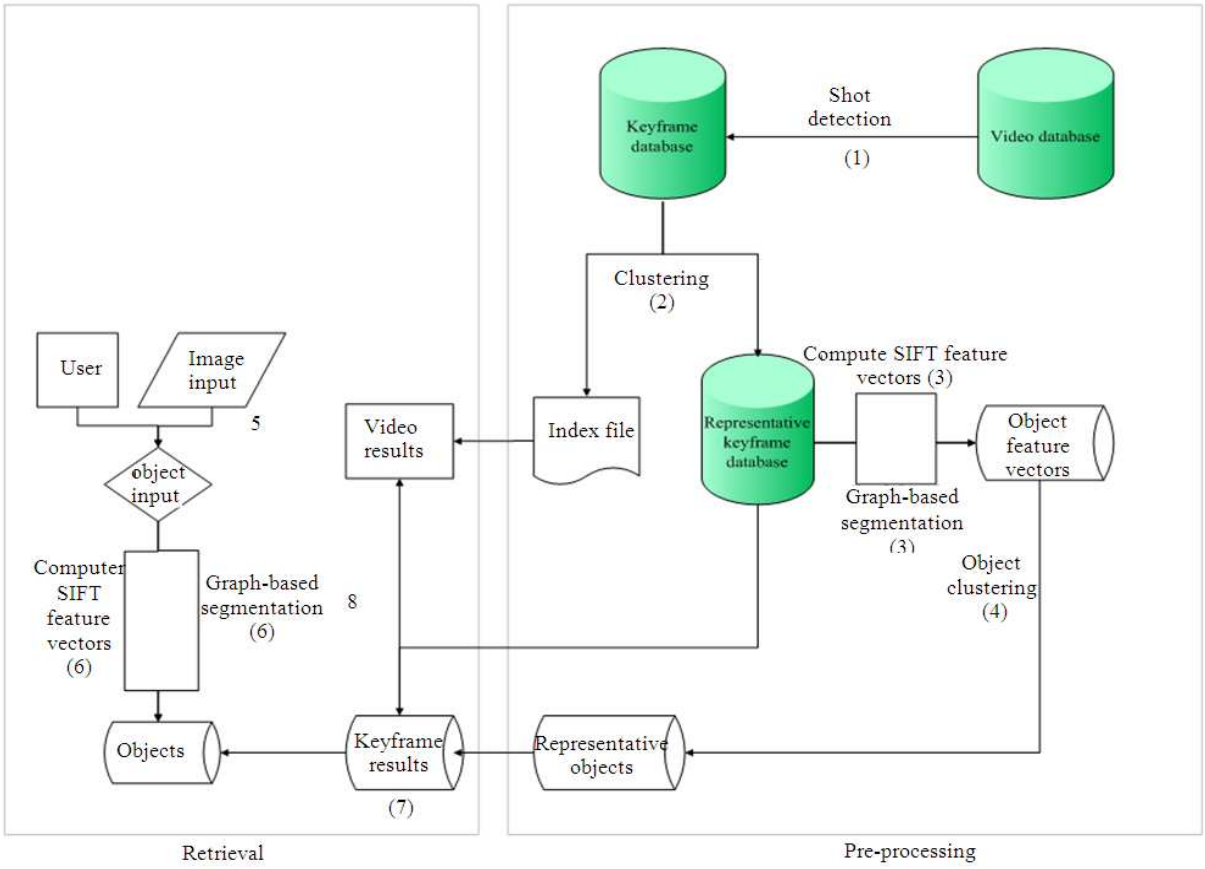


Fig. 1: General model of video retrieval system. We present step (1) in part 3, step (2) in part 4, step (3) and (6) in part 5.1, step (4) in part 5.2 and 5.3, step (7) in part 5.3

Keyframe clustering: Due to the shot detection method (Anh *et al*., 2011) the length of shots is usually short (about 1-5 sec), so choosing the first frame in each shot as the only keyframe for the shot is enough to preserve the shot's content. At the same time, an index file is created to save information about each keyframe (the cover video, its position in the video). In order to reduce the size of keyframe database, these keyframes are clustered as following:

