

A Novel Video Content Understanding Scheme Based on Feature Combination Strategy

Xinghao Jiang^{1,2}

¹School of Information Security Engineering, Shanghai Jiao-tong University, Shanghai, China

²Shanghai Information Security Management and Technology Research Key Lab, Shanghai, China
Email: xhjiang@sjtu.edu.cn,

Tanfeng Sun^{1,2}, Bin Chen¹, Rongjie Li¹, Bing Feng¹

¹School of Information Security Engineering, Shanghai Jiao-tong University, Shanghai, China

²Shanghai Information Security Management and Technology Research Key Lab, Shanghai, China
Email: {tfsun, virgoshaka}@sjtu.edu.cn

Abstract—With the development of the multimedia technology, there are more and more video resources on the Internet, which are difficult to automatically recognize, classify and index. So solve these problems, we present a novel video content understanding scheme in this paper. This scheme is based on the combination strategy of different video features. To represent these video features, we use nine standard MPEG-7 descriptors, including color, texture, region and motion descriptors. We extract and combine these descriptors together to represent the whole video character. After that, we use an SVM as the classifier to train the model. The traditional 1-1 method of the SVM is modified by a Second-Prediction Strategy to gain higher classification accuracy. Finally, the videos are classified into five genres, including cartoons, commercial, music, news, and sports. We compare our classification results with some of the results in the recent papers, and demonstrate the effectiveness of our scheme.

Index Terms— content understanding, video content classification, MPEG-7 descriptors, second-prediction, support vector machine

I. INTRODUCTION

With the development of network and multimedia technology, a tremendous amount of video data is springing up on the Internet. Because there is little structural information in the video stream data, it is difficult to analyze the video content, and how to analyze, classify and retrieve videos becomes an important issue. Among these, content-based video classification is a critical technique.

The general process of automatic video classification is as follows. First, video data is analyzed to obtain video features, for example, color and texture, then a classifier is used to classify different videos with these features. Commonly used classifiers include hidden Markov model(HMM), support vector machine(SVM), Bayesian network and others. As it is hard to obtain suitable and effective features for semantic classification, most approaches of video classification do not achieve

high accuracy. Many researches have begun on the video classification area. Yi Haoran used HMM to classify videos in compressed domain^[1]. Li-Qun Xu used PCA method to reduce video feature dimensions^[2]. Ankush Mittal compared common classifiers and classified videos using Bayesian network and high-dimensional features^[3]. W.J. Gillespie used RBF network to classify videos^[4]. M.Kalaiselvi Geetha combined several video features and used HMM to do the classification^[5], and achieved acceptable results.

However, as different approaches extract different kind of video features, it is hard to scale these approaches' effect from a benchmark. What is more, the extracted features such as color histogram and texture are difficult to be comprehended directly by human beings, and those video content, which humans can easily comprehend, are difficult to extract. Due to these problems, some researchers introduced the descriptors in the MPEG-7 standard. Evaggelos Spyrou combined three MPEG-7 descriptors to classify the image content^[6], and Wensheng Zhou also used some MPEG-7 descriptors in the basketball scene classification^[7]. However, they only applied some of the MPEG-7 descriptors in the classification and did not achieve a high accuracy. Based on the previous research work, we present a video classification approach which applies MPEG-7 descriptors on a large scale. In this approach, we extract the descriptors and combine them as a whole video feature. Then we use SVM as the classifier, and enable the second-prediction strategy to improve SVM's 1-1 vote method. Experiments show that the classification accuracy is improved.

The rest of this paper is organized as follows: Section II introduces the MPEG-7 descriptors and how to extract them. Section III presents the principle of SVM and its second-prediction strategy. Experiments and results of video classification are discussed in Section IV. The conclusion follows in Section V.

