SIViP (2010) 4:61–73
DOI 10.1007/s11760-008-0097-9

ORIGINAL PAPER

Classification of sport videos using edge-based features and autoassociative neural network models

C. Krishna Mohan · B. Yegnanarayana

Received: 30 November 2006 / Revised: 10 May 2008 / Accepted: 15 November 2008 / Published online: 10 December 2008

Abstract In this paper, we propose a method for classification of sport videos using edge-based features, namely edge direction histogram and edge intensity histogram. We demonstrate that these features provide discriminative information useful for classification of sport videos, by considering five sports categories, namely, cricket, football, tennis, basketball and volleyball. The ability of autoassociative neural network (AANN) models to capture the distribution of feature vectors is exploited, to develop class-specific models using edge-based features. We show that combining evidence from complementary edge features results in improved classification performance. Also, combination of evidence from different classifiers like AANN, hidden Markov model (HMM) and support vector machine (SVM) helps improve the classification performance. Finally, the performance of the classification system is examined for test videos which do not belong to any of the above five categories. A low rate of misclassification error for these test videos validates the effectiveness of edge-based features and AANN models for video classification.

Keywords Video classification · Edge-based features · Autoassociative neural network model · Hidden Markov model · Support vector machines

C. Krishna Mohan (✉)
Indian Institute of Technology Madras,
Chennai 600036, Tamil Nadu, India
e-mail: ckm@cs.iitm.ernet.in

B. Yegnanarayana
International Institute of Information Technology,
Hyderabad 500032, Andhra Pradesh, India
e-mail: yegna@iiit.ac.in

1 Introduction

Classification of digital videos into various genres, or categories is an important task, and enables efficient cataloging and retrieval with large video collections. The objective of video classification is to classify a given video clip into one of the predefined video categories. Many approaches have been proposed for content-based classification of video data. The problem of content-based classification of video can be addressed at different levels in the semantic hierarchy. For instance, video collections can be categorized into different program genres such as news, commercials and sports. Then, videos of a particular genre, such as sports, can be further classified into sub-categories like soccer, hockey, cricket, etc. A video sequence of a given sub-category can then be partitioned into smaller segments, and these segments can be classified into semantically meaningful classes.

In this paper, we address the problem of sport videos classification for five classes, namely, cricket, football, tennis, basketball and volleyball. Sports videos represent an important application domain due to their commercial appeal. Classification of sports video data is a challenging problem, mainly due to the similarity between different sports in terms of entities such as playing field, players and audience. Also, there exists significant variation in the video of a given category collected from different TV programs/channels. This intra-class variability contributes to the difficulty of classification of sports videos.

Content-based video classification is essentially a pattern classification problem [1] in which there are two basic issues, namely, feature extraction and classification based on the selected features. Feature extraction is the process of extracting descriptive parameters from the video, which will be useful in discriminating between classes of video. The classifier operates in two phases: training and testing phase.

Training is the process of familiarizing the system with the video characteristics of a given category, and testing is the actual classification task, where a test video clip is assigned a class label.

Several audio–visual features have been described for characterizing semantic content in multimedia [2]. The general approach to video classification involves extraction of visual features based on color, shape, and motion, followed by estimation of class-specific probability density function of the feature vectors [3,4]. A criterion based on the total length of edges in a given frame is used in [5]. The edges are computed by transforming each block of 8×8 pixels using discrete cosine transform (DCT), and then processing the DCT coefficients. A rule-based decision is then applied to classify each frame into one of the predefined semantic categories. Another edge-based feature, namely, the percentage of edge pixels, is extracted from each keyframe for classifying a given sports video into one of the five categories, namely, badminton, soccer, basketball, tennis, and figure skating [6]. The k-nearest neighbor algorithm was used for classification. Motion is another important feature for representation of video sequences. A feature, called motion texture, is derived from motion field between video frames, either in optical flow field or in motion vector field in [7]. These features are employed in conjunction with support vector machines (SVMs) to devise a set of multicategory classifiers.

The approach described in [8] defines local measurements of motion, whose spatio-temporal distributions are modeled using statistical nonparametric modeling. To exploit the strong correlation between the camera motion and the actions taken in sports, sports videos are categorized on the basis of camera motion parameters [9]. The camera motion patterns such as fix, pan, zoom, and shake are extracted from the video data. Motion dynamics such as foreground object motion and background camera motion are extracted in [10] for classification of a video sequence into three broad categories, namely, sports, cartoons and news. Transform coefficients derived from DCT and Hadamard transform of image frames are reduced in dimension using principal component analysis (PCA) [11]. The probability density function of the compressed features is then modeled using a mixture of Gaussian densities. Dimension reduction of low-level features such as color and texture, using PCA, has also been attempted in [12,13]. Another approach, described in [14], constructs two hidden Markov models (HMMs), one from the principal motion direction, and the other from the principal color of each frame. The decisions from both the models are combined to obtain the final score for classification. Apart from the above statistical models, rule-based methods have also been applied for classification. A decision tree method is used in [15] to classify videos into different genres. For this purpose, several attributes are derived from the video sequences, such as length of the video clip, number of shots, average shot length and percentage of cuts. A set of decision rules is derived using these attributes.

Edges constitute an important feature to represent the content of an image. Human visual system is sensitive to edge-specific features for image perception. In sports video classification, images that contain the playing field are significant for distinguishing among the classes of sports. This is because, each sport has its own distinct playing field, where most of the action takes place. Also, the interaction among subjects (players, referees and audience) and objects (ball, goal, basket) is unique to each sport. A few sample images of each sports category are shown in Fig. 1. The corresponding edge images are shown in Fig. 2. Each playing field has several distinguishing features such as lines present on the playing field, and regions of different textures. The subjects are also prominent in the images, thus helping in distinguishing different sports. From Fig. 2, we can observe that edge features are important for representing the content of sports video, and also these features carry sufficient information for discriminating among classes. These observations suggest that features derived to represent the edge information can be of significant help for discriminating various categories of sports.

In this paper, we propose to make use of two edge-based features which provide complementary information for classification of sports videos. We exploit the capability of auto-associative neural network models to capture the distribution of the feature vectors. Two other classifier methodologies, namely, HMMs and support vector machines (SVMs), are employed for their ability to capture the sequence information and discriminative learning, respectively. Evidences from these classifiers are combined to improve the performance of video classification.

The paper is organized as follows: In Sect. 2, edge direction histogram (EDH) and edge intensity histogram (EIH) are extracted for representing the visual features inherent in a video class. Section 3 gives a brief introduction to the classifier methodologies used for video classification. The section also describes a method to combine the evidences from multiple classifiers. Section 4 describes experiments for classification of videos of the five sports categories, and discusses the performance of the system. Section 5 summarizes the study.