- First, from each keyframe, the mean of all SIFT descriptor vectors is calculated and considered as a mean SIFT feature of the keyframe
- The above mean SIFT vectors are cluster into groups based on the complete-link algorithm (Jain and Dubes, 1988) and our metric
- The first keyframe in each group is taken as representative keyframe of the group
- At the same time, a second index file is created to link between representativekeyframes, all keyframes and videos to inform videos which each representative keyframe "belong to" (corresponding keyframe in group belongs to) as well as its position

Keyframe segmentation and feature vectors clustering:

Keyframe segmentation: One of the most important processes for a keyframe database is to compute feature vectors. We don't describe each representativekeyframeby a feature vector, but each object segmented from a representativekeyframe by one vector. We start with representative keyframes and output groups of the feature vectors.

Although using an image for input, users often focus on one particular object in the image such as actor, item, animal, rather than the whole. To satisfy this demand, we segment every keyframe into regions (objects). We use Pedro F. Felzenszwalb and Daniel P. Huttenlocher's graph-based image segmentation

method (Felzenszwalb and Huttenlocher, 2004) after an image is segmented by this algorithm, there is always evidence for a boundary between every pair of objects in image. Besides the algorithm satisfies two global properties, runs in time nearly linear in the number of edges of graph, a representation of the segmented image and preserves detail in low-variability image regions while ignoring detail in high-variability regions (Felzenszwalb and Huttenlocher, 2004).

Feature vectors clustering: In the SIFT framework (Lowe, 1999) interest points on objects in an image are called keypoints and there is a descriptor vector corresponding to each key point. And this approach often generates large numbers of descriptor vectors from an image, so to use it we must solve a problem: matching process is slow. In study (Anh *et al*., 2010) authors propose an idea to overcome this difficulty. They replace N descriptor vectors corresponding to N keypoints on an object with mean of the vectors. By using this method each object is represented by one mean descriptor vector.

After completing the above processes we get a large collection of feature vectors. In order to retrieval processing run more quickly, we cluster these vectors. We also use complete-link algorithm (Jain and Dubes, 1988) for this study. A representative vector of one cluster is mean of all vectors in that cluster.

A new metric: To applying the clustering algorithm and the matching process, we created a new metric on $\mathbb{R}^{128}$ based on SIFT descriptor vectors' characteristic. Some SIFT descriptor vector's components are always large and some other components are always small. For example, for one descriptor vector, 9th component, 17th component, 41st component and 49th component are almost more large than 0.1 and sometimes more larger than 0.2, but 4th component, 6th component, 7th component 8th component are almost smaller than 0.5.

If we choose 9th component as a landmark and set its value to 3.25 (in order to $\sum_{i=1}^{128} ai = 128$ then value of other components in the above example is approximated alternately as follow.

Denoting a_i as the approximated value of i^{th} thcomponent.After some experiments we find out that for two descriptor vectors x, y, if a_i is small then $|x_i-y_i|$ is often small and if a_i is large then $|x_i-y_i|$ is often large, too. So, we define a new metric Eq. 3:

$$d_p : \mathbb{R}^{128} \times \mathbb{R}^{128} \to \mathbb{R}$$

$$(x,y) \mapsto \sqrt[p]{\sum_{i=1}^{128} a_i \left|x_i - y_i\right|^p} \quad (3)$$

For every:

$$p \in [1,\infty), x = (x_1,...,x_{128}), y = (y_1,...,y_{128}) \in \mathbb{R}^{128}$$

