SVM-based Relevance Feedback for semantic video retrieval

Hadi Sadoghi Yazdi*, Malihe Javidi
and Hamid Reza Pourreza

Department of Computer Engineering,
Ferdowsi University of Mashhad,
P.O. Box 91775-1111, Mashhad, Iran
E-mail: h-sadoghi@um.ac.ir
E-mail: sadoghi_y@yahoo.com
E-mail: malihejavidi@gmail.com
E-mail: hpourreza@um.ac.ir
*Corresponding author

Abstract: This paper presents a novel method for efficient key frame extraction from video shot representation and employs a Support-Vector-Machine-based Relevance Feedback (SVM-RF) to bridging semantic gap between low-level feature and high-level concepts of shots. We introduce a new approach for key frame extraction using a hierarchical approach based on clustering. Using this key frame representation, the most representative key frame is then selected for each shot. Furthermore, our system incorporates user to judge about the result of retrieval and labelled retrieved shot in two groups, relevant and irrelevant. Then, by mean feature of relevant and irrelevant shots train an SVM classifier. In the next step, video database is classified in two groups, relevant and irrelevant shots. Suitable Graphic User Interface (GUI) is shown for capturing RF of user. This process continued until user satisfied with results. The proposed system is checked over collected shots from Trecvid2001 database and home videos include 800 shots of different concepts (10 semantic groups). Experimental results demonstrate the effectiveness of the proposed method.

Keywords: relevance feedback; semantic gap; support vector machine; video retrieval; key frame; semantic gap; RF; relevance feedback.

Reference to this paper should be made as follows: Yazdi, H.S., Javidi, M. and Pourreza, H.R. (2009) 'SVM-based Relevance Feedback for semantic video retrieval', *Int. J. Signal and Imaging Systems Engineering*, Vol. 2, No. 3, pp.99–108.

Biographical notes: Hadi Sadoghi Yazdi received the BS Degree in Electrical Engineering from Ferdowsi University of Mashad in 1994, and then he received to the MS and PhD Degrees in Electrical Engineering from Tarbiat Modarres University of Tehran, Iran, in 1996 and 2005 respectively. He works in Computer Department as an Assistant Professor at Ferdowsi University of Mashhad. His research interests include adaptive filtering, image and video processing, and optimisation in signal processing. He has more than 140 journal and conference publications in subject of interesting area.

Malihe Javidi obtained the BS and MS Degrees in Computer Engineering from Ferdowsi University of Mashad, Iran, in 2006 and 2008, respectively. Her research interests include image and video processing, machine learning and fuzzy systems. Recently, she professes lectures in Azad University of Mashhad. She has some journal and conference publications in subject of interesting area.

H.R. Pourreza received his BSc in Electronic Engineering from Ferdowsi University of Mashhad, Iran in 1989. He obtained his MSc in Electronic Engineering in 1992 and his PhD in Artificial Intelligence in 2003 from Amirkabir University of Technology, Iran. He is an Associate Professor of Computer Enigineering Department of Ferdowsi University of Mashhad. His research interest includes computer vision, signal processing and Intelligent Transportation System (ITS).

1 Introduction

Video retrieval is one of the important design issues in the development of multimedia information systems, such as digital video library. In recent years, many researchers have dedicated to the study of Content-Based Video Retrieval (CBVR) (Tsutsumi and Nakajima, 2001; Lei and Wu, 2004;

Zampoglou et al., 2007; Fan et al., 2007). Content-based video indexing relies on the processing of a set of features extracted from a video sequence. First, the video sequence is segmented into groups of related frames called ‘shots’ by means of shot detection (Huang and Liao, 2001; Lee et al., 2000). Shot boundary determination has been widely studied for the last decade. Some of the early work can be found in Zhang et al. (1993), Shahraray (1995) and Wang et al. (2000). Further information about shot boundary determination can be found in Liu et al. (2007). The second step to feature extraction is the selection of one or more representative frames from a video shot (known as key frames). The visual contents of these key frames are used to represent the video shots, for indexing and retrieval. Key frame(s) must be able to represent video shot effectively, because this ability influences precision of retrieval, directly. However, the existed techniques of CBVR still suffer from the semantic gap between low-level visual features and high-level semantic visual concepts. To improve retrieval accuracy of CBVR systems and bridging semantic gap, it is necessary that user provide some guidance to the machine. Various techniques exist for such a purpose. A well-known technique is Relevance Feedback (RF), which is widely used in Content-Based Image Retrieval (CBIR). The RF is a powerful tool traditionally used in text-based information retrieval systems. The user is incorporated into the retrieval systems to provide his evaluation on the retrieval results. On the basis of these opinions, the learning mechanism tries to refine the retrieval result in the next iteration. The process iterates until a satisfactory result is obtained for the user. In this paper, two important innovations are proposed in the following fields:

- key frame extraction for video content representation
- RF for interactive retrieval.

Because of applied innovations, different approaches in key frame extraction and ways of applying RF are studied in the two next subsections in the literature.

1.1 Related works on key frame extraction

For effective video browsing and retrieval, the selected key frames should be able to represent the content of the entire video sequence (Sato et al., 1998). There has recently been many works related to the problem of key frame selection and several surveys on the automatic indexing of video data are presented in Sato et al. (1998). After the video is segmented into shots, an easy way is to use the first frame of each shot as the key frame (Nagasaki and Tanaka, 1992). Although simple, the number of key frames for each shot is limited to one, regardless of the shot’s visual complexity. Furthermore, the first frame normally is not stable and does not capture the major visual content. In other proposed approach (Zhang et al., 1997), the current frame of the shot will be compared against the last key frame. If significant content change occurs, the current frame will be selected as a new key frame, note that first frame will be selected as the first key frame. A motion-based approach to key frame extraction is proposed by Wolf (1996). The optical flow for each frame is obtained, first (Horn and Schunck, 1981), then a simple motion metric based on the optical flow is computed, and finally Wolf analyses the metric as a function of time to select key frames at the local minima of motion. Another approach to key frame extraction is a shot-activity-based approach, which is proposed in Gresle and Huang (1997). They first compute an activity indicator. On the basis of the activity curve, the local minima are selected as the key frames.