2 Extraction of edge-based features

We consider two features to represent the edge information, namely, EDH and EIH. Edge direction histogram is one of the standard visual descriptors defined in MPEG-7 for image and video, and provides a good representation of nonhomogeneous textured images [16]. This descriptor captures the spatial distribution of edges. Our approach to compute the EDH is a modified version of the approach given in

Fig. 1 Sample images from five different sports video categories: **a** Basketball, **b** cricket, **c** football, **d** tennis and **e** volleyball

[16]. A given image is first segmented into four sub-images. The edge information is then calculated for each sub-image using Canny algorithm [17]. The range of the edge directions (0°–180°) is quantized into 5 bins. Thus, an image partitioned into four sub-images results in a 20-dimensional EDH feature vector for each frame of a video clip. The choice of partitioning an image into four sub-images and quantization of edge directions into 5 bins are found to be appropriate based on experimental evidence. Figure 3 shows 20-dimensional EDHs for five different categories. Each histogram is obtained by averaging the histograms obtained from individual frames of a clip. The clips were selected randomly from five different classes. The figure shows that the pattern of EDH is different for different classes, and that the selected features carry discriminative information among the different video classes.

We also consider the distribution of the edge intensities to evaluate the degree of uniformity of the edge pixels. This feature is derived from the magnitude information of the edge pixels. The range of magnitudes (0–255) is quantized into 16 bins, and a 16-dimensional EIH is derived from each frame of a video clip. Figure 4 shows the 16-dimensional EIH for five different categories. Each histogram is obtained by averaging the histograms obtained from individual frames of a

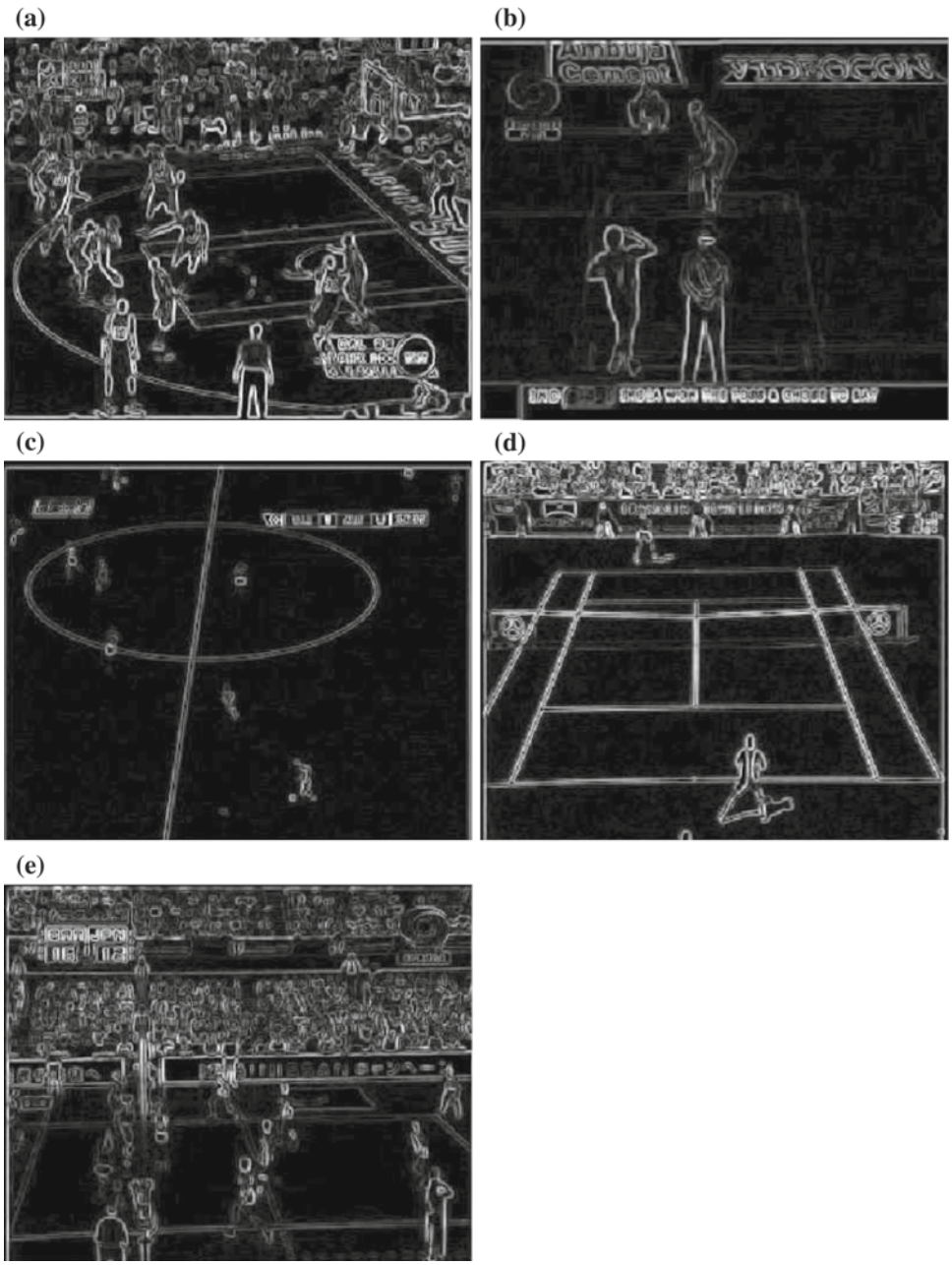

Fig. 2 Edge images corresponding to the five different images shown in Fig. 1, for the following categories: **a** Basketball, **b** cricket, **c** football, **d** tennis and **e** volleyball

clip. The clips were selected randomly from the five different classes. From Figs. 3 and 4, we observe that EDH carries more discriminative information among the classes than EIH.

3 Classifier methodologies

Once features are extracted, the next step is to model the behavior of the features for performing classification. We consider three classifier methodologies for our study, namely, autoassociative neural networks (AANN), HMMs, and SVMs. The AANNs are useful to model the video content, due to their ability to capture the distribution of the feature vectors [18]. Given the temporal nature of the video, HMMs are effective for modeling the time-varying patterns [19]. Support vector machines are useful for their inherent discriminative learning ability and good generalization performance [20]. In the following subsections, a brief introduction to the three classifier methodologies is presented.