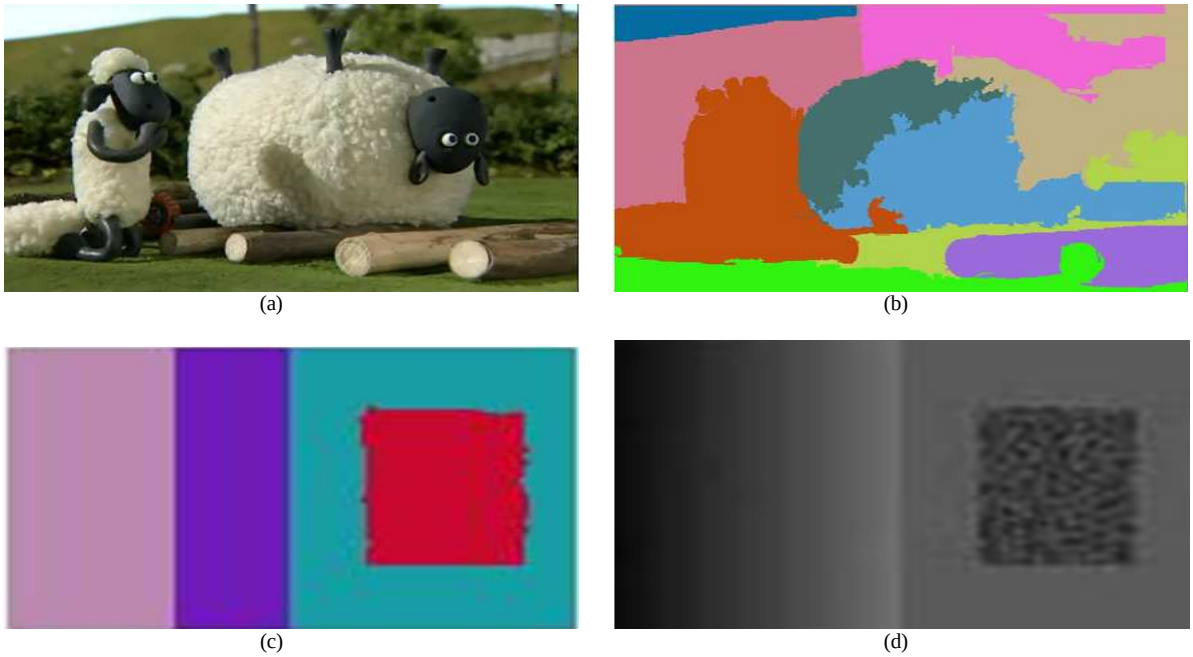
In comparing with Euclidean metric, this metric "increases distance" between two descriptor vectors x, y by increasing large components and decreasing small component. Therefore, we can easily choose clustering threshold and get a better result of this process.

To evaluate the performance of our system, we performed experiments on a medium video database (200G) of elevencategories which represent distinct contents rather than a scene. Since many keyframes are blurred (due to the effect of films, fast movement of objects…) or just contain a part of a real object (an actor, an animal…), the results are influenced a lot.

For query keyframes fromdatabase, the results are high accurate (more than 90% in our experiments). For query images not in database and their content are different a little from the content of keyframes in database, the query result precision is about 30%. We test for 100 images of 10 different categories of interest. The following are our detailed experiments:

CONCLUSION

In a movie, the movement of main objects (people, vehicle,) and the variation of background create different shots, although many shot contains same main objects. Therefore, clustering a main object at different shots (if this object does not change much) into a cluster is an important request to reduce the largeness of keyframe database. Because of the ability of the segmentation process to separate main objects from their correlative background with acceptable accuracy and the ability of being invariable under the changing of geometry transforming and rate, the scheme of keyframe segmentation, calculating SIFT feature and object retrieving can recognize similar main objects from different shots with good accuracy (Fig. 2-4). Or we can say that the schemeis a good choice to solve the above request. Moreover, since SIFT feature is unchanged under the varying of light intensity; it rejects the lighting effects used in movie in clustering process (see the first cluster in Fig. 2). In summary, our algorithm study fairly well on retrievalling query images with some geometry, light variations from some keyframes. But that is different with other variations such as feeling variations, changing of background.