The first two approaches for key frame extraction are relatively fast. However, they do not effectively capture the visual content of the video shot, since the first frame is not necessarily a key frame. The last two approaches are more sophisticated owing to their analysis of motion and activity. However, they are computationally expensive and their underlying assumption of local minima is not necessarily correct. Also, colour- and motion-based criteria have been employed for key frame selection (Ferman et al., 2002; Lee and Kim, 2002). Although these methods are simple and computationally efficient, they may not provide the most powerful video shot representation.

Representative frame selection in Sze et al. (2005) is based on the probability of occurrence of the pixels at the same position in the frames within a shot. In other words, a shot is represented by a constructed frame, whose value at each pixel position corresponds to that of the pixel with the largest probability of occurrence. This is a time-consuming progress and unsuitable for online application.

In Zhuang et al. (1998), a technique based on clustering for key frame extraction was proposed. If a frame is important, the camera will focus more on this frame. This is the basic assumption in this approach. This method has local view to shot for clustering, and it extracts non-representative frames, for some shots. In this paper, we modify this problem and proposed a clustering-based approach, which is both efficient and effective. It provided more details about these modifications in Subsection 3.1.

1.2 Related works on Relevance Feedback

Many RF methods have been developed in recent years. They either adjust the weights of various features to adapt to the user’s perception (Rui et al., 1997). Another approach estimates the density of the positive feedback samples (Chen et al., 2001). Moreover, discriminate learning has also been used as a feature selection method for RF (Lee and Kim, 2002). These methods work well with certain limitations. The method in Rui et al. (1997) is heuristic. The density estimation method in Chen et al. (2001) loses information contained in negative samples. The discriminate learning in Zhou and Huang (2001) often suffers from the matrix singular problem.

Regarding the positive samples and the negative samples as two different groups and aiming at finding a classifier to identify these two groups from each other, RF in information retrieval systems becomes an online learning problem. In other words, it is a real-time classification problem (Dacheng and Xiaoou, 2006). Recently, classification-based RF has become a popular topic in CBIR. Among classifiers, the SVMs (Hong et al., 2000; Zhang et al., 2001; Tao and Tang, 2004; Tong and Chang, 2001) based RF has shown promising results owing to its good generalisation ability. SVM has a very good performance for pattern classification problems by minimising the Vapnik-Chervonenkis dimensions and achieving a minimal structural risk (Vapnik, 1995). Munesawang and Guan (2005) incorporated a self-learning neural network to implement an automatic RF method. Since neural network models perform effectively when matching given patterns against a large number of possible templates, he can adopt this organisation for similarity matching in video retrieval.

Nonetheless, we have not seen noticeable works on SVM RF in CBVR systems. The difficultly is that RF requires video representation to capture sequential information to allow analysis. In this paper, we use SVM RF in our CBVR system and employ an effective method for video representation to have suitable train data for SVM and high precision in retrieval. Main notes are summarisation of shot based on content, into few suitable frames. Matching of shots with unequal key frames is a problem. Also applying of SVM in RF for semantic video retrieval has not presented yet.

The structure of the paper is organised as follows. Section 2 includes studying of SVM. Section 3 describes the proposed system in detail and experimental results are illustrated in Section 4. Section 5 draws conclusion.

2 Background of Support Vector Machine

SVM is a supervised learning method used for classification and regression. Viewing input data as two sets of vectors in an n-dimensional space, an SVM will construct a separating hyperplane in that space, one that maximises the margin between the two data sets. To calculate the margin, two parallel hyperplanes are constructed, one on each side of the separating hyperplane, which are ‘pushed up against’ the two data sets. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the neighbouring data points of both classes, since in general the larger the margin the lower the generalisation error of the classifier (Vapnik, 1995). We omit the detailed theory and illustrate the algorithm of SVM here. Further information can be found in Vapnik (1995), Gunn (1998) and Wang (2005).

Consider a linearly separable problem:

$$\{(x_i, y_i)\}_{i=1}^{N} \text{ and } y_i = \{+1, -1\} \tag{1}$$

where x_i is an n-dimension vector and y_i is the label of the class that the vector belongs to. SVM separates the two classes of points by a hyper-plane,

$$w^T x + b = 0 \tag{2}$$

where x is an input, w is the weight vector, and b is the bias. SVM finds parameters w and b for the optimal hyperplane to maximise the geometric margin $2/\|w\|$, subject to:

$$y_i(w^T x + b) \geq +1. \tag{3}$$

The solution can be found through a Wolfe dual problem with the Lagrangian multiplied by α_i:

$$Q(\alpha) = \sum_{i=1}^{m} \alpha_i - \sum_{i.j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)/2. \tag{4}$$

Subject to

$$\alpha_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{m} \alpha_i y_i = 0.$$

In the dual format, the data points only appear in the inner product. To get a potentially better representation of the data, the data points are mapped into the Hilbert Inner Product space through a replacement:

$$x_i \cdot x_j \to \phi(x_i) \cdot \phi(x_j) = K(x_i, x_j) \tag{5}$$

where $K(\cdot)$ is a kernel function. With a suitable kernel, SVM can separate in the feature space the data that in the original input space was not separable. There are many kernel functions that can be used, for example:

Gaussian Radial Basis Function (RBF) kernel:

$$K(x_i, x_j) = \mathrm{e}^{-\|x_i - x_j\|^2 \big/ 2\sigma^2}. \tag{6}$$

The polynomial kernel:

$$K(x_i, x_j) = (x_i \cdot x_j + m)^p. \tag{7}$$

We then get the kernel version of the Wolfe dual problem:

$$Q(\alpha) = \sum_{i=1}^{m} \alpha_i - \sum_{i.j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i . x_j)/2. \tag{8}$$

Thus, for a given kernel function, the SVM classifier is:

