Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition Letters 26 (2005) 1221–1231
Pattern Recognition Letters
www.elsevier.com/locate/patrec

A new motion histogram to index motion content in video segments

Haoran Yi, Deepu Rajan *, Liang-Tien Chia

Center for Multimedia and Network Technology, School of Computer Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798, Singapore

Received 17 March 2004; received in revised form 7 September 2004
Available online 15 December 2004

Abstract

A new motion feature for video indexing is proposed in this paper. The motion content of the video at pixel level, is represented as a Pixel Change Ratio Map (PCRM). The PCRM enables us to capture the *intensity* of motion in a video sequence. It also indicates the spatial location and size of the moving object. The proposed motion feature is the motion histogram which is a non-uniformly quantized histogram of the PCRM. We demonstrate the usefulness of the motion histogram with three applications, viz., video retrieval, video clustering and video classification.

Keywords: Motion histogram; Video retrieval; Expectation-maximization; Video classification

1. Introduction

The abundance of multimedia content in various forms like digital libraries and broadcast media calls for efficient techniques for its analysis and management. While initial research activities in this area were directed at image databases, it did not take long to address similar issues like retrieval and classification for video data. These tasks are guided by suitable indexing methods based on the content of the video itself and/or semantic descriptors that could be extracted from the data. Early video indexing methods were based on detecting shot boundaries followed by extracting key frames from which visual features like color, texture, shape, edge etc. were extracted to be used as indices (Smoliar and Zhang, 1994; Deng et al., 2001; Manjunath and Ma, 1996; Rui et al., 1996; Park et al., 2000). However, such indexing techniques do not take into account the essential characteristic of video, viz., its temporal dimension. Recent works have used the spatio-temporal

* Corresponding author. Tel.: +65 6790 4933; fax: +65 6792 6559.
E-mail address: asdrajan@ntu.edu.sg (D. Rajan).

0167-8655/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2004.11.011

relationship among video frames by extracting motion information inherent in them (Ngo et al., 2002; Chang et al., 1998; Fablet et al., 2000; Sahouria and Zakhor, 1999). Ngo et al. (2002) use texture features extracted from temporal slices to index motion content. While patterns in spatio-temporal slices reveal camera motions (pan and zoom) and direction of motion, they do not indicate the *intensity* of motion in a video sequence. Chang et al. (1998) use the trajectory of the moving object as the index to motion content. However, it is very difficult to extract the trajectories of moving objects under complex scene. Fablet et al. (2000) use Markov Random Fields to characterize the optical flow field of video clips. This method is computationally intensive and is, therefore, not suitable for long video clips in large video databases. Sahouria and Zakhor (1999) use block based motion vectors and principal component analysis to represent motion content, but this is done at a very coarse level. The MPEG-7 standard provides motion descriptor to describe motion activity. The extraction of this descriptor is based on aggregate motion vectors contained in the compressed bitstream (MPEG, 2002). However, as in (Sahouria and Zakhor, 1999), motion vectors in MPEG compressed video streams are computed over macroblocks of size 16×16 and hence they are only a coarse representation of the actual motion. It is desirable to consider motion features that can be determined at the pixel level in order to obtain motion information at a finer resolution.

In this paper, we propose a simple and efficient method to extract motion features at the pixel level in order to index video segments on the basis of motion content. Through a simple procedure that uses frame differencing, we create what we call a Pixel Change Ratio Map (PCRM) that indicates moving regions in a particular video sequence. The histogram of each of the PCRMs, with appropriate quantization as explained later, is the proposed new motion histogram. This definition of a motion histogram is different from an intuitive interpretation, which might consider it to be a distribution of motion vectors with respect to their magnitudes and directions (Deng and Manjunath, 1997). However, since our motion histogram is extracted from a PCRM which in turn is representative of the motion in the video sequence at the pixel level, the terminology is justified. We illustrate the utility of the proposed motion histogram through applications in video retrieval, video clustering and video classification.

The paper is organized as follows: in Section 2, we describe the method to compute the PCRM. Section 3 describes the construction of the motion histogram from PCRM. In Section 4, we illustrate the applications of the proposed motion histogram and finally, conclusions are presented in Section 5.