3.1 AANN models for estimating the density of feature vectors

Autoassociative neural network models are feedforward neural networks, performing an identity mapping of the input space [21,22]. From a different perspective, AANN models

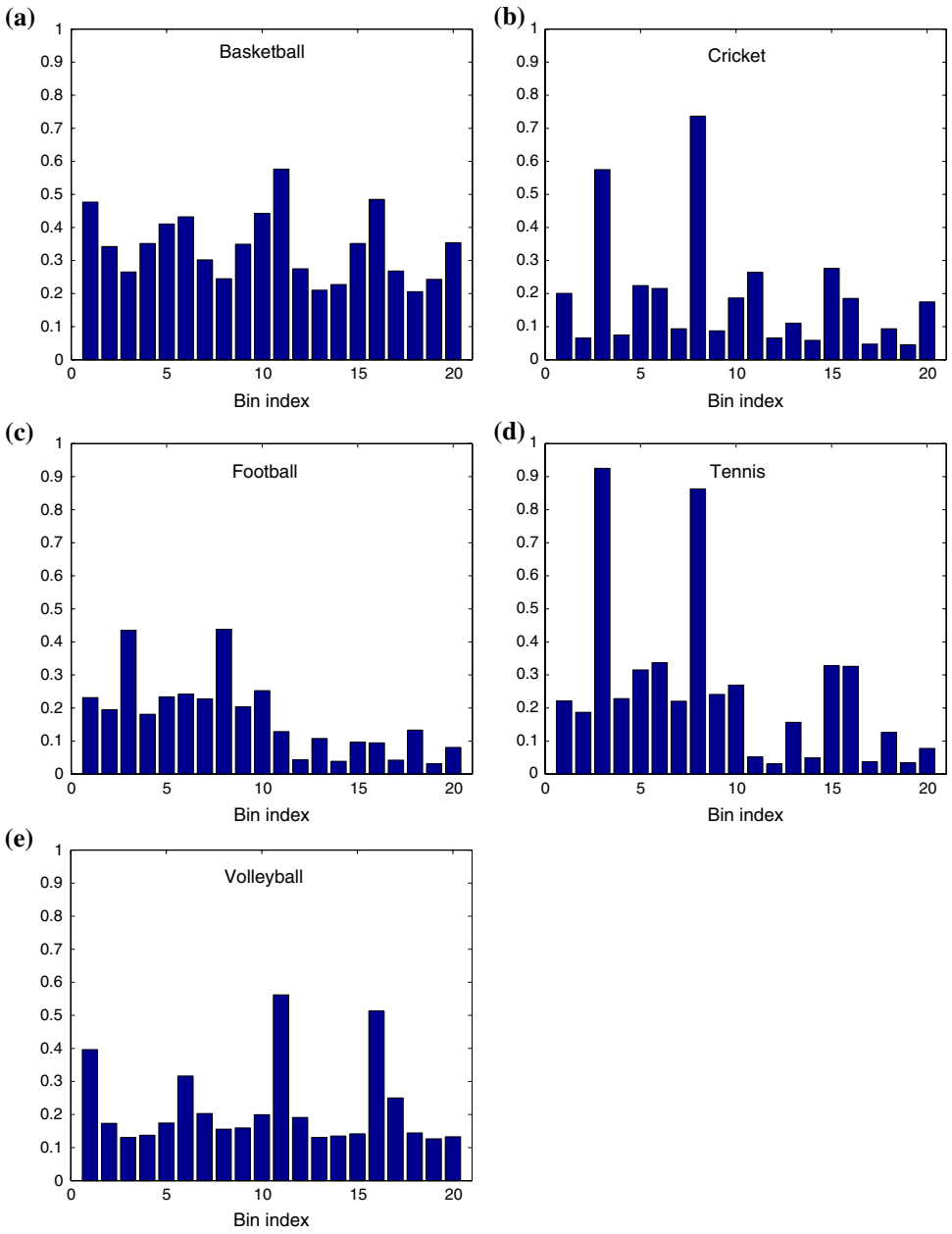

Fig. 3 Average edge direction histogram feature vectors of 20 dimension for sample clips selected randomly from the five different classes: **a** Basketball, **b** cricket, **c** football, **d** tennis and **e** volleyball

can be used to capture the distribution of input data [18]. The distribution capturing ability of the AANN models is discussed in detail in Appendix. In this study, separate AANN models are used to capture the distribution of feature vectors of each sports video category. A five-layer AANN model is shown in Fig. 5. The structure of the AANN model used in the present studies is 20*L* 40*N* 6*N* 40*N* 20*L*, where *L* denotes linear units and *N* denotes nonlinear units. This structure is arrived at experimentally. The activation function of the nonlinear unit is a hyperbolic tangent function. The network is trained using error backpropagation learning algorithm for 500 epochs [21]. One epoch denotes the presentation of all the training examples (of a given class) to the neural network exactly once. The number of epochs is chosen using cross-validation for verification, to obtain the best performance for the experimental data.

The block diagram of the proposed sports video classification system based on EDH is shown in Fig. 6. For each video category, a separate AANN model is developed. The model giving the strongest evidence for a given test clip is hypothesized as the category of the test clip. Similar classification system is developed for features based on EIH. The EDH and EIH feature vectors corresponding to each sports category are used to train two separate AANN models for each feature type. The AANN models are trained using backpropagation learning algorithm in the pattern mode [21,22]. The learning