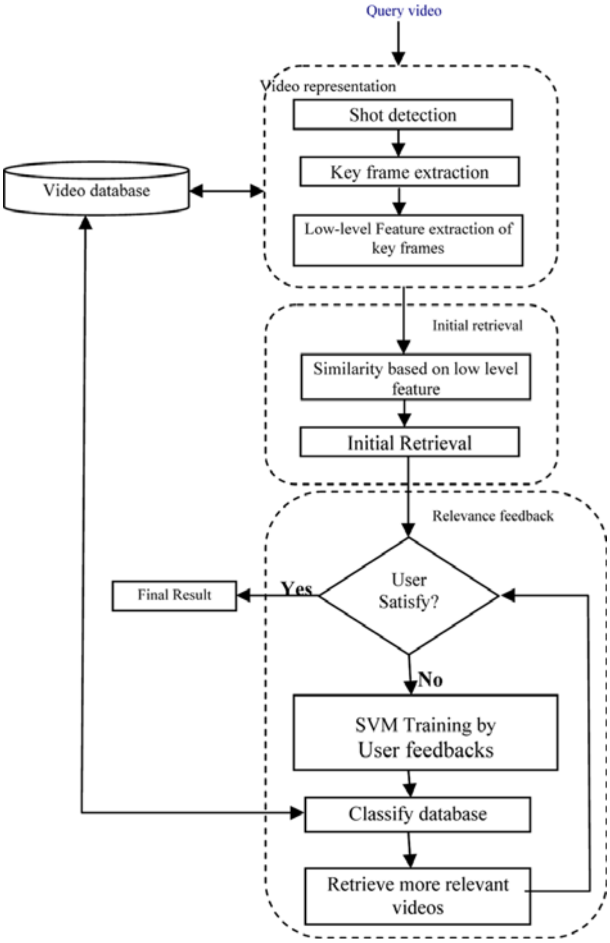
$$F(x) = \operatorname{sgn}(f(x)). \tag{9}$$

$f(x) = \sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b,$ is the output hyperplane decision function of the SVM. In general, when $|f(x)|$ for a given pattern is high, the corresponding prediction confidence will be high. Meanwhile, a low $|f(x)|$ of a given pattern means that the pattern is close to the decision boundary and its corresponding prediction confidence will be low. Consequently, the output of SVM has been used to measure the dissimilarity between a given pattern and the query shot, in SVM-based CBVR RF.

3 The proposed system

The block diagram of proposed system for semantic video retrieval is shown in Figure 1, which consists of three modules namely video representation, initial retrieval and RF module. The following subsections explain the proposed system in detail.

Figure 1 The block diagram of the proposed system (see online version for colours)



3.1 Video content representation

A video sequence is made of several shots, and each shot corresponds to continuous records of a single camera operation, thereupon a shot detection algorithm is applied first. Effective shot boundary extraction existed in literature. In this paper, a new approach for key frame extraction is introduced. Since key-frame extraction directly influences accuracy of retrieval, a hierarchical approach based on clustering is proposed.

First, the colour histogram of whole frames of shot is extracted (i.e., 12 bins Hue colour histogram in the HSV colour space is used). As mentioned in Zhuang et al. (1998), important frames are which ones camera focuses more over them. Therefore, the correlation of Hue is calculated between consecutive frames as follows:

$$\operatorname{corr}(H_k, H_{k-1}) = \frac{\underset{<i>}{\text{——}}(H_k^i - \bar{H}_k)(H_{k-1}^i - \bar{H}_{k-1})}{\underset{<i>}{\text{——}}(H_k^i - \bar{H}_k)^2 \text{——} (H_{k-1}^i - \bar{H}_{k-1})^2} \quad (10)$$

$$\bar{H}_k = \underset{<i>}{\text{——}} H_k^i$$

$$\bar{H}_{k-1} = \underset{<i>}{\text{——}} H_{k-1}^i.$$

H_k and $H_{k–1}$ are colour histogram of frame k and k–1, respectively, and i denotes histogram bins. The value of corr(H_k, $H_{k–1}$) is a number between 0 and 1, whatever about to 0 that means frames k and k – 1 are different.

After that, one or more representative frames from a video shot, known as key frames, are selected. The proposed method to select key frames includes three steps:

a Initially, fuzzy 2-means clustering is applied to eliminate the transient frames from other ones.

b In the second step, linkage clustering is applied to select frames across the more frequent frames. In this step, the key clusters are formed.

c For each key cluster, the frame that is closest to the centroid of cluster is selected as key frame.

Camera motion in each shot gives many frames as edge frames, which have small correlation with its previous and they do not represent the salient content of the shot. Since in the second step of key frame extraction, the proposed approach finds frequent frames, we are encountered with edge frames as key frames wrongly. So, all frames are clustered to two groups' edge frames (uncorrelated frames) and others (correlated frames) by fuzzy 2-means algorithm. Fuzzy 2-means divide frames to correlated and uncorrelated frames. Correlated frames are related frames that camera more focus on them and centroid of this cluster about to 1. After the clusters are formed, an unsupervised clustering is applied on Hue histogram of correlated frames. In this step, the linkage clustering with single-link method (Jain and Dubes, 1988) is applied to cluster the correlated frames. The following subsection explains the unsupervised clustering in detail.

3.1.1 Hierarchical clustering

In cluster analysis, single linkage or nearest neighbour is a method of calculating distances between clusters in hierarchical clustering. In hierarchical method, several mechanisms can be used to obtain the distance of two clusters. One of them is single-link method. In this method, the distance between two clusters is defined as the minimum distance of their samples.

Mathematically, the linkage function, the distance $D(x, y)$ between clusters X and Y, is described by the following expression:

$$D(x, y) = \min(d(x, y)) \tag{11}$$

where $d(x, y)$ is the distance between elements $x \in X$ and $y \in Y$. X and Y are two sets of elements (clusters).

Algorithm

The following algorithm is an agglomerative scheme that erases rows and columns in a proximity matrix as old clusters are merged into new ones. The $N \times N$ proximity matrix D contains all distances $d(i, j)$. The cluster is assigned sequence numbers 0, 1, …, $n-1$ and $L(k)$ is the level of the kth cluster. A cluster with sequence number m is denoted as (m) and the proximity between clusters (r) and (s) is denoted as $d[(r), (s)]$.

The algorithm is composed of the following steps:

a Begin with the disjoint clustering having level $L(0) = 0$ and sequence number $m = 0$.

b Find the least dissimilar pair of clusters in the current clustering, say pair (r), (s), according to $d[(r),(s)] = \min d[(i),(j)]$, where the minimum is overall pairs of clusters in the current clustering.

c Increment the sequence number: $m = m + 1$. Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to $L(m) = d[(r),(s)]$.

d Update the proximity matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The proximity between the new cluster, denoted (r,s), and old cluster (k) is defined as $d[(k),(r,s)] = \min d[(k),(r)], d[(k),(s)]$.

e If all objects are in one cluster, stop. Else, go to step b.