II. MPEG-7 DESCRIPTORS EXTRACTION

MPEG-7, formally named "Multimedia Content Description Interface", is an ISO/IEC standard developed

Supported by the National Natural Science Foundation of China (No.60702042, 60802057), National 863 Plan of China (2007AA01Z455), and Shanghai Research Scholar Plan of China (08XD14023).

by the MPEG committee. Its aim is to provide a set of standardized tools to describe multimedia content, thus to enable people to search and retrieve the multimedia information they are interested in. MPEG-7 standard includes many descriptors such as color, texture, shape and motion, and each descriptor defines the particular syntax and semantics of a certain fundamental visual feature. MPEG-7 also allow users to define their own descriptors. However, MPEG-7 only standardize the presentation of descriptors, and the extraction algorithm of descriptors is not part of the standard. So applications have to implement the extraction algorithm based on their own needs. In this paper, we extract seven MPEG-7 standard descriptors: dominant color descriptor (DCD), color layout descriptor (CLD), color structure descriptor (CSD), GoF/GoP color descriptor (GoP), homogenous texture descriptor (HTD), edge histogram descriptor (EHD), region shape descriptor (RSD), and define two new descriptors: chunk color descriptor (CCD), motion intensity descriptor (MID).

Since many descriptors are dealing with images, we first decode the video data to a set of image frames, then extract the descriptors in each image frame and take the median value of these descriptors as the whole video's descriptor. The alternative option is to take the average value, but we choose the median value because it has better robustness in case of some noises.

The nine descriptors will be introduced in the following, and the details of MPEG-7 standard descriptors can be found in the MPEG-7 official document^[8].

A. MPEG-7 Standard Descriptors

1) Dominant Color Descriptor

This descriptor specifies a set of dominant colors in an arbitrarily shaped region. In this region, the color information will be represented as a few dominant colors.

This descriptor is defined by

$$F(DCD) = \{ \{ C_k, P_k, V_k \}, S \}, k = 1, 2, \dots, N \quad (1)$$

where C_k represents the i th dominant color, P_k is its percentage value, V_k is its color variance, which is an option, and S represents its spatial coherency.

We extract C_k , which is the index of the dominant color. The extraction algorithm is as follows:

$$C_k = \text{Median} \{ I_0^{\max k}, \dots, I_{n-1}^{\max k} \}, k = 1, 2, 3 \quad (2)$$

where 0 to $n-1$ represent the frame number in the video, and $I_i^{\max k}$ represents the k th dominant color's index of the i th frame.

Since we extract the first three dominant colors, and each dominant color consists of three components of R, G and B, the dimension of DCD we extract is 9.

2) Color Layout Descriptor

This descriptor is designed to capture the spatial distribution of color in an image or an arbitrary-shaped region. It can be used for sketch-based image retrieval and content-based image filtering. It not only has a simple representation, but also has good effectiveness.

CLD is extracted in the YCbCr color space. The feature extraction process consists of two parts: grid based representative color selection and DCT transform with quantization. An input picture is divided into 64 blocks and their average colors are derived. Then the derived average colors are transformed into a series of coefficients by performing DCT. Finally a few low-frequency coefficients are selected using zigzag scanning and quantized to form a CLD.

We extract 6 Y coefficients, 3 Cb coefficients, 3 Cr coefficients, so the dimension of CLD is 12.

3) Color Structure Descriptor

This descriptor specifies both color content (similar to that of a color histogram) and the structure of this content. It does this via the use of a structuring element. Unlike the color histogram, this descriptor can distinguish between two images in which a given color is present in identical amounts but where the structure of the groups of pixels having that color is different in the two images. Its main function is image matching and retrieval, where an image may consist of either arbitrarily shaped, possibly disconnected, regions or a single rectangular frame.

CSD is extracted in the HMMD color space. In order to compute the CSD, an 8×8 -structuring element is used. The spatial extent of the structuring element is determined by the following simple rule:

$$p = \max \{ 0, \text{round}(0.5 \log_2 WH - 8) \} \quad (3)$$

$$K = 2^p, E = 8K$$

where W, H are image width and height, respectively, $E \times E$ is the spatial extent of the structuring element, and K is the sub-sampling factor.

We extract a 256-bin color structure histogram, so the dimension of CSD is 256.

4) GoF/GoP Color Descriptor

This descriptor specifies a structure required for representing the color features of a collection of images or video frames by means of the scalable color descriptor. The collection of video frames can be a contiguous video segment or a non-contiguous collection of similar video frames.

GoP is extracted in the HSV color space. Its extraction algorithm is as follows:

$$F(\text{GoP})[j] = \text{Haar}(\text{Median}\{\text{Histogram}_0[j], \dots, \text{Histogram}_{n-1}[j]\}), j = 0, \dots, 255 \quad (4)$$

where $\text{Histogram}_0[j]$ is the color histogram value of the j th bin in the first frame.