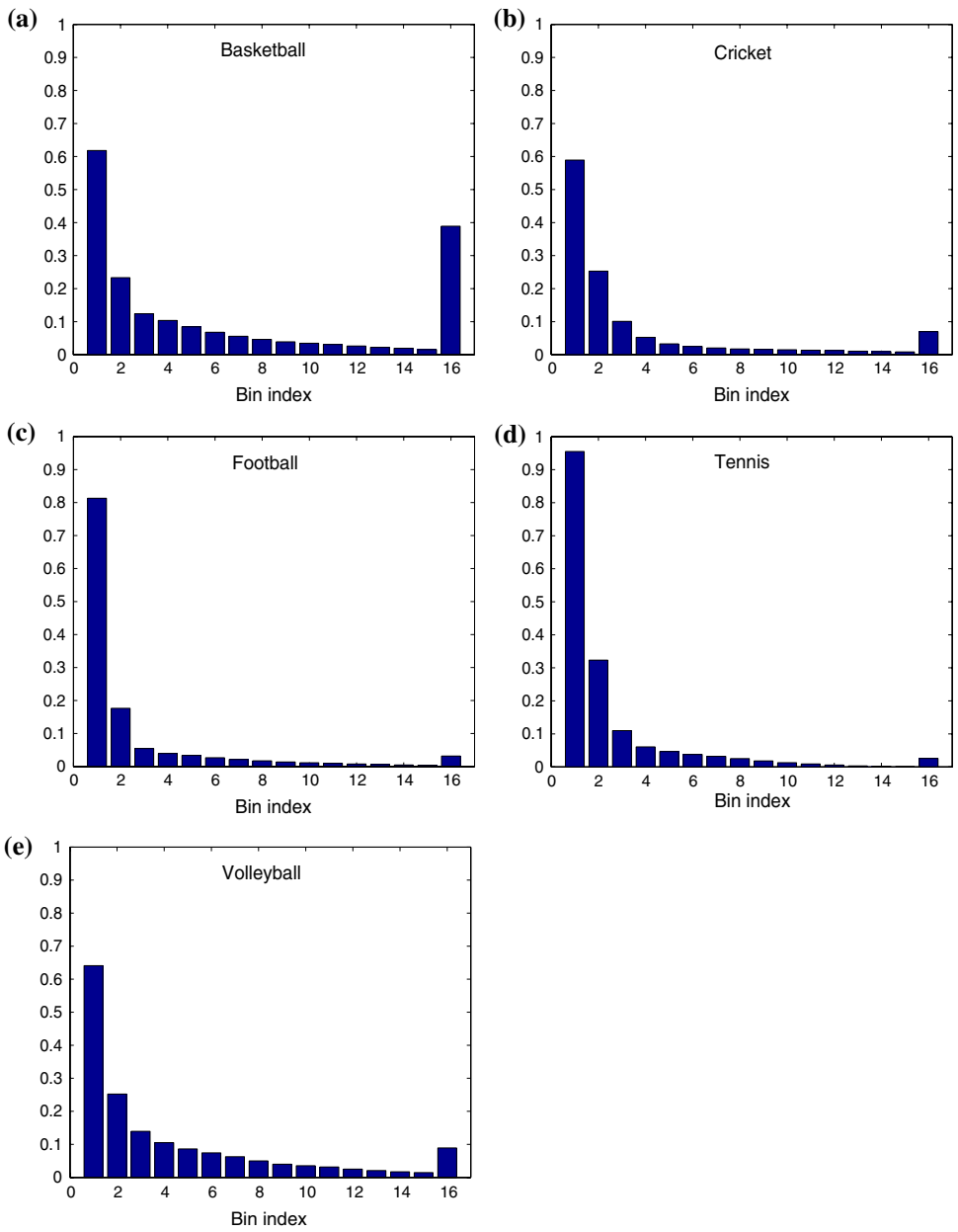

Fig. 4 Average edge intensity histogram feature vectors of 16 dimension for sample clips selected randomly from the five different classes: **a** Basketball, **b** cricket, **c** football, **d** tennis and **e** volleyball

algorithm adjusts weights of the network to minimize the mean squared error obtained for each feature vector.

A test video clip is processed to extract the EDH and EIH features. These features are presented as input to the AANN models of all the categories. The output of each model is compared with its input to calculate the squared error for each frame. The error E_k for the kth frame is transformed into a confidence value by using the relation $C_k = \exp(-E_k)$. For a given test clip, the confidence value is given by $C = \frac{1}{N}\sum_{k=1}^{N} C_k$, where N is the total number of frames in the test clip. For each category, two confidence values are obtained, one for each feature type. These two scores are combined using linear weighted average rule to obtain a combined score $\hat{C}$ given by

$$\hat{C} = w \times C_d + (1 - w) \times C_i, \tag{1}$$

where C_d and C_i denote the confidence scores for EDH and EIH features, respectively. The value of w ($0 \leq w \leq 1$) is chosen to maximize the classification performance for the given data set. Thus, for each test video clip, five scores are obtained. The category whose model gives the highest confidence value is hypothesized as the sports category of the test clip. Experimental results are discussed in Sect. 4.

3.2 Hidden Markov models

The hidden Markov model consists of finite number (N) of states. The state of the system at each time step is updated

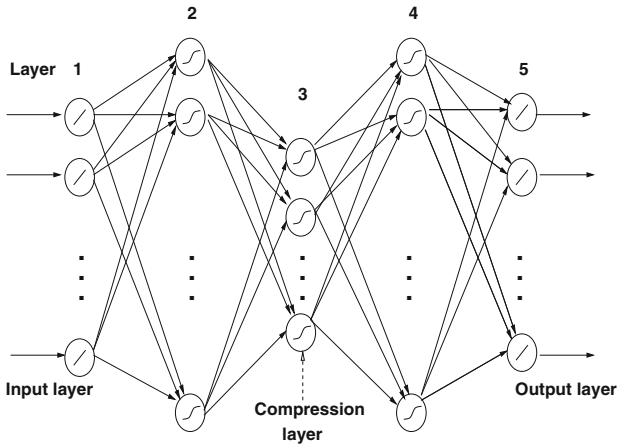

Fig. 5 Structure of five-layer AANN model used for video classification

according to a probability distribution that depends only on the previous state. Additionally, a symbol is generated in each state according to a probability distribution that depends on that state. The parameters of the HMM are adjusted using the training data set [23]. Given a HMM model λ and an observation sequence $\mathbf{O}$, the probability $P(\mathbf{O}/\lambda)$ that this observation sequence comes from the model λ is calculated as a sum over all possible state sequences. The hidden Markov model toolkit (HTK) was used for developing class-specific models [24]. The number of states ($N = 7$) and number of mixtures ($M = 1$) per state are chosen experimentally to obtain the best classification performance. During testing phase, the HMM gives the log probability, representing the likelihood that the given clip belongs to that particular class. The test methodology is similar to the block schematic shown in Fig. 6. Experimental results are discussed in Sect. 4.

3.3 Support vector machines for video classification

Support vector machines provide a new approach to the pattern classification problems with underlying basis in statistical learning theory, in particular the principle of structural risk minimization [25]. The SVM models learn to separate the boundary regions between patterns belonging to two classes by mapping the input patterns onto a high-dimensional space, and seeking a separating hyperplane in this space. The separating hyperplane is chosen in such a way as to maximize its distance (margin) from the closest training examples. We consider SVM models for classification due to their ability to generalize from limited amount of training data, and also due to their inherent discriminative learning [22]. The SVM-Torch-II [26] was used for developing class-specific SVM models. When a given feature vector corresponding to a test clip is presented to an SVM model, the result is a measure of the distance of the feature vector from the hyperplane constructed as a decision boundary between a given class and the remaining classes.

The performance of the pattern classification problem depends on the type of kernel function chosen. Possible choices of kernel function include polynomial, Gaussian and sigmoidal functions. In this work, we have used Gaussian kernel, since it was empirically observed to perform better than the other two. This class of SVMs involves two parameters, namely, the kernel width σ and the penalty parameter P. In our experiments, the value of σ represents the dynamic range of the features. The value of P was chosen