3.1.2 Key frame extraction

Afterwards, the clusters, which are most important, are considered as key clusters, and a key frame(s) is extracted from one(s). A cluster is more important, if its size is bigger than $2N/M$, where N is the number of frames in a shot that was remained from fuzzy 2-means clustering and M is the number of clusters. For each key cluster, the frame that is closest to the centre of cluster is selected as key frame. Key frame is the frame that can represent the salient content of the shot. The proposed approach for key frame extraction extracts representative frames more effectively in comparison with Zhuang et al. (1998), because the proposed approach benefits both local and global view of shot. In the local view, the fuzzy 2-means clustering eliminates transient frames, and remaining frames are correlated with previous frames. Besides, in the global view, there is a linkage clustering of choice frames across the more frequent frames, whereas the proposed method in Zhuang et al. (1998), it is probable that the transient frames form a key cluster and extract a key-frame, which cannot represent salient content of shot, obviously.

The result of proposed method is shown in Figures 2–4. Figure 2 shows some of frames of shot. Frames 1–55 have not been the focus of camera, so they are the underlying shot. The key cluster, which is obtained using unsupervised clustering (proposed approach), is shown in Figure 3, and Figure 4 shows extracted key frame from them. Key frames extracted by Zhuang et al. (1998) have been shown in Figure 5. As shown, two frames 17 and 45 are extracted, whereas camera has no focus on them and they have no roll in shot representation.

Figure 2 Some frames of shot (see online version for colours)

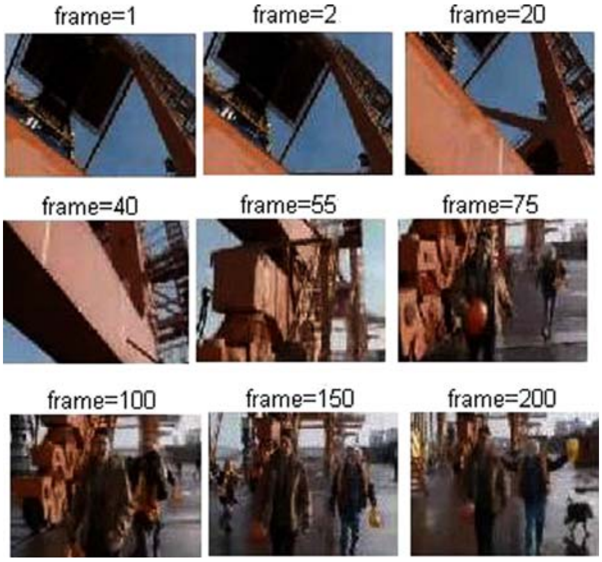


Figure 3 Frames into key cluster (see online version for colours)

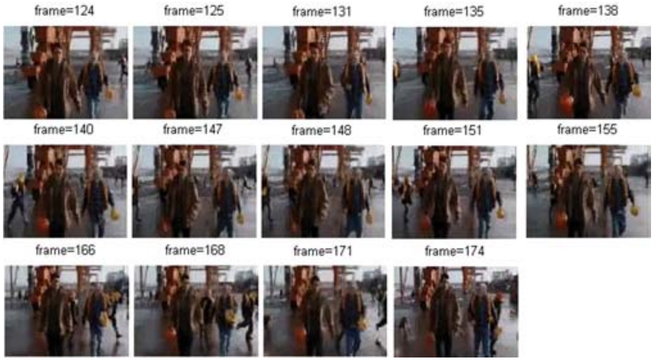
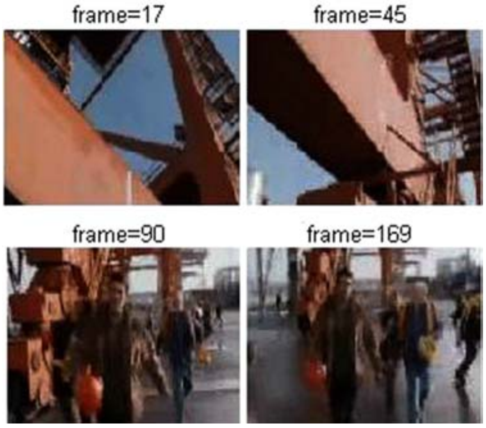


Figure 4 Key frame extracted by proposed method (see online version for colours)

Figure 5 Key frames extracted by proposed method in Zhuang et al. (1998) (see online version for colours)



3.2 Initial retrieval

In this subsection, a distance function is introduced to measure the distance between query shot and each shot in database. Matching of shots with unequal key frames is a problem. To solve this problem, a new distance function is proposed. A new distance function is similar to Hausdorff distance, but instead of maximum operator, summation is used. Twelve bins Hue histogram is extracted from key frame(s) to represent each shot in the feature space. Assume two shot $S_i = \{f_{i1}, f_{i2}, ..., f_{iN}\}$ and $S_j = \left\{\hat{f}_{j1}, \hat{f}_{j2}, \ldots, \hat{f}_{jM}\right\}$ that f is a vector with length of 12 and indicates a key frame in feature space. The new distance function between two shot is defined as follows:

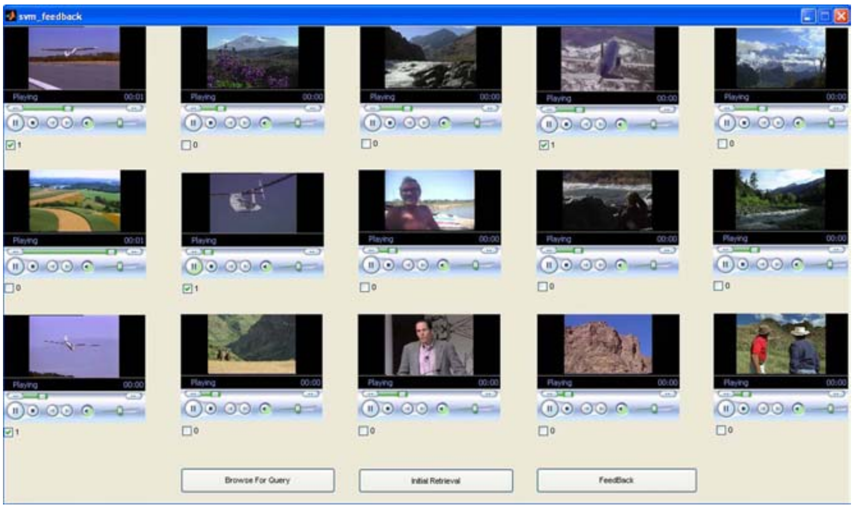
$$\text{Dist}(S_i, S_j) = \frac{\sum_{k=1}^{N} \min\left(\|f_{ik} - \hat{f}_{j1}\|, \|f_{ik} - \hat{f}_{j2}\|, \ldots, \|f_{ik} - \hat{f}_{jM}\|\right)}{|N - M| + 1}. \quad (12)$$