(a) (b)

(c) (d)

Fig. 2: Two images (a) and (c) are segmented into objects (images (b) and (d)) with acceptable accuracy

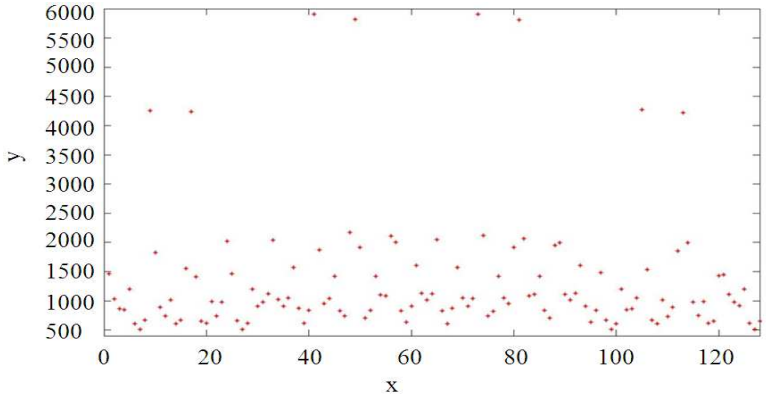
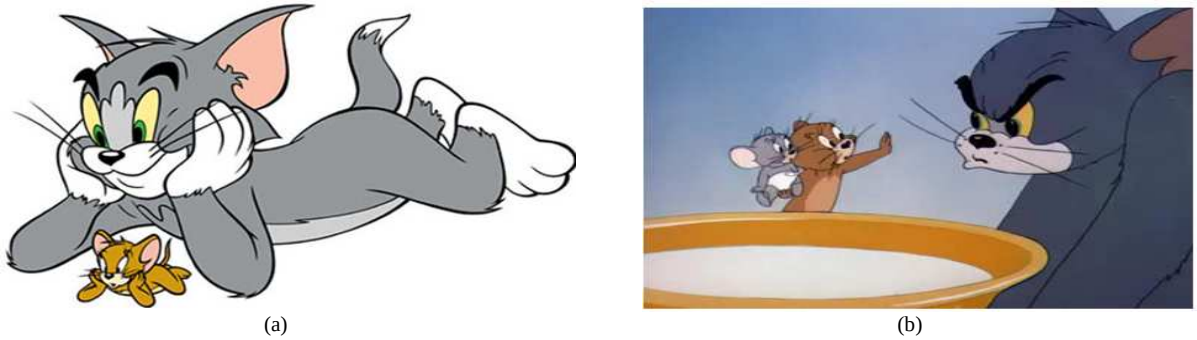


Fig. 3: Sum of representative descriptor vectors of all objects in 2000 randomrepresentativekeyframes. x-axis contains 1,… 128 and y-axis is value of each component of the sum vector



(a) (b)

Fig. 4: (a) a query image, (b) a corresponding result (a representative keyframe) from a movie “Tom and Jerry” in the database

Table 1: Approximated value of 128 components (1st component is 1, 2nd component is 0.75, 3rd component is 0.75, so on)

1	0.75	0.75	0.5	1	0.5	0.5	0.5	3.25	1.5	0.5	0.5	0.75	0.5	0.5	1.25
3.25	1.00	0.50	0.50	0.75	0.50	0.75	1.50	1.0	0.50	0.50	0.50	1.00	0.75	0.75	0.75
1.5	0.75	0.75	0.75	1.25	0.75	0.50	0.50	4.5	1.50	0.75	0.75	1.00	0.50	0.50	1.50
4.5	1.50	0.50	0.50	1.00	0.75	0.75	1.50	1.5	0.50	0.50	0.75	1.25	1.00	0.75	0.75
1.5	0.50	0.50	0.75	1.25	0.75	0.75	0.75	4.5	1.50	0.50	0.50	1.00	0.75	0.75	1.50
4.25	1.50	0.75	0.75	1.00	0.50	0.50	1.50	1.5	0.75	0.75	0.75	1.25	0.75	0.50	0.50
1	0.50	0.50	0.50	1.00	0.50	0.50	0.75	3.0	1.00	0.50	0.50	0.75	0.50	0.75	1.50
3.25	1.50	0.75	0.50	0.75	0.50	0.50	1.00	1.0	0.75	0.75	0.75	1.00	0.50	0.50	0.50

Table 2: Experiment result. The columns show the accuracy and average query time of the three methods on three rows