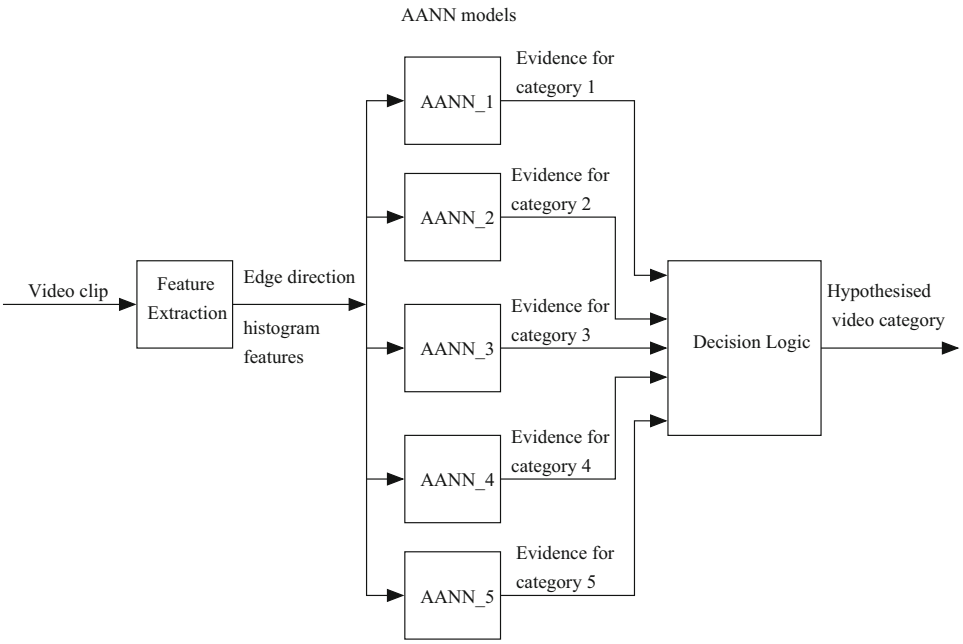

Fig. 6 Block diagram of the proposed video classification system using edge direction histogram features. Categories 1–5 are cricket, football, tennis, basketball and volleyball respectively

corresponding to the best classification performance. The SVMs were originally designed for two-class classification problems. In our work, multi-class ($M = 5$) classification task is achieved using one-against-rest approach, where an SVM is constructed for each class by discriminating that class against the remaining ($M - 1$) classes. The test methodology is similar to the block schematic shown in Fig. 6. Experimental results are discussed in Sect. 4.

3.4 Combining evidence due to multiple classifiers

It has been shown in the literature [27–30] that combination of evidence obtained from several complementary classifiers can improve the performance of classification. The reasons for combining evidence from multiple classifiers/features are as follows: (a) For a specific pattern recognition application, each classifier methodology can attain only a certain degree of success, but combining evidence from different methodologies can reinforce a positive decision and minimize the incidence of misclassification. (b) It is hard to lump different features together to design one single classifier, due to the curse of dimensionality. (c) Combining evidence from different features which provide complementary information about a given class may help in improving classification.

There are numerous types of features that can be extracted from the same raw data. Based on each of these features, a classifier or several different classifiers can be trained for the same classification task. As a result, we need to combine the results from these classifiers to produce an improved result for the classification task. The output information from a classifier reflects the degree of confidence that the specific input belongs to the given class. First, the evidence due to two different features, namely, the EDH and EIH, are combined using the rule of linear weighting, as described in Eq. 1. At the next level, evidence obtained from different classifiers are combined using linear weighting. The outcome of such a combination of evidences is discussed in the next section.

4 Results and discussion

4.1 Data set

Experiments are carried out on about 5 h and 30 min of video data (1,000 video clips, 200 clips per sports category, and each clip of 20 s duration) comprising of cricket, football, tennis, basketball and volleyball video categories. The video clips were captured at the rate of 25 frames/s, at 320 × 240 pixel resolution. The data were collected from different TV channels in different sessions to capture the variability due to sessions. For each sports video category, 100 clips are used for training, and the remaining 100 clips are used for testing.

4.2 Performance of different classifiers

The performance of the AANN based classification system using EDH, EIH, and combined evidence from EDH and EIH is shown in Table 1. The performances of the classification systems based on HMMs and SVMs are given in Tables 2 and 3, respectively. From the results, it can be observed that the classification performance is poorer for video clips of cricket and football categories, compared to those of tennis, basketball and volleyball categories. This is because, in the latter three categories, the playing fields have well-defined lines, and they appear in a majority of frames of a video clip. Moreover, a few well-defined camera views dominate the broadcast. For example, such views may cover the full court in tennis or volleyball. Thus, a large area of an image frame comprises the playing field. On the other hand, in cricket and football categories the camera view tends to change from one position to another depending on the action. Thus, continuous motion along with lack of well-manifested edge-specific information results in poorer classification. It is also evident that the edge direction is a stronger feature for discriminating between the classes, compared to the edge intensity. This may be due to the fact that one can visually perceive the content of an image from its binary edge image, which preserves only the edge directions but not the magnitudes. Performance of

Table 1 Performance of AANN based sports video classification system using EDH, EIH, and combined evidence (correct classification in %)

	Cricket	Football	Tennis	Basketball	Volleyball	Average performance
EDH	81	84	95	94	95	89.8
EIH	54	57	93	93	92	77.8
Combined	84	88	100	100	100	94.4

Table 2 Performance of HMM based sports video classification system using EDH, EIH, and combined evidence (correct classification in %)

	Cricket	Football	Tennis	Basketball	Volleyball	Average performance
EDH	77	86	92	95	94	88.8
EIH	45	58	84	93	92	74.4
Combined	80	87	93	98	96	90.8

Table 3 Performance of SVM based classification system using EDH, EIH, and combined evidence (correct classification in %)

	Cricket	Football	Tennis	Basketball	Volleyball	Average performance
EDH	81	84	92	93	95	89.0
EIH	68	86	32	89	92	73.4
Combined	83	86	100	100	100	93.8

Table 4 Classification performance obtained by combining evidence from different classifiers (correct classification in %)

	Cricket	Football	Tennis	Basketball	Volleyball	Average performance
AANN	84	88	100	100	100	94.0
SVM	83	86	100	100	100	93.8
HMM	80	87	93	98	96	90.8
AANN + SVM	96	94	100	100	100	98.0
AANN + HMM	92	92	100	100	100	96.8
HMM + SVM	90	92	100	100	100	96.4
AANN + HMM + SVM	96	94	100	100	100	98.0

the SVM based classifier is particularly poor for EIH features compared to AANN and HMM-based classifiers for the same feature. This is due to lack of discriminative information in EIH, and also due to the fact that the SVMs are chosen for their discriminative ability. Since edge direction and edge intensity features can be viewed as complementary sources of information, the evidence due to these features can be combined. Tables 1, 2, and 3 also show the performance of classification obtained using weighted combination of evidences using EDH and EIH from different classifiers. There is an improvement in the performance of classification due to the combination of evidences, from all the classifiers.

4.3 Effect of duration and quality of test video data

The duration of the test data (test video clip) has significant bearing on the classification performance. Several techniques for video classification typically use test clips with durations varying from 60 to 180 s [6,5,9,13,14,31]. The classification performance was found to improve with increase in the duration of the test clip. The average edge ratio used in conjunction with k-nearest neighbor algorithm requires 120 s of test data to yield a classification performance of 92.4% on a five-class problem [6]. The AANN based classifier has better generalizing ability than the k-nearest neighbor classifier. Similarly, a time-constrained clustering algorithm [13] using compressed color features requires a minimum of 50 s of test data to yield a classification performance comparable to the proposed method. The proposed method was applied on test clips of 20 s duration in all the experiments on video classification. The performance given in Tables 1, 2 and 3 is comparable to the result obtained using larger duration of test clips. Apart from the duration of the test data, the quality of the test data also influences the classification performance. Some methods retain only the class-specific frames in the test data by editing out images related to crowd/audience or off-field action [13]. Such editing results in an improved performance. In our experiments, no such editing of the test data was done.