$\| \ \|$ is the Euclidean distance and $|N - M| + 1$ is the normalisation coefficient. If $N = M$, then normalisation coefficient is equal to one otherwise distance normalise based on difference between number of key frames of two shots. According to the distance function, the distance between query shot and each shot in the database is calculated, and sorted in ascending order. Then, L top results are retrieved as initial retrieval, where L is depth of retrieval system. With the new distance function, it is possible to measure similarity between two shots, which have different numbers of key frames.

3.3 Relevance Feedback module

As shown in Figure 1, the RF module consists of three sections: SVM training, database classification and retrieval section. The system first computes the features of the query shot and then returns L top shots ones with the highest similarity scores to the user. The system solicits the user to judge the relevance of the retrieved images. The user provides his evaluation by labelling each displayed image as relevant and irrelevant. Figure 6 shows a scheme of the proposed interactive retrieval system. Features of key frames corresponding to each shot extracted and formed train matrix for SVM training. Each row in train matrix corresponds to a key frame of retrieved shots. For example, if user determines six shots as relevant and nine shots as irrelevant and each shot consists of two key frames, the first 12 rows of training matrix corresponding to 12 relevant key frames and 18 next rows corresponding to 18 irrelevant key frames. After termination of training, the optimal hyperplane is obtained. Now, all key frames in the database can be classified in two groups, relevant and irrelevant key frames. If at least one key frame of shot belongs to relevant group, it is considered as relevant shot. In next step, the shots in relevant group are sorted, based on the distance of hyperplane in descending order and return L top shots of order. Distance between shots and hyperplane is maximum of distance between key frames that belong to each shot and hyperplane, where L is depth of retrieval results. If user satisfies, these retrieved shots are final result. Otherwise, user applies his judgements and run relevance again. The relevant and non-relevant shots will be used to train the SVM for finding the optimal hyperplane to compute the similarity measure. By each user feedback, the number of relevant retrieved shots is increased and provide more positive sample for better training of SVM and so improve precision in the next retrieval results.

Figure 6 Graphic User Interface for proposed system (see online version for colours)



4 Experimental result

In this section, the performance of proposed system for video indexing and retrieval is demonstrated. The proposed system has been tested on general-purpose videos with 800 shots from Trecvid2001 (http://www.open-video.org) and home videos. Video shots database includes Fly of airplanes, Jungles, Rivers, Mountains, Wild Life, Basketball, Roads. They are in AVI format. Depth of retrieval is 15. Figure 7 and Table 1 have shown Average precision (APR) vs. number of user feedbacks.

$$\text{Precision} = \frac{\text{Number of Relevant Retrieved Shots}}{\text{Total Number of Retrieved Shots}}. \quad (13)$$

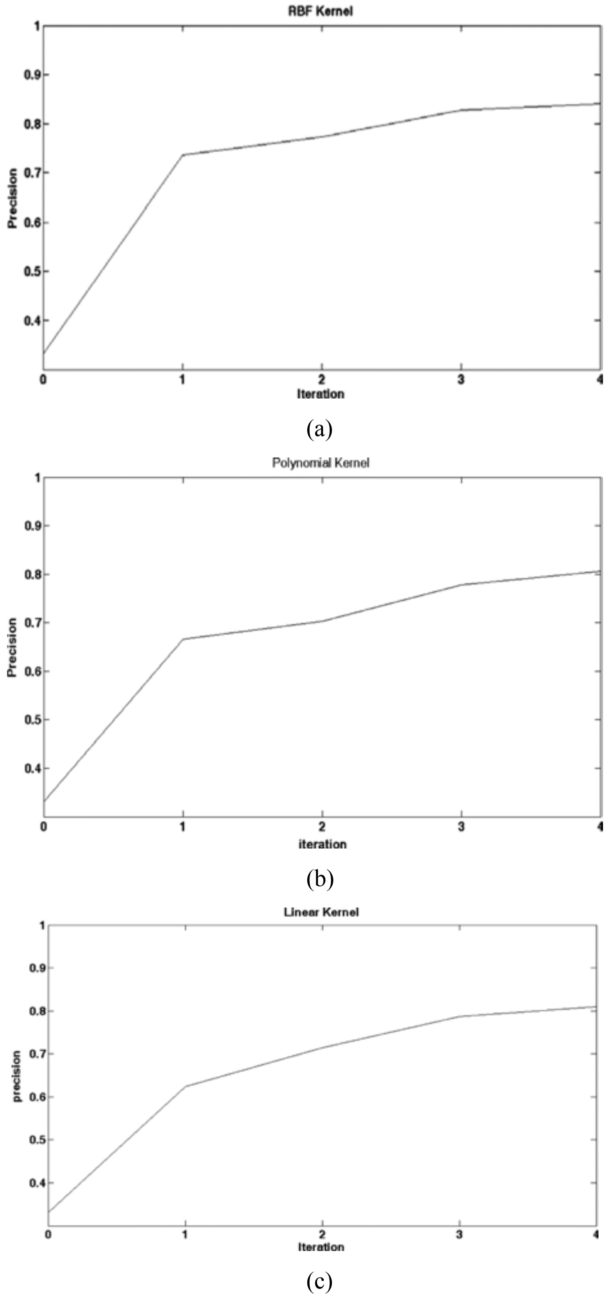
A total of 100 random queries were selected and judgements on the relevance of each shot to each query shot were evaluated. Different kernels for SVM-based learning in RF module are used. Experimental results show that SVM with RBF as kernel has better performance rather than linear or polynomial kernel. Because feature space is non-linear, so there is no distinction boundary between feature related to key frames of relevant and irrelevant shots. The average of precision retrieval in initial iteration (without RF), is 33% and after four user feedback, the proposed system (SVM with RBF) achieves 84% of precision, whereas, after seven iterations the retrieval precision of the ARFN proposed system (Munesawang and Guan, 2005) only reaches 79%. Thus, the proposed approach is able to reach the retrieval goal in only a few iterations. This improvement is preferred in video retrieval since the user aims to retrieve the desired video in as few feedback steps as possible. This precision is better rather than results of ARFN proposed system in Munesawang and Guan (2005). In Munesawang and Guan (2005) for depth of 15 and after 20 user feedback achieve to 79% of precision.