In this paper, we extract a 256-bin color histogram, so the dimension of GoP is 256.

5) Homogeneous Texture Descriptor

Artificially made videos such as cartoons usually have different texture from those ordinary videos. This descriptor provides a quantitative characterization of texture for similarity based search and retrieval applications.

The computation of this descriptor is as follows. The frequency space is partitioned into 30 channels with equal divisions in the angular direction and octave division in the radial direction. Then the center frequencies of the feature channels are spaced equally in

30 degrees in angular direction, and the center frequencies of the neighboring feature channels are spaced one octave apart in the radial direction. Then the individual feature channels are modeled using the following 2-D Gabor functions,

$$G_{p,s,r}(\omega, \theta) = \exp\left[-\frac{(\omega - \omega_s)^2}{2\sigma_{\rho_s}^2}\right] \cdot \exp\left[-\frac{(\theta - \theta_r)^2}{2\sigma_{\theta_r}^2}\right] \quad (5)$$

where $\sigma_{\theta_r} = \frac{15^\circ}{\sqrt{2\ln 2}}, \sigma_{\rho_s} = \frac{B_s}{2\sqrt{2\ln 2}}$

Then calculate the energy

$$e_i = \log_{10}[1 + p_i] \quad (6)$$

where $p_i = \int_{\omega=0^+}^1 \int_{\theta=(0^\circ)^+}^{360^\circ} [G_{p,s,r}(\omega, \theta) \cdot P(\omega, \theta)]^2$

$$i = 6 \times s + r + 1$$

and the energy deviation

$$d_i = \log_{10}[1 + q_i] \quad (7)$$

where $q_i = \sqrt{\int_{\omega=0^+}^1 \int_{\theta=(0^\circ)^+}^{360^\circ} \{[G_{p,s,r}(\omega, \theta) \cdot P(\omega, \theta)]^2 - p_i\}^2}$

$$i = 6 \times s + r + 1$$

In this paper, we extract 62 dimensions of HTD, including average intensity, standard deviation of intensity, energy and energy deviation.

6) Edge Histogram Descriptor

This descriptor specifies the spatial distribution of different types of edges in local image regions. The distribution of edges is a good texture signature that is useful for image matching even when the underlying texture is not homogeneous.

The computation of this descriptor is as follows. A given image is first sub-divided into sub-images, and local edge histograms for each of these sub-images is computed. Edges are broadly grouped into five categories: vertical, horizontal, 45 diagonal, 135 diagonal, and isotropic. Thus, each local histogram has five bins corresponding to the above five categories. The image partitioned into 16 sub-images results in 80 bins. So the dimension of EHD is $16 \times 5 = 80$.

7) Region Shape Descriptor

This descriptor specifies the region-based shape of an object. The shape of an object may consist of either a single region or a set of regions, as well as some holes in the object. Since the region-based shape descriptor makes use of all pixels constituting the shape, it can describe any shape, i.e. not only a simple shape with a single connected region but also a complex shape that consists of several disjoint regions.

Region Shape Descriptor utilizes a set of ART (Angular Radial Transform) coefficients, including 12 angular basis functions and 3 radial basis functions. While there is no definition when angle and radius are both zero, the dimension of RSD is 35.

B. Chunk Color Descriptor

We define the Chunk Color Descriptor to specify the color distribution in different regions of a video

frame. Different genres of videos often have different regional color distribution. For example, green grass is often in the bottom of a soccer video, and the stands are often on the top.

We extract CCD in the HSV color space. The process is as follows:

1) Divide the video frame into $4 \times 4 = 16$ chunks.