2. Pixel change ratio map

In this section, we present the algorithm to compute the Pixel Change Ratio Map (PCRM), which characterizes motion content in a video segment. We assume that each segment consists of a single shot. However, if the segment consists of more than one shot, a shot detection algorithm can be used to locate the shot boundaries before determining the PCRM for each shot. The algorithm to generate the PCRM is motivated by the fact that the human visual system perceives motion if the intensity of motion is high and the motion continues for a reasonably long duration. By intensity of motion, we mean how fast a certain object moves, implying that a high intensity of motion results in a large change in pixel intensities over the frames. Similarly, object motion that is spread over several frames, i.e., one that has a long duration, is more perceptible than motion over a few frames.

Based on the above observations, we accumulate the changes in pixel intensity over all the frames in a video segment to generate the PCRM for that segment. We initialize the PCRM, which is of the same size as that of the frame, to all zeros. If the frame size is $M \times N$, then the PCRM is simply a matrix of size $M \times N$ initialized to zeros. For the current frame i, we add the absolute values of the frame differences $p_i - p_{i-1}$ and $p_{i+1} - p_i$, i.e., $DI_i = |p_i - p_{i-1}| + |p_{i+1} - p_i|$. For each pixel in this frame, if the sum DI_i is greater than a threshold, the corresponding location in the PCRM is incremented by 1. This procedure is carried out for all the frames in a video segment. The PCRM values are then divided by the number of frames and nor-

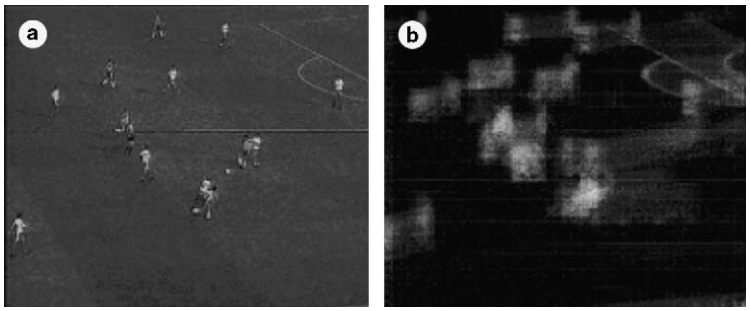

Fig. 1. (a) Key frame (middle frame) of a 'Football' sequence, (b) Pixel Change Ratio Map of the video sequence in (a).

malized to lie in [0, 1]. Thus, it represents the ratio of the number of pixels whose intensities have changed as a result of significant motion, where significant qualifies intensity and duration, as explained earlier. The comparison of DI_i with a threshold is simply to undo the effect of any noise associated with the camera or the decoding process when dealing with compressed video. In this work, the threshold is chosen to be 10 for all the experiments discussed and hence it is not dependent on any particular type of video. Fig. 1(a) shows the 50th frame of a video segment from a soccer sequence consisting of 100 frames. The PCRM for this segment, shown as an intensity image in Fig. 1(b), clearly indicates motion regions. Note that both the local motion (of the players) as well as the global motion (of the camera) is captured in the PCRM. It is the camera motion component of the total motion that causes the lines of the penalty box to show up. The brighter the pixel intensity in the PCRM, higher is the intensity of motion there.

3. The motion histogram

The representation of the PCRM as an intensity image enables us to consider its histogram as a reliable feature to index motion content. However, a histogram with uniform quantization of the bins will not achieve the purpose. This is because most of the values in the PCRM will be cluttered at low values. This is indicated in Fig. 2 which shows the histogram of the PCRM images accumulated for all the 643 video segments in the database that we use in our experiments. Here, the range [0, 1] of the PCRM values are quantized uniformly into 1000 bins. Clearly, in order for the proposed motion histogram to have a high discriminative ability among different classes, it is required to requantize the bins non-uniformly such that the step size is finer where the distribution is high and coarser where the distribution is low. Thus, in Fig. 2, we would like 'narrower' bins to be placed close to the origin while the bins get larger as we move away from it. Once the quantization step sizes are determined for the accumulated histogram, the same is used to requantize the histograms of individual PCRM images, which is indeed the proposed motion histogram for each video sequence.