For visual example, a query shot of database is selected and its key frame is shown in Figure 8. This query related to fly of an airplane. Figure 9 shows result of retrieval in initial retrieval. Figures 10–13 have shown result of retrieval in first, second, third and fourth feedback of user, respectively. There are 4 relevant shots in initial retrieval, then number of relevant shots increases to 10 shots, by applying first user feedback. By more user feedbacks, number of relevant shots reached to 12 and 13 in second and third user feedbacks, respectively. By fourth feedback, there is no change in the number of relevant shots and remain in 13 shots. In fact, by user feedback, system learns which colours must be dominant in retrieved shots.

Table 1 Average of retrieval precision for different kernels

	Average precision of retrieval		
	RBF kernel	*Polynomial kernel*	*Linear kernel*
No RF	0.3311	0.3311	0.3311
First RF	0.7364	0.6654	0.6243
Second RF	0.7727	0.7027	0.7147
Third RF	0.8273	0.7781	0.7872
Fourth RF	0.8409	0.8065	0.8094

Figure 7 APR of the proposed system: (a) RBF kernel with sigma = 2; (b) polynomial kernel with $p = 3$ and (c) linear kernel (dot product)



(a)

(b)

(c)

Figure 8 Key frame of selected query (see online version for colours)

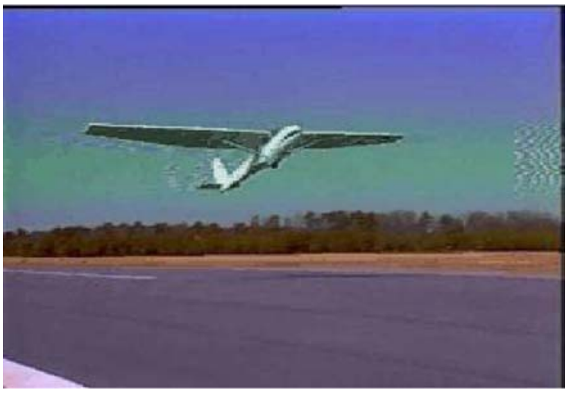


Figure 9 Initial retrieval (see online version for colours)

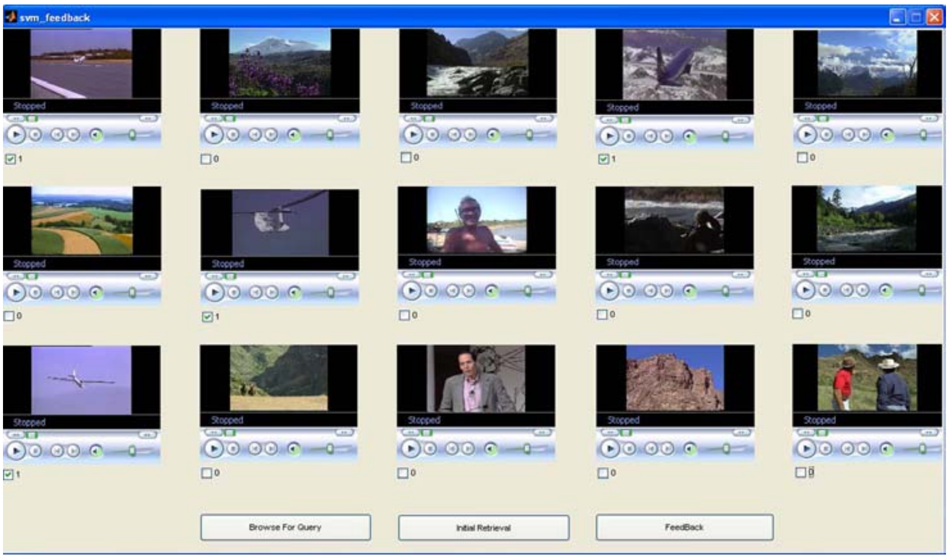


Figure 10 First Relevance Feedback (see online version for colours)

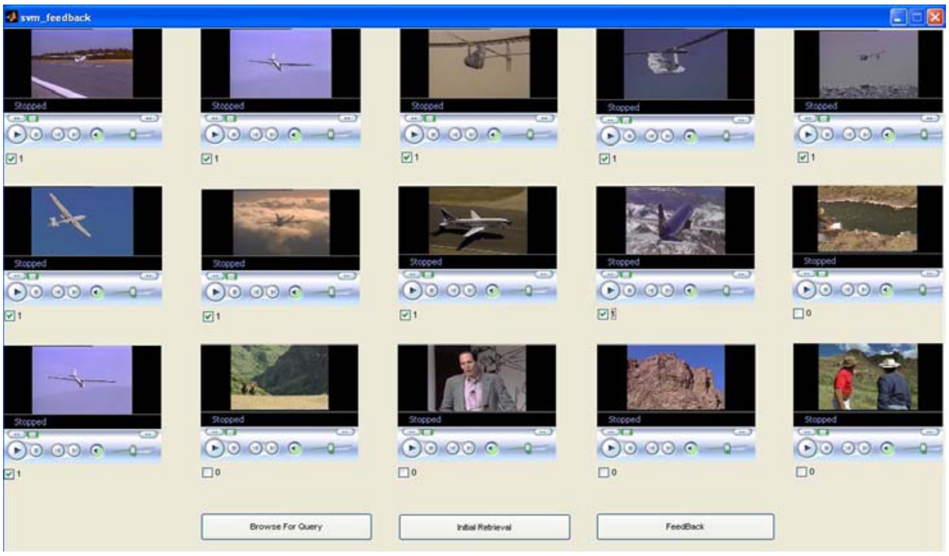


Figure 11 Second Relevance Feedback (see online version for colours)