2) Compute a 36-bin non-uniform color histogram in each chunk. The details of color histogram is defined as follows^[9]:

For $v \in [0, 0.2)$, it is a black area, $l = 0$

For $s \in [0, 0.2]$ and $v \in [0.2, 0.8)$, it is a gray area, $l = \text{floor}((v - 0.2) \times 10) + 1$

For $s \in [0, 0.2]$ and $v \in [0.8, 1.0]$, it is a white area, $l = 7$

For $s \in (0.2, 1.0]$ and $v \in (0.2, 1.0]$, it is a color area,

$$H = \begin{cases} 0, & h \in (330, 22] \\ 1, & h \in (22, 45] \\ 2, & h \in (45, 70] \\ 3, & h \in (70, 155] \\ 4, & h \in (155, 186] \\ 5, & h \in (186, 278] \\ 6, & h \in (278, 330] \end{cases}$$

$$S = \begin{cases} 0, & s \in (0.2, 0.65] \\ 1, & s \in (0.65, 1] \end{cases} \quad V = \begin{cases} 0, & v \in (0.2, 0.7] \\ 1, & v \in (0.7, 1] \end{cases}$$

$$l = 4H + 2S + V + 8$$

where l is the bin index of the color histogram, h, s, v are the three components of the HSV color, and floor rounds the elements to the nearest integers less than or equal to the original value.

3) Do the above operation in each frame of a video, then take the median value as the CCD value.

Since there are 16 chunks in a frame, and each chunk has a 36-bin color histogram, the dimension of CCD is 576.

C. Motion Intensity Descriptor

We define this descriptor to specify the intensity of motion in a video. As we know, sports and vocal concerts usually have high motion intensity, while news videos usually have low motion intensity. This descriptor can distinguish between different genres of videos using this time-dependent information.

For a video clip, which is a contiguous sequence of n frames, $Video = \{f_0, f_1, \dots, f_{n-1}\}$, we compute δ_i^u and δ_i^v jointly from frames f_{i-1} and f_i . Specifically, $\delta_i^u = |f_i^u - f_{i-1}^u|$ and $\delta_i^v = |f_i^v - f_{i-1}^v|$, where δ_i^u and δ_i^v are the mean and variance of luminance values of pixels in frame f_i , respectively.

We extract MID using the following expressions:

TABLE I. SUMMARY OF DESCRIPTORS

Descriptors	Description	Dimension
DCD	specifies a set of dominant colors	9
CLD	captures the spatial distribution of color	12
CSD	specifies both color content and the structure of the content	256
GoP	specifies a structure for representing video frames' color features	256
HTD	provides a quantitative characterization of texture	62
EHD	specifies the spatial distribution of different types of edges	80
RSD	specifies the region-based shape of an object	35
CCD	specifies the color distribution in different regions of a video frame	576
MID	specifies the intensity of motion in a video	3

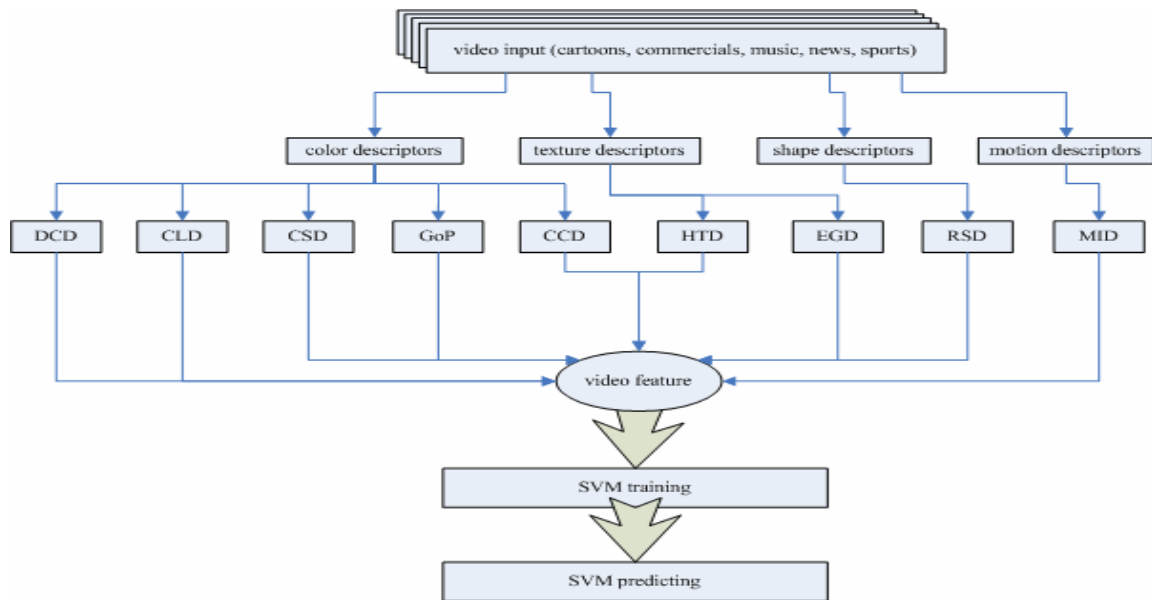


Figure 1. Video Classification Scheme using MPEG-7 descriptors and SVM

$$MID_1 = \frac{|\Omega_1|}{n-1}, \tag{8}$$

where $\Omega_1 = \{\delta_i^v \mid \delta_i^v > Th_1\}$, and Th_1 is a certain threshold.