Suppose there are m quantization levels and it is required to determine the step size for each of them. We obtain the cumulative histogram from the accumulated histogram and traverse it, marking off points at $\frac{1}{m}, \frac{2}{m}, \ldots, \frac{m-1}{m}, 1$, where m is the number of bins. The abscissa of each of these points marks the boundary between two bins. This is a way of simply ensuring that the distribution of the PCRM values among the bins is uniform. Fig. 3(a) shows the histogram of Fig. 1(b) with uniform quantization; this histogram is then subjected to a non-uniform quantization and the result is shown in Fig. 3(b). We see that the histogram obtained from nonuniform quantization has higher bins that are more populated than those in the histogram with uniform quantization. This is desirable since the 'Football' sequence of Fig. 1 has high motion content. We note that although the non-uniform quantization in Fig. 3 is illustrated for a

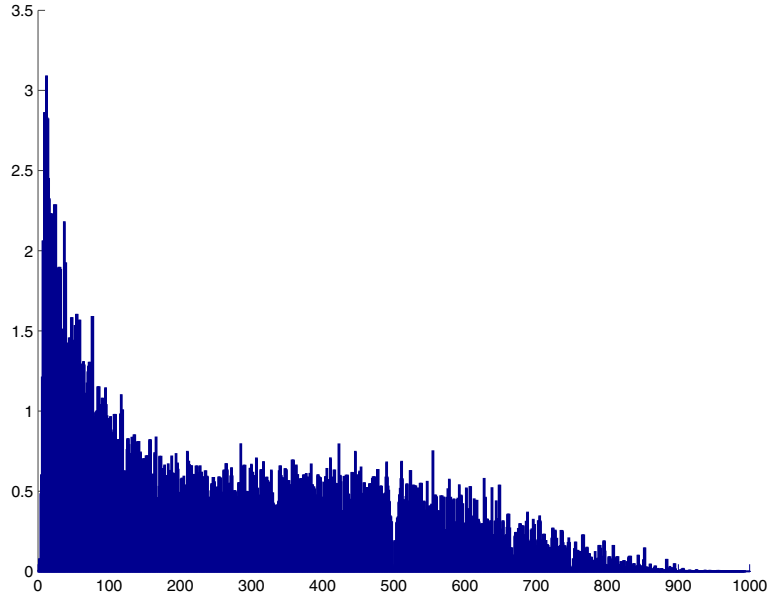

Fig. 2. Histogram of PCRM images accumulated over all the sequences in the database.

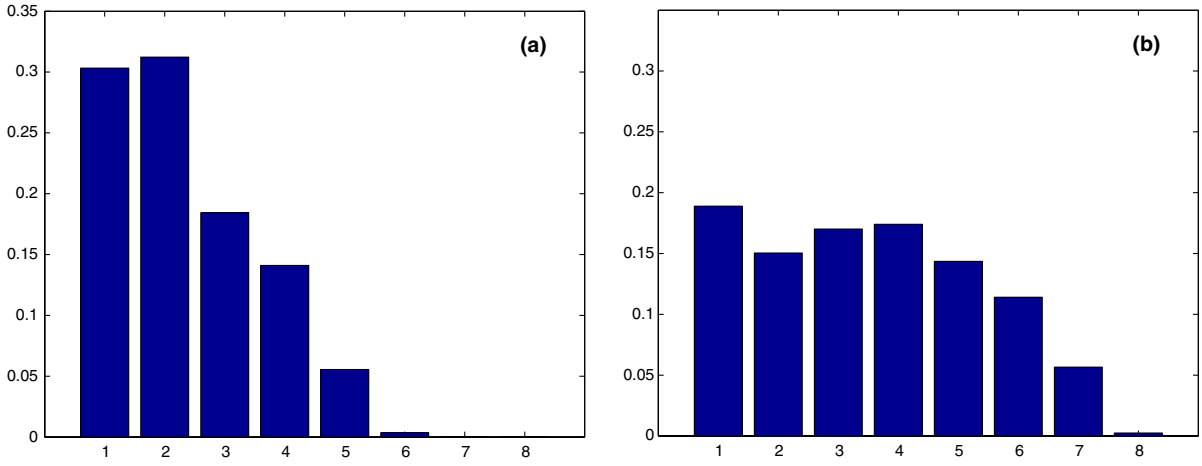

Fig. 3. (a) Histogram of Fig. 1(b) before re-quantization, (b) histogram of Fig. 1(b) after 8-level nonuniform quantization.

single video sequence, the actual requantization is carried out on the histogram obtained from the entire database.