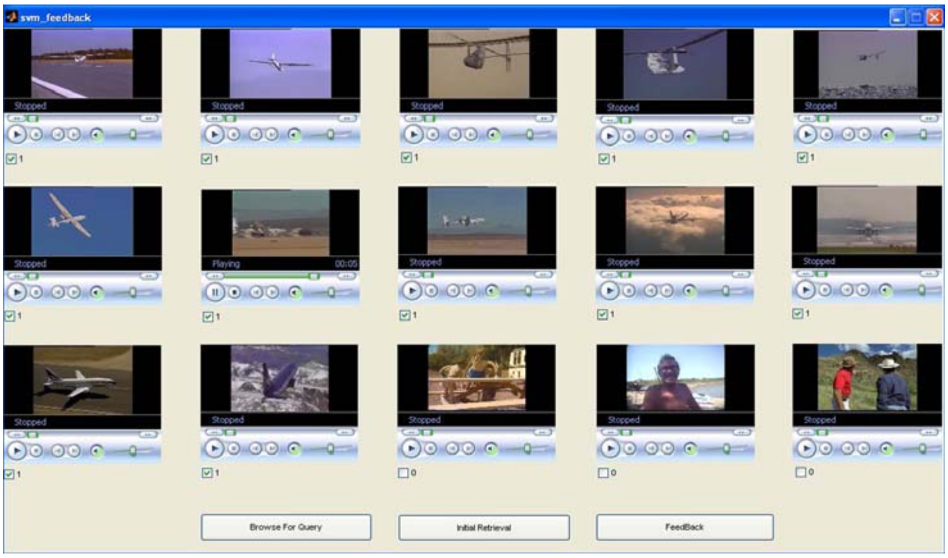


Figure 12 Third Relevance Feedback (see online version for colours)

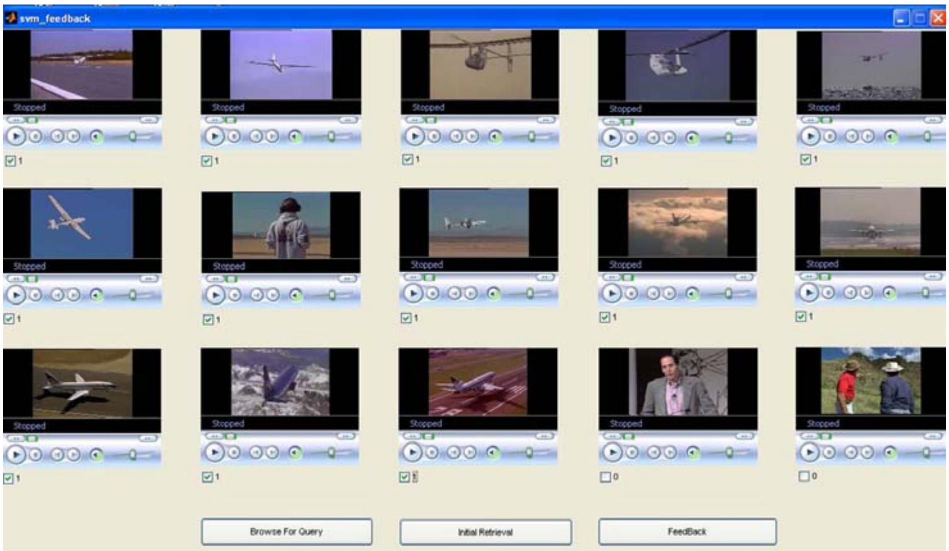
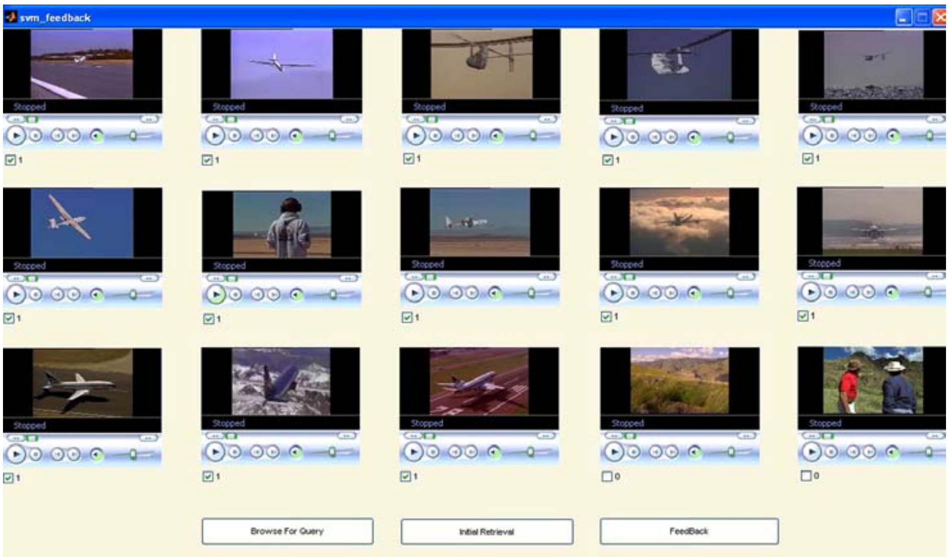


Figure 13 Fourth Relevance Feedback (see online version for colours)



References

Chen, Y., Zhou, X. and Humg, T.S. (2001) 'One-class SVM for learning in image retrieval', *IEEE International Conference on Image Processing 2001*, Thessaloniki, Greece, Vol. 1, pp.34–37.

Dacheng, T. and Xiaoou, T. (2006) 'Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval', *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 28, No. 7, July, pp.1088–1099.

Fan, J., Luo, H., Gao, Y. and Jain, R. (2007) 'Incorporating concept ontology for hierarchical video classification, annotation, and visualization', *IEEE Transactions on Multimedia*, Vol. 9, No. 5, August, pp.939–957.