Shot detection/ Retrieving	Recall (%)	Precision (%)	The average query time
Shot detection	61.0000000	39.0000000	5.4s/MB
Retrieving based on an object	65.3061224	18.7683284	38.83861s
Retrieving based on entire image	46.3917526	22.0588235	77.980265s

In this study, we developed a video retrieval system combining between histogram; SIFT algorithm, graph-based segmentation method and complete-link algorithm which has advantage of simplicity and efficiency in searching distinct objects rather than a scene. Users can use an input image or an object of that image to retrieve (Table 1-2). Moreover, the system can be applied easily to the specific data domains, for instance, video shot retrieval for face sets (Lowe, 1999) events. However, our system has two main disadvantages: long query time, surpluses in detecting gradual shot transitions. So, our future study is to overcome those disadvantages to have a better video retrieval system.

REFERENCES

Anh, N.D., P.T. Bao, B.N. Nam and N.H. Hoang, 2010. A new CBIR system using SIFT combined with neural network and graph-based segmentation. Lecture Notes Comput. Sci., 5990: 294-301. DOI: 10.1007/978-3-642-12145-6_30

Anh, T.Q., P. Bao, T.T. Khanh and B.N.D Thao, 2011. Shot Detection Using Histogram Comparison and Image Subtraction. GESTS Int. Trans. Comput. Sci. Eng.

Boreczky, J.S. and L.D. Lynn, 1998. A hidden Markov model framework for video segmentation using audio and image features. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, May 12-15, IEEE Xplore Press, Seattle, pp: 3741-3744. DOI: 10.1109/ICASSP.1998.679697

Cao, Y., W. Tavanapong, K. Kim and J.H. Oh, 2003. Audio-assisted scene segmentation for story browsing. Lecture Notes Comput. Sci., 2728: 446-455. DOI: 10.1007/3-540-45113-7_44

Felzenszwalb, P.F. and D.P. Huttenlocher, 2004. Efficient graph-based image segmentation. Int. J. Comput. Vision, 59: 167-181. DOI: 10.1023/B:VISI.0000022288.19776.77

Flickner, M., H. Sawhney, W. Niblack, J. Ashley and Q. Huang, 1995. Query by image and video content: The QBIC system. IEEE Comput., 28: 23-32. DOI: 10.1109/2.410146

Geetha, P. and V. Narayanan, 2008. A survey of content-based video retrieval. J. Comput. Sci., 4: 474-486. DOI: 10.3844/jcssp.2008.474.486

Han, B., G. Xinbo and J. Hongbing, 2005. A shot boundary detection method for news video based on rough-fuzzy sets. Int. J. Inform. Technol., 11: 101-111.

Jain, A.K. and R.C. Dubes, 1988. Algorithms for Clustering Data. 1st Edn., Prentice Hall, Englewood Cliffs, New Jersey, ISBN-10: 013022278X, pp: 320.

Li, S. and Lee, 2005. An improved sliding window method for shot change detection. Proceeding of the 7th IASTED International Conference Signal and Image Processing, Aug. 15-17, USA., pp: 464-468.

Lowe, D.G., 1999. Object recognition from local scale-invariant features. Proceedings of the 7th IEEE International Conference on Computer Vision, Sep. 20-27, IEEE Xplore Press, Kerkyra, Greece, pp: 1150-1157. DOI: 10.1109/ICCV.1999.790410

O'Toole, C., A.F. Smeaton, N. Murphy and S. Marlow, 1999. Evaluation of automatic shot boundary detection on a large video test suite. Proceeding of the 2nd U.K. Conference Image Retrieval: The Challenge of Image Retrieval, Feb. 25-26, UK., pp: 1-12.

TRECVID, 2006. An overview of up-to-date methods in content-based video retrieval --- by examining top performances in TREC video retrieval evaluation. TRECVID.

Zhu, X., X. Wu, A.K. Elmagarmid, Z. Feng and L. Wu, 2005. Video data mining: Semantic indexing and event detection from the association perspective. IEEE Trans. Knowl. Data Eng., 17: 665-677. DOI: 10.1109/TKDE.2005.83