The use of the accumulated histogram to determine the step sizes implies that we use all the available information in the database. However, addition of a new sequence to the database does not require the recomputation of quantization levels since we are already considering a wide variety of sequences whose properties are included in the accumulated histogram. The recomputation needs to be done only in the case where the accumulated histogram is seen to be altered significantly. The determination of the step sizes as

described above can also be justified from an information theoretic point of view wherein we would like the requantized histogram to provide the maximum information about the dataset. Since uniform distribution is the one that maximizes the entropy of a source in the absence of constraints other than those imposed by laws of probability, we would like a uniform distribution of the PCRM values among the bins.

4. Applications of motion histogram

In this section, we illustrate three applications of the new motion histogram proposed in this paper. We utilize the motion histogram for video retrieval, clustering and classification. The video database consists of 643 video sequences taken from the MPEG-7 test set. The videos include sports videos (e.g., soccer, basketball, golf, cycling), news videos, sitcoms, concert videos, documentaries etc. The duration of each video segment varies from 7 to 30 s and there are a total of 299,110 frames. We expect various types of motion content in the different videos and therefore, they form a suitable data set in which to test the proposed motion histogram.

4.1. Video retrieval

The objective here is to retrieve video sequences with motion content similar to that in a query video. The motion content is represented by the motion histogram described above. In this paper, we consider three instances of the number of quantization levels viz., 8, 16 and 32. While the choice of these numbers are merely illustrative, it may be pointed out that they can be represented in binary format with full usage of bits and that they enable the histograms to be feature scalable, i.e, a histogram with a high number of quantization levels can be converted to one with a lower number by simply merging the adjacent bins. Moreover, the MPEG-7 Scalable Color Descriptor which is based on histograms also uses these quantization levels. The Minkowski distance between two histograms h_q and h_d is defined as

$$d(h_q,h_d) = \left(\sum_{m=1}^{N} |h_q(m) - h_d(m)|^u \right)^{1/u}, \quad (1)$$

where N is the number of quantization levels (or bins). In our experiment, we choose Euclidean distances ($u = 2$ in Eq. (1)) to compute the similarity between motion histograms.

Fig. 4 shows examples of video retrieval experiments based on the motion histogram. The first column consists of the key frames of the query video shots. In this case, we simply consider the middle frame in the sequence as the key frame. The second, third and fourth columns show the key frames of the first, second and third retrieved video sequences. It is evident that sequences belonging to the same category are present among the top returns for different types of videos. We also observe that the motion histogram is able to distinguish between large and small moving objects. For example, the soccer sequence (first row in Fig. 4) consists of moving regions that are small while the sequence containing walking people (second row in Fig. 4) consists of moving regions that are large.

To quantify the performance of the motion histogram for video retrieval, we consider the precision of retrieval. However, first, we would like to emphasize on the difficulty of obtaining the ground truth when motion histogram is used as a feature for retrieval. Intuitively, it would seem that the ground truth could be obtained by categorizing the database according to motion content, e.g., low, mid and high motion sequences. But this is a very subjective task and it is not easy to define these categories exactly. The alternative is to categorize the database on the basis of semantics in which case construction of the ground truth is not hard. But the problem here is that semantics does not always relate to motion content in a direct manner, e.g. in a golf sequence, the swing of the golfer is high motion while the ball slowly rolling towards the hole is low motion. Even though 'golf' is the semantic class, the motion is not coherent throughout. However, we observe that in soccer, basketball and talking head videos, the motion content is coherent in that there is significant motion throughout the sequence, e.g., in soccer

Fig. 4. Examples of video retrieval showing the key frame of each sequence.