Ferman, A.M., Tekalp, A.M. and Mehrotra, R. (2002) 'Robust color histogram descriptors for video segment retrieval and identification', *IEEE Trans. Image Process.*, Vol. 11, No. 5, May, pp.497–508.

Gresle, P.O. and Huang, T.S. (1997) 'Gisting of video documents: a key frames selection algorithm using relative activity measure', *The 2nd Int. Conf. on Visual Information Systems*, December, San Diego, California, USA.

Gunn, S.R. (1998) *Support Vector Machines for Classification and Regression*, Technical Report of Faculty of Engineering, Science and Mathematics School of Electronics and Computer Science, 10 May, University of Southampton.

Hong, P., Tian, Q. and Huang, T.S. (2000) 'Incorporate support vector machines to content-based image retrieval with relevant feedback', *IEEE International Conference on Image Processing (ICIP'2000)*, Vancouver, Canada, Vol. 3, September, pp.750–753.

Horn, B.K.P. and Schunck, B.G. (1981) 'Determining optical flow', *Artificial Intelligence*, Vol. 17, pp.185–203.

Huang, C.L. and Liao, B.Y. (2001) 'A robust scene change detection method for video segmentation', *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 11, No. 12, December, pp.1281–1288.

Jain, A.K. and Dubes, R.C. (1988) *Algorithms for Clustering Data*, Prentice-Hall, Michigan State University.

Lee, H.C. and Kim, S.D. (2002) 'Rate driven key frame selection using temporal variation of visual content', *Electron. Lett.*, Vol. 38, No. 5, February, pp.217–218.

Lee, S.W., Kim, Y.M. and Choi, S.W. (2000) 'Fast scene change detection using direct feature extraction from MPEG compressed videos', *IEEE Trans. Multimedia*, Vol. 2, No. 12, December, pp.240–254.

Lei, Z. and Wu, L.D. (2004) 'A new video retrieval approach based on clustering', *Proceedings of the Third International Conference on Machine Learning and Cybernetics*, 26–29 August, Shanghai, Vol. 3, pp.1733–1738.

Liu, Z., Gibbon, D., Zavesky, E., Shahraray, B. and Haffner, P. (2007) 'A fast, comprehensive shot boundary determination system', *IEEE International Conference on Multimedia and Expo (ICME'07)*, July, Beijing, China, pp.1487–1490.

Munesawang, P. and Guan, L. (2005) 'Adaptive video indexing and automatic/semi-automatic relevance feedback', *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 15, No. 8, August, pp.1032–1046.

Nagasaki, A. and Tanaka, Y. (1992) 'Automatic video indexing and full-video search for object appearances', *Proceedings of the IFIP, Second Working Conference on Visual Database Systems*, Vol. 11, pp.113–127.

Rui, Y., Huang, T.S. and Mehrotra, S. (1997) 'Content-based image retrieval with relevance feedback in MARS', *Proc. IEEE Int'l Conf. Image Processing*, Washington DC, USA, Vol. 3, pp.815–818.

Sato, T., Kanade, T., Hughes, E.K. and Smith, M.A. (1998) 'Video OCR for digital news archive', *Proceedings of IEEE International Workshop on Content-Based Access of Image and Video Databases*, 3 January, Bombay, India, pp.52–60.

Shahraray, B. (1995) 'Scene change detection and content-based sampling of video sequences', *Digital Video Compression: Algorithms and Technologies 1995, Proc. SPIE*, San Diego, CA, USA, Vol. 2419, pp.2–13.

Sze, K.W., Lam, K.M. and Qiu, G. (2005) 'A new key frame representation for video segment retrieval', *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 15, No. 9, September, pp.1148–1155.

Tao, D. and Tang, X. (2004) 'Random sampling based SVM for relevance feedback image retrieval', *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, Washington, DC, USA, Vol. 2, pp.647–652.

Tong, S. and Chang, E. (2001) 'Support vector machine active learning for image retrieval', *Proc. ACM Int'l Conf. Multimedia*, Ottawa, Canada, Vol. 9, pp.107–118.

Tsutsumi, F. and Nakajima, C. (2001) 'Hybrid approach of video indexing and machine learning for rapid indexing and highly precise object recognition', *ICIP'2001, IEEE*, Thessaloniki, October, Greece, Vol. 2, pp.645–648.

Vapnik, V. (1995) *The Nature of Statistical Learning Theory*, Springer-Verlag, Berlin.

Wang, L. (2005) *Support Vector Machines: Theory and Applications, Studies in Fuzziness and Soft Computing*, Vol. 177, Springer-Verlag, Berlin, Heidelberg.

Wang, Y., Liu, Z. and Huang, J. (2000) 'Multimedia content analysis using audio and visual information', *IEEE Signal Processing Magazine*, Vol. 17, No. 6, November, pp.12–36.

Wolf, W. (1996) 'Key frame selection by motion analysis', *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, Atlanta, Georgia, Vol. 2, May, pp.1228–1231.

Zampoglou, M., Papadimitriou, Th. and Diamantaras, K.I. (2007) 'Support vector machines content-based video retrieval based solely on motion information', *Proc. 17th Int. Workshop on Machine Learning for Signal Processing (MLSP-2007), IEEE*, August, Thessaloniki, Greece, pp.176–180.

Zhang, H., Wu, J., Zhong, D. and Smoliar, S.W. (1997) 'An integrated system for content based video retrieval and browsing', *Pattern Recognition*, Vol. 30, No. 4, pp.643–658.

Zhang, H.J., Kankanhalli, A. and Smoliar, S.W. (1993) 'Automatic partitioning of full-motion video', *ACM Multimedia System*, Vol. 1, No. 1, pp.10–28.

Zhang, L., Lin, F. and Zhang, B. (2001) 'Support vector machine learning for image retrieval', *Proc. IEEE Int'l Conf. Image Processing*, Thessaloniki, Greece, October, pp.721–724.

Zhou, X. and Huang, T.S. (2001) 'Small sample learning during multimedia retrieval using biasmap', *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, Vol. 1, pp.11–17.

Zhuang, Y., Rui, Y., Huang, T.S. and Mehrotra, S. (1998) 'Adaptive key frame extraction using unsupervised clustering', *Proc. of Int'l Conf. on Image Processing*, Chicago, IL, October, pp.886–870.

Website

Online Available: http://www.open-video.org