4.4 Performance on combining evidence from multiple classifiers

The normalized measurement values obtained from the three classifiers are combined using linear weighting. Table 4 shows classification performance obtained by combining evidence from different combinations of the three classifiers. It is observed that the combination of evidences from any two classifiers results in a better performance than those of the individual classifiers. The confusion matrix for the final classifier (combined AANN, HMM, and SVM) is given in Table 5. The improvement in classification due to combination of evidence can be attributed to the merits in different classifier methodologies, which emphasize different types

Table 5 Confusion matrix of video classification results (in %) corresponding to the score obtained by combining evidence due to all the three classifiers (in %) (AANN, HMM, and SVM)

	Cricket	Football	Tennis	Basketball	Volleyball
Cricket	96	00	04	00	00
Football	02	94	04	00	00
Tennis	00	00	100	00	00
Basketball	00	00	00	100	00
Volleyball	00	00	00	00	100

of information present in the features, such as their spatial distribution and temporal sequence.

4.5 Verification of test video sequences using the classifiers

It is necessary to examine the response of a classifier for test inputs of a different class. More specifically, if a test video clip belongs to a class other than the above five classes, the system is not expected to assign the label of any of the five classes. Instead, the system should assign a separate label to all such test cases. This, however, depends on two factors: (a) the nature of evidence/measurement output by a classifier, and (b) the decision logic for assigning a class label to a test video clip. In SVM-based classifiers, one-against-rest approach is used for decomposition of multi-class pattern classification problem into several two-class pattern classification problems. Hence, one should ideally get all negative confidence scores as output of the SVM model for a test clip which does not belong to any of the predefined categories. Thus, a natural threshold of 0 helps in decision making in the case of SVM, although such a decision could also be in error.

In the case of AANN models and HMMs, the training process attempts to capture only the within-class properties, and no specific attempt is made to distinguish a given class from others. Thus, a nonclass test input to these models still results in positive measurements, although small. Figure 7 shows the histogram of in-class confidence scores along with that of nonclass confidence scores, for AANN models, SVMs and HMMs. The scores are normalized between 0 and 1. The in-class scores are obtained by presenting the test video clips of a given category to the models of the same category. The nonclass scores are obtained by presenting the test video clips of a given category to the models of other categories. A total of 100 test video clips of each class are used to obtain the in-class and nonclass confidence scores. The extent of separation of the histograms indicates the ability of the model to discriminate between in-class and nonclass examples. The area of overlap of the two histograms is a measure of the minimum classification error. From Fig. 7, we observe that this area of overlap is least for SVM-based classifier, followed by AANN-based classifier. If the confidence score corresponding to the intersection of the two histograms is chosen as threshold for decision, then such a choice results in minimum classification error on the training data. The same threshold is used for decision in the case of test data. Tables 6, 7, and 8 indicate the outcome of presenting test video clips of cartoon, commercial and news categories, to the models based on AANN, SVM, and HMM, respectively, trained on cricket, football, tennis, basketball, and volleyball. The entries in the tables denote the percentage of misclassification. For instance, if a test video clip of cartoon category, when presented to the model of cricket

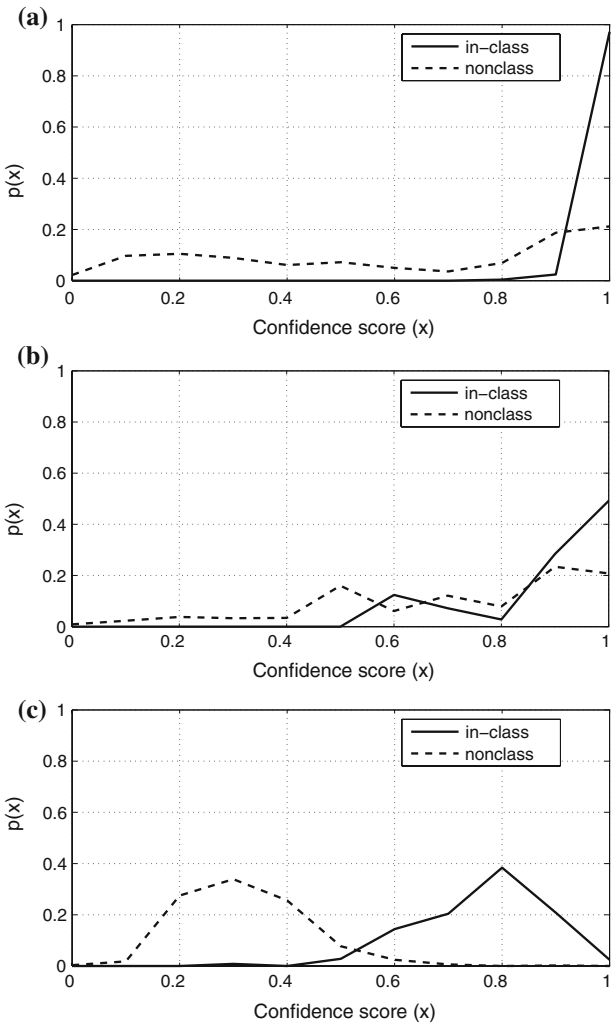

Fig. 7 Histograms of in-class confidence scores along with nonclass confidence scores for **a** AANN models, **b** HMMs, and **c** SVM models

category, is labeled as cricket, then the test video clip is said to be misclassified. For verification, 100 test video clips of each of cartoon, commercial and news categories are used. The average misclassification is less than 15% for classifiers based on AANN and SVM. Classifier based on HMM does not seem to be very useful for discrimination. The misclassification error may be reduced further by extracting features specific to a given class.

4.6 Performance comparison of proposed approach with existing approaches

The performance of the proposed approach is compared against some existing approaches in the literature [5–15,31]. In [5], the DCT edge features are used to classify video sequences into meaningful semantic segments of 24 s duration. For edges and their duration as features, the correct

Table 6 Performance of misclassification (in %) obtained from AANN models, for test clips which do not belong to any of the five sports category

	Cricket	Football	Tennis	Basketball	Volleyball	Average performance
Cartoon	08	06	02	01	01	3.60
Commercial	19	12	08	03	02	8.80
News	24	17	21	05	04	14.20

Table 7 Performance of misclassification (in %) obtained from SVM models, for test clips which do not belong to any of the five sports category

	Cricket	Football	Tennis	Basketball	Volleyball	Average performance
Cartoon	39	16	02	01	04	12.00
Commercial	34	03	02	01	01	8.40
News	55	12	01	02	01	14.00

Table 8 Performance of misclassification (in %) obtained from HMM models, for test clips which do not belong to any of the five sports category

	Cricket	Football	Tennis	Basketball	Volleyball	Average performance
Cartoon	47	16	27	01	25	22.20
Commercial	59	02	28	02	33	24.80
News	11	08	01	02	01	4.40

classification is about 65%. In [6], the average edge ratio is used in conjunction with *k*-nearest neighbor algorithm. The method requires 120 s of test data to yield a classification performance of 92.4% on a five-class (badminton, soccer, basketball, tennis, and skating) problem. In [7], a motion pattern was used to classify the video contents at the semantic level using SVM. A 10-h long video program including science and educational films, sight-seeing videos, stage performances, and sports video were segmented into shots, and each shot was classified into semantic concepts. An average classification performance of 94% was achieved. The approach described in [8] uses statistical nonparametric modeling of motion information to classify video shots into temporal texture (rivers, grass motion, trees in the wind), sequences of pedestrians and traffic video shots. The method achieved mean recognition rate higher than 90%.