sequence, even when the ball goes out of play and there is not much motion among the players, the camera will be zooming into the area from where the ball is thrown in. Thus, we calculate precision of retrieval on two semantic classes—the first one is called 'soccer & basketball' and the second one is called 'talking head'. Note that soccer and basketball sequences are categorized into one class—we could as well consider them as separate classes. Fig. 5(a) shows the average precision for different values of top n retrievals for 'soccer & basketball'. The average precision for a particular n is obtained by querying the database with each video sequence belonging to this class and averaging the precision over all the queries. The 32-level motion histogram performs better than the 8-level and the 16-level histograms because the variation in pixel intensities in the PCRM image for these sequences is very large and hence, a finer quantization is required. Note that our measure of precision is stricter than the one based only on semantic ground truth because the retrieved video sequences are considered relevant only when they are not only similar in motion content to the query video sequence, but also share the same semantics. Fig. 5(b) shows the average precision for top n retrievals for the 'talking head' class of videos. Here, the 8-level histogram performs better since the sequences in this class contain very low motion resulting in the intensity values of the PCRM image to be concentrated near zero. Since the variation of the pixel values in the PCRM is very small, the 32 level motion histogram may over quantize the PCRM values. The lower average precision in this case compared to 'soccer & basketball' is a consequence of our definition of precision. In the database, there are many sequences from sitcoms which have similar motion content to talking head videos, e.g. conversation between two people. Although they are similar in motion content, they are considered as false posi-

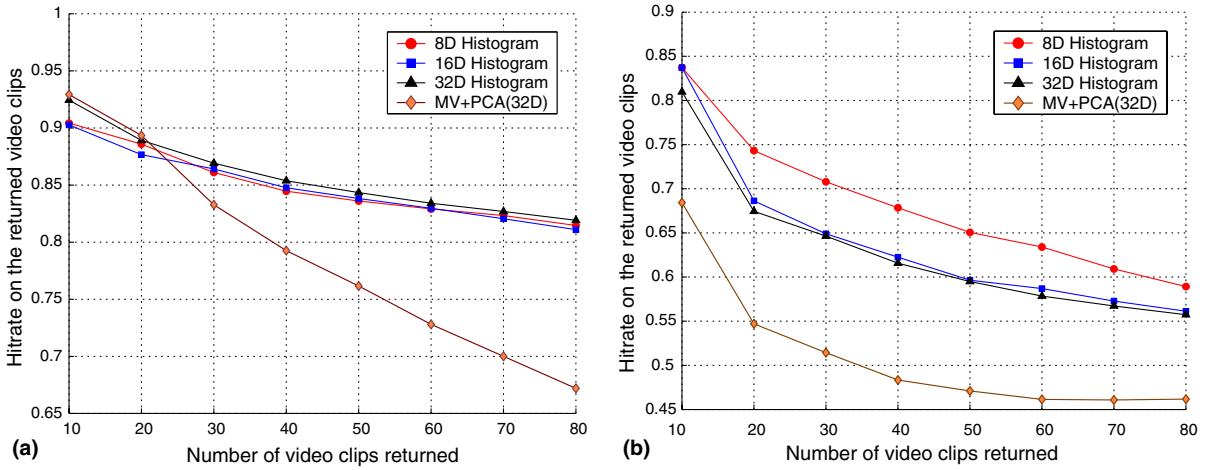

Fig. 5. Average precision of retrieval for (a) 'soccer & basketball' and (b) 'talking head' sequences.

tives since they do not belong to the same semantic class. We also compare the efficiency of the proposed feature with motion vectors used in (Sahouria and Zakhor, 1999); the precision and recall for a 32 dimensional motion feature are illustrated by the MV + PCA curves in Fig. 5(a) and 5(b). Clearly, the feature proposed in this paper is superior.

4.2. Video clustering

We wish to group together video sequences with similar motion content using the proposed motion histogram. However, the clusters thus formed should also be semantically meaningful. If so, the clusters can be initialized to be the starting point for content browsing by the user. We assume that the video sequences in the database are generated by a Gaussian Mixture Model (GMM). Since each sequence is represented by the motion histogram, its N dimensional feature vector $\boldsymbol{x}$ (recall that N is the number of bins in the motion histogram) is denoted by

$$f(\boldsymbol{x}|\theta) = \sum_{k=1}^{M} \alpha_k f_k(\boldsymbol{x}|\theta_k), \tag{2}$$

where M is the number of mixtures, α is the mixing co-efficient and θ is the parameter set. The assumption of Gaussian mixture models allows us to write $f_k(\boldsymbol{x}|\theta_k)$ as

$$f_k(\boldsymbol{x}|\theta_k) = \frac{1}{\sqrt{(2\pi)^N |\Sigma_k|}} \times \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\mu_k)^{\mathrm{T}} \Sigma_k^{-1} (\boldsymbol{x}-\mu_k)\right\}. \tag{3}$$

The model is specified in terms of the parameter set $\theta_k = \{\alpha_k, \mu_k, \Sigma_k\}$, for $k = 1$ to M, where $\sum_{k=1}^{M} \alpha_k = 1$ and Σ_k is a $N \times N$ positive definite covariance matrix. The clustering problem is then the estimation of these parameters followed by determining those video sequences that have the most similar sets of parameters. Here, M is the number of clusters. A maximum likelihood (ML) approach is followed to carry out this task. If there are Q observations of the random vector $\boldsymbol{x}$, then the ML estimate of the parameters, θ_{ML_k}, is obtained by maximizing the likelihood function $f(\boldsymbol{x}|\theta_k)$, i.e.,

$$\theta_{\mathrm{ML}_k} = \arg\max_{\theta_k} f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_Q|\theta_k). \tag{4}$$

Here, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_Q$ are the feature vectors corresponding to motion histograms of the Q video sequences in the database. We use the Expectation–Maximization (EM) algorithm (Dempster et al., 1977) to estimate the parameter θ_k.

It is important that the EM algorithm is initialized with parameters so that it does not get trapped in a local maxima of the likelihood

function. Therefore, we utilize the K-means algorithm (Duda and Hart, 1973) to perform data-driven initialization instead of a random one. In the experiment, we perform K-means clustering 5 times (here the initialization is random). The K-means clustering result with the smallest sum of squared distance (SSD) is chosen as the initialization for the EM algorithm. An optimal criterion for the number of clusters, M, is based on a trade-off between the performance and number of parameters used for describing the mixture distribution. We choose the Bayes information criterion (BIC) (Burnham et al., 1994) to determine the optimum number of clusters,

$$\mathrm{BIC} = -2 \times \log \mathrm{Lik} + q \times \log L, \tag{5}$$

where $\log \mathrm{Lik}$ is the log-likelihood evaluated at the maximum likelihood estimates of the model parameters, and q is the number of parameters in the model and L is the sample size. Fig. 6(a) shows the BIC for different number of clusters. We see that 7 is the optimal number of clusters as chosen by the BIC.

Although we would like to show the discriminatory capability of the motion histogram visually, it is not possible to plot the feature vectors because of their high dimensionality. However, we use the Principal Components Analysis to reduce the dimension of the feature vectors (from 8, 16 or 32) to 2 and plot them in Fig. 6(b). We observe that the features are separated into two clusters. One of the clusters represents 'high' motion and the other represents 'low' motion as perceived by the human visual system. Next, we apply the EM based clustering algorithm on the 643 video sequences in the database to cluster them into 7 groups. Fig. 7 shows the key frames from 3 representative video sequences for each of the clusters. It is interesting to note that each cluster represents a particular type of motion. For example, cluster 2 represents sequences in which the camera is stationary while there is little object motion. Such sequences include news videos, interviews etc. However, in cluster 3, there is indeed a smooth camera motion although motion of objects is very little or non-existent, for example as in the camera pan of a landscape. The descriptions of motion in the video sequences in each cluster and their examples are listed as follows:

Cluster 1. Irregular camera motion, large object motion with small moving object size (e.g., soccer (long shot), basketball (long shot), etc.)

Cluster 2. Still camera and little object motion (e.g., talking head, dialogue, interview, etc.)