Motion dynamics such as foreground object motion and background camera motion are extracted in [10] for classification of video sequences into three categories, namely, sports, cartoon and news using Gaussian mixture models (GMMs). Using 30 s clips, this method gives a classification performance of about 94%. The approach described in [11] uses statistical models (GMM) of reduced DCT or Hadamard transform coefficients, and gives 87% correct classification rate for six different video shots consisting of presentation graphics, long shots of the projection screen both lit and unlit, long shots of the audience and medium close-ups of human figures on light and dark backgrounds.

In [12], GMM was used to model low-level audio/video features for the classification of five different categories, namely, sports, cartoon, news, commercial, and music. An average correct classification rate of 86.5% was achieved with 1 h of recordings per genre, consisting of continuous sequences of 5 min each and 40 s decision window. In [13], vector quantization and HMM are used to characterize sports videos such as basketball, ice hockey, soccer, and volleyball based on motion information. The length of the video sequences ranged from 43 s to 1 min. The method achieved an average performance of about 82 and 88% using vector quantization and HMM, respectively.

Another approach described in [14] uses HMMs and motion and color features for classification of four different sports videos, namely, ice hockey, basketball, football and soccer. The method achieved an overall classification accuracy of 93%. Each testing sequence is of 100 s duration. In [15], a decision tree method was used to classify video shots into different genres such as movie, commercial, music, and sports based on motion and color information. This method achieved an average prediction accuracy of nearly 75%. In [31], the C4.5 decision tree was used to build the classifier for genre labeling using a set of features that embody the visual characteristics of video sequences, such as news, commercial, music, sports, and cartoon. The average classification performance was between 80 and 83% for video clips of 60 s duration.

The method proposed in this paper uses AANNs to classify five sports categories, namely, cricket, football, tennis, basketball, and volleyball based on edge-specific features. Despite using shorter duration test clips (20 s), the proposed method yields an average classification performance of 94.4% which compares favorably with existing approaches that use longer duration test clip.

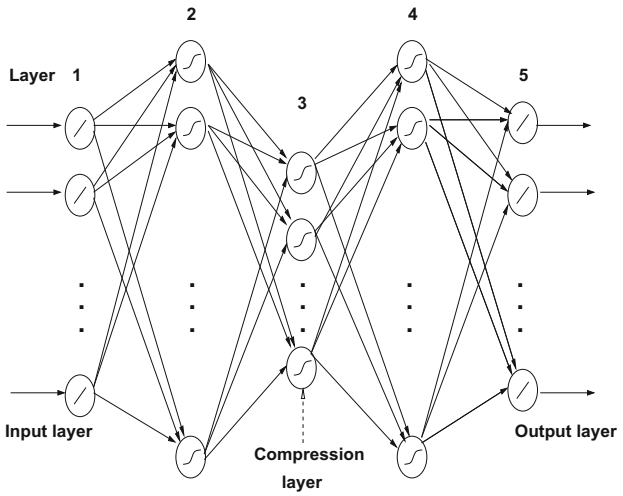

Fig. 8 A five-layer AANN model

5 Summary

We have presented an approach for classification of sports video based on edge-specific features, namely, EDH and EIH. We have exploited the ability of AANNs to capture the distribution of feature vectors. Two other classifiers, namely, HMMs and SVMs were also examined. A video database of TV broadcast programs containing five sports video categories, namely, cricket, football, tennis, basketball and volleyball was used for training and testing the models. Experimental results indicate that the edge-based features can provide useful information for discriminating the classes considered, and that the EDH is a superior feature compared to the EIH. It was shown that combining evidences from complementary edge features and from different classifiers results in an improvement in the performance of classification. It is also observed that the classification system is able to decide, whether a given test video clip belongs to one of the five predefined video categories or not. In order to achieve better classification performance, evidence from audio and visual features may be combined.

Appendix

Autoassociative neural network models

Autoassociative neural network models are feedforward neural networks performing an identity mapping of the input space, and are used to capture the distribution of the input data [18,32]. The distribution capturing ability of the AANN model is described in this section. Let us consider the five-layer AANN model shown in Fig. 8, which has three hidden layers. In this network, the second and fourth layers have more units than the input layer. The third layer has fewer units than the first or fifth. The processing units in the first and third hidden layer are nonlinear, and the units in the second compression/hidden layer can be linear or nonlinear. As the error between the actual and the desired output vectors is minimized, the cluster of points in the input space determines the shape of the hypersurface obtained by the projection onto the lower dimensional space. Figure 9b shows the space spanned by the 1-dimensional compression layer for the 2-dimensional data shown in Fig. 9a for the network structure *2L 10N 1N 10N 2L*, where *L* denotes a linear unit and *N* denotes a nonlinear unit. The integer value indicates the number of units used in that layer. The nonlinear output function for each unit is $\tanh(s)$, where s is the activation value of the unit. The network is trained using backpropagation algorithm [21,22]. The solid lines shown in Fig. 9b indicate mapping of the given input points due to the 1-dimensional

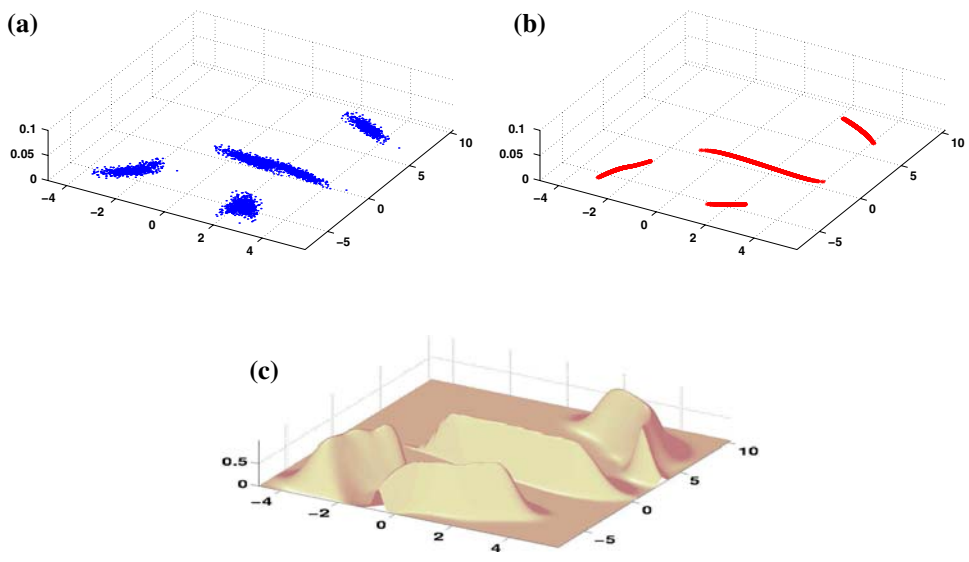

Fig. 9 Distribution capturing ability of AANN model. **a** Artificial 2-dimensional data. **b** 2-dimensional output of AANN model with the structure *2L 10N 1N 10N 2L*. **c** Probability surfaces realized by the network structure *2L 10N 1N 10N 2L*