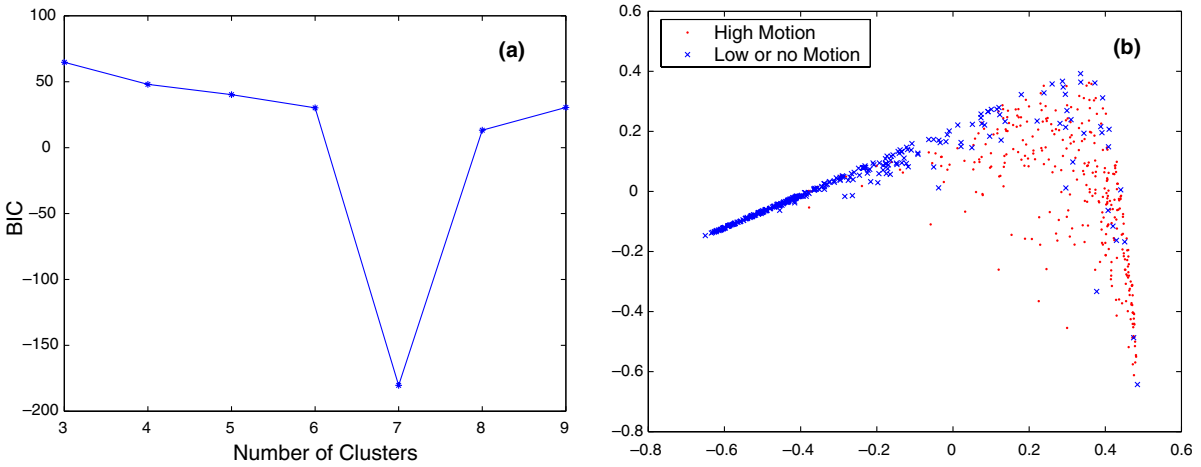

Fig. 6. (a) Bayes Information Criterion (BIC) for different number of clusters and (b) scatter plot of feature vectors reduced from 8 dimensions to 2 dimensions using PCA.

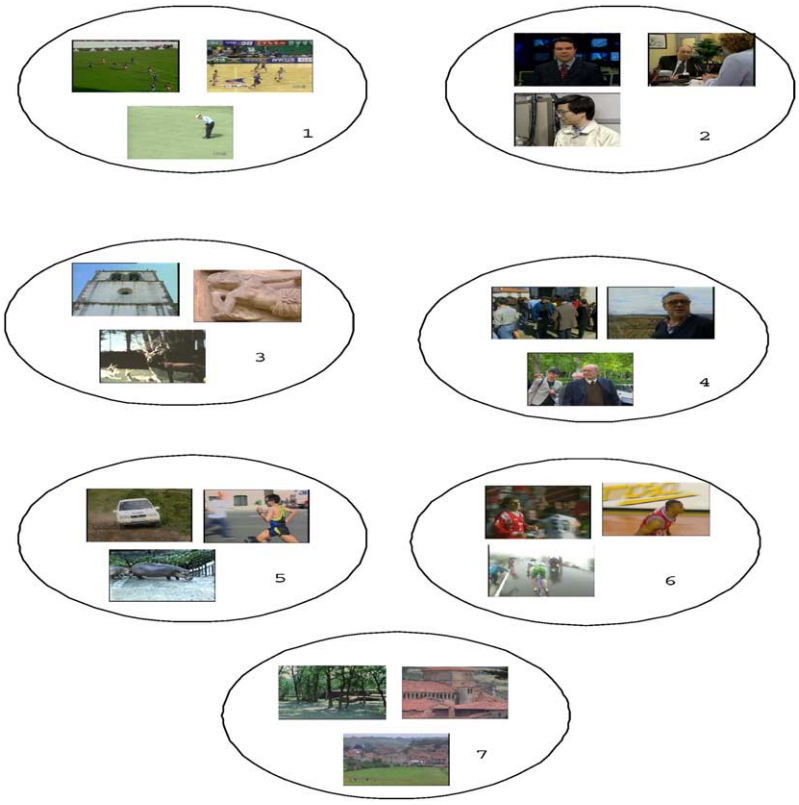

Fig. 7. Video clustering using 8-level motion histogram (the key frames are shown).

Cluster 3. Smooth camera motion with little or no object motion (e.g., scenery, etc.)
Cluster 4. Little camera motion and little object motion (e.g., outdoor interview, out door news shots, etc.)
Cluster 5. Little camera motion and large object motion (e.g., marathon, racing cars, animal shows, etc.)
Cluster 6. Irregular camera motion and large moving object(e.g., soccer(close-up), basketball(close-up), cycling, etc.)
Cluster 7. Little or no motion (scenery, etc.).

4.3. Video classification

In this section, we classify video sequences into predefined categories according to their motion content. We choose two categories—'high motion' and 'low motion'. As before, we use the 643 sequences in the database and label them as belonging to one of the two classes. There are 337 sequences labelled as 'high motion' and the rest 306 are labelled as 'low motion'. Fig. 8(a) shows some of the representative video sequences from the two classes; the top row shows sequences containing high motion and the bottom row those containing low motion. We use Support Vector Machines (SVM) for classification because of their discriminating power and simplicity. In this case, the SVM performs a binary classification. Radial basis functions are chosen as the kernel for training. For details about SVM (see Christianini and Shawe-Taylor, 2000).