compression layer. Thus, one can say that the AANN captures the distribution of the input data depending on the constraints imposed by the structure of the network, just as the number of mixtures and Gaussian functions do in the case of GMM.

In order to visualize the distribution better, one can plot the error for each input data point in the form of some probability surface as shown in Fig. 9c. The error E_i for the data point i in the input space is plotted as $p_i = \exp^{(-E_i/\alpha)}$, where α is a constant. Note that p_i is not strictly a probability density function, but we call the resulting surface as probability surface. The plot of the probability surface shows a large amplitude for smaller error E_i, indicating better match of the network for that data point. The constraints imposed by the network can be seen by the shape the error surface takes in both the cases. One can use the probability surface to study the characteristics of the distribution of the input data captured by the network. Ideally, one would like to achieve the best probability surface, best defined in terms of some measure corresponding to a low average error.

References

1. Jain, A., Duin, R., Mao, J.: Statistical pattern recognition: a review. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 4–37 (2000)
2. Wang, Y., Liu, Z., Huang, J.C.: Multimedia content analysis using both audio and visual clues. IEEE Signal Process. Mag. **17**, 12–36 (2000)
3. Assflag, J., Bertini, M., Colombo, C., Bimbo, A.D.: Semantic annotation of sports videos. IEEE Multimedia **9**, 52–60 (2002)
4. Kokaram, A., Rea, N., Dahyot, R., Tekalp, A.M.: Browsing sports video. IEEE Signal Process. Mag. **5021**, 47–58 (2006)
5. Lee, M.H., Nepal, S., Srinivasan, U.: Edge-based semantic classification of sports video sequences. In: International Conference on Multimedia and Expo, Baltimore, MD, USA (2003)
6. Yuan, Y., Wan, C.: The application of edge features in automatic sports genre classification. In: Proceedings of IEEE Conference on Cybernetics and Intelligent Systems, Singapore (2004)
7. Ma, Y.F., Zhang, H.J.: Motion pattern based video classification using support vector machines. In: Proceedings of the IEEE International Symposium on Circuit and Systems, Arizona, USA (2002)
8. Fablet, R., Bouthemy, P.: Statistical modeling for motion-based video classification and retrieval. In: Proceedings of the Workshop on Multimedia Content Based Indexing and Retrieval, France (2001)
9. Takagi, S., Hattori, S., Yokoyama, K., Kodate, A., Taminaga, H.: Sports video categorizing method using camera motion parameters. In: International Conference on Multimedia and Expo, Baltimore, MD, USA (2003)
10. Roach, M.J., Mason, J.S.D., Pawlewski, M.: Video genre classification using dynamics. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Utah, USA (2001)
11. Girgensohn, A., Foote, J.: Video classification using transform coefficients. In: Proceedings of the Internatioanl Conference on Acoustics, Speech and Signal Processing, Arizona, USA (1999)
12. Xu, L.Q., Li, Y.: Video classification using spatial–temporal features and PCA. In: International Conference on Multimedia and Expo, pp. 345–348 (2003)
13. Sahouria, E., Zakhor, A.: Content analysis of video using principal components. IEEE Trans. Circuits Syst. Video Technol. **9**, 1290–1298 (1999)
14. Gibert, X., Li, H., Doermann, D.: Sports video classification using HMMs. In: International Conference on Multimedia and Expo, Baltimore, MD, USA (2003)
15. Yuan, Y., Song, Q.B., Shen, J.Y.: Automatic video classification using decision tree method. In: Proceedings of the IEEE International Conference on Machine Learning and Cybernetics, Beijing (2002)
16. (UPM-GTI, ES), J.M.M.: MPEG-7 Overview (version 8). ISO/IEC JTC1/SC29/WG11 N4980, Klangenfurt (July, 2002)
17. Canny, J.: A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. **8**, 679–698 (1986)
18. Yegnanarayana, B., Kishore, S.: AANN: an alternative to GMM for pattern recognition. Neural Netw. **15**, 459–469 (2002)
19. Rabiner, L., Juang, B.: An introduction to hidden Markov models. IEEE Acoust. Speech Signal Process. Mag. **3**, 4–16 (1986)
20. Collobert, R., Bengio, S.: SVMTorch: support vector machines for large-scale regression problems. J. Mach. Learn. Res. **1**, 143–160 (2001)
21. Yegnanarayana, B.: Artificial Neural Networks. Prentice-Hall India, New Delhi (1999)
22. Haykin, S.: Neural Networks: A Comprehensive Foundation. Prentice-Hall International, New Jersey (1999)
23. Baum, L.E., Petrie, T., Soules, G., Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Stat. **41**, 164–171 (1970)
24. HMM Tool Kit. Available at: http://htk.eng.cam.ac.uk/
25. Burges, C.: A tutorial on support vector machines for pattern recognition. Data Mining Knowl. Discov. **2**, 121–167 (1998)
26. SVM Tool Kit. Available at: http://www.idiap.ch/~bengio/projects/SVMTorch.html
27. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. **20**, 226–239 (1998)
28. Rohlfing, T., Russakoff, D.B., Maurer, C.R.: Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. IEEE Trans. Med. Imaging **23**, 983–994 (2004)
29. Lam, L., Suen, C.Y.: Optimal combinations of pattern classifiers. Pattern Recognit. Lett. **16**, 945–954 (1995)
30. Xu, L., Krzyzak, A., Suen, C.Y.: Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Trans. Syst. Man, Cybern. **2**, 418–435 (1992)
31. Truong, B.T., Venkatesh, S., Dorai, C.: Automatic genre identification for content-based video categorization. In: Proceedings of the International Conference on Pattern Recognition, Barcelona, Spain (2000)
32. Yegnanarayana, B., Gangashetty, S.V., Palanivel, S.: Autoassociative neural network models for pattern recognition tasks in speech and image. In: Soft Computing Approach to Pattern Recognition and Image Processing, World Scientific Publishing Co. Pte. Ltd, Singapore (2002) 283–305