In order to train the SVM, we randomly pick n samples from the database. We consider the effect of n on the classification rate by performing the training/classification 30 times for each n and then

Fig. 8. (a) Examples of video sequences (key frames) from two classes—'high motion' (top row) and 'low motion' (bottom row), (b) average classification rate with different number of training samples.

finding the average of the classification rate. The performance of the classifier for various values of n are shown in Fig. 8(b). We note that the classifier needs very less training samples to converge. In our experiment, only 20 samples are enough to train the SVM, which is just 3% of the entire data. The small size of the training samples indicates that these two classes are quite separated by the proposed motion histogram. The average classification rate for 8-, 16- and 32-level motion histograms are above 86%. Considering that labelling of the video clips into 'high motion' and 'low motion' is a highly subjective task, the high classification rate is encouraging. Comparing the PCRM based feature with the features proposed in (Sahouria and Zakhor, 1999), we find the former outperforms that latter by a significant margin as shown in Fig. 8(b).

5. Conclusions

In this paper, we have introduced a new method for motion indexing. We propose the formation of a *Pixel Change Ratio Map* based on motion content in a video sequence and extract the motion histogram from it. The bins in the motion histogram are adaptively quantized so as to have a high discriminating power among different classes. We illustrate the efficacy of the proposed motion histogram through applications in video retrieval, clustering and classification. Not only does it retrieve sequences having similar motion as the query sequence, but it is also able to provide an indication of the size of the moving objects. The clusters formed by using motion histogram as the feature are very similar in motion content to what is perceived by the human visual system. Classification of video sequences using the proposed motion histogram results in a high classification rate. Moreover, there is need for only a few training samples indicating that it has a high discriminating ability. We plan to investigate other features that can be extracted from the PCRM which could enable description of textures.

References

Burnham, K.P., Anderson, D.R., White, G.C., 1994. Evaluation of the Kullback–Leibler discrepancy for model selection in open population capture-recapture models. Biometrical J. 36, 299–315.

Chang, S.F., Chen, W., Meng, H.J., Sundaram, H., Zhong, D., 1998. A fully automatic content-based video search engine supporting multiple object spatio-temporal queries. IEEE Trans. Circuit Syst. Video Technol. 8 (Sep.), 602–615.

Christianini, N., Shawe-Taylor, J., 2000. An introduction to Suppor Vector Machines. Cambridge University Press.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Royal Statist. Soc. 39 (1), 1–38.

Deng, Y., Manjunath, B.S., 1997. Content-based search of video using color, texture and motion. IEEE Int. Conf. Image Process. 2, 534–537.

Deng, Y., Manjunath, B.S., Kenney, C., Moore, M.S., Shin, H., 2001. An efficient color representation for image retrieval. IEEE Trans. Image Process. 10 (1), 140–147.

Duda, R.O., Hart, P.E., 1973. Pattern Classification and Scene Analysis. John Wiley and Sons Inc., New York, USA.

Fablet, R., Bouthemy, P., Perez, P., 2000. Statistical motion-based video indexing and retrieval. In: International Conference on Content Based Multimedia Information Access, pp. 602–619.

Manjunath, B.S., Ma, W.Y., 1996. Texture features for browsing and retrieval of image data. IEEE Trans. Pattern Anal. Machine Intell. 18 (8), 837–842.

MPEG, 2002. Multimedia content description interface-part 8: extraction and use of MPEG-7 descriptors. ISO/IEC 15938-8:2002.

Ngo, C.-W., Pong, T.-C., Zhang, H.-J., 2002. On clustering and retrieval of video shots through temporal slices analysis. IEEE Trans. Multimedia 4 (4), 446–458.

Park, D.K., Jeon, Y.S., Won, C.S., 2000. Efficient use of local edge histogram descriptor. In: Proceedings of the ACM workshops on Multimedia. Los Angeles, California, USA, pp. 51–54.

Rui, Y., She, A.C., Huang, T.S., 1996. Modified fourier descriptors for shape representation a practical approach. In: Proceedings of First International Workshop on Image Databases and Multimedia search.

Sahouria, E., Zakhor, A., 1999. Content analysis of video using principle components. IEEE Trans. Circuit Syst. Video Technol. 9, 1290–1298.

Smoliar, S., Zhang, H., 1994. Content-based video indexing and retrieval. IEEE Trans. Multimedia 1, 62–72.