$$MID_2 = \frac{|\Omega_2|}{n-1}, \tag{9}$$

where $\Omega_2 = \{\delta_i^v \mid \delta_i^u < Th_2, \delta_i^v < Th_3\}$, and Th_2 and Th_3 are certain thresholds.

$$MID_3 = \frac{\sum_{i=2}^n \delta_i^c}{n-1}, \tag{10}$$

where

$$\delta_i^c = \frac{1}{w \cdot h} \sum_{k=1}^{wh} \sqrt{(f_i^R(k) - f_{i-1}^R(k))^2 + (f_i^G(k) - f_{i-1}^G(k))^2 + (f_i^B(k) - f_{i-1}^B(k))^2}$$

w and h are width and height of the video frames.

As described above, the dimension of MID we extract is 3.

D. Summary of Descriptors

The most important thing we care about is how to narrow the huge gap between the low-level features and high-level semantics of the video. However, the existing approaches do not solve this problem perfectly. So in this paper we suggest an approach which combines MPEG-7 descriptors as the video features. Each descriptor's contribution to video classification is shown Section IV. The overall classification accuracy comparison in Section IV demonstrates that we can achieve a good classification result by combining all those descriptors together.

TABLE II. DESCRIPTORS SELECTED ACCORDING TO THE TWO CLASSES

Descriptors	Cartoons	Commercials	Music	News
Sports	CSD,HTD,EHD,RSD	CSD,GoP,RSD,CCD	CSD,RSD,MID	CSD,RSD,MID
News	GoP,HTD,EHD,MID	GoP,CSD,MID	CSD,GoP,MID	(-)
Music	GoP,HTD,EHD,MID	GoP,CSD,MID	(-)	(-)
Commercials	GoP,HTD,EHD	(-)	(-)	(-)

III. VIDEO CLASSIFICATION SCHEME USING SECOND-PREDICTION STRATEGY

A. Video Classification Scheme

Figure 1 shows the video classification scheme in this paper. Videos are classified into five genres such as cartoons, commercials, music, news, and sports. First we extract nine descriptors including color, texture, shape and motion from the input video data, then combine them as the whole video feature, and put it into a classifier to train and test. Finally we get classification result.

Classifier is one of the most important part in video classification. In this paper we use SVM as the video classifier.

B. SVM and Its Second-Prediction Strategy

Support Vector Machine^[10] (SVM) is a useful technique for data classification. In recent years it has been widely used in pattern recognition, data mining and other related areas. SVMs are originally designed for binary classification, and we can extend them for multi-class classification. There are three commonly used methods: 1-r (one-against-rest), 1-1 (one-against-one) and DAGSVM. Among the three methods, the 1-1 method usually has the highest accuracy.

For a case of n classes, the 1-1 method constructs $n(n-1)/2$ classifiers where each one is trained on data from two classes. After testing data on each classifier, it gets $n(n-1)/2$ results. Then it uses the following vote strategy: if a classifier says the result is class x , then the vote for class x is added by one. Finally it gets $n(n-1)/2$ votes, and the class with the highest votes is the final prediction result.

Though the 1-1 method is known as its high accuracy, it has some drawbacks. In case that two classes have identical votes, it is difficult to select the final result, and it will lower the accuracy. To solve this problem, we suggest a technique called second-prediction strategy. In this strategy, we add a second-prediction step based on 1-1 method if there are two classes with identical votes.

The detail steps of the Second-Prediction Strategy is as follows:

- (1) Construct $n(n-1)/2$ classifiers for each two classes.
- (2) Train the two classes of data in the

corresponding classifier.

(3) When predicting, put the test data in all the classifiers and get $n(n-1)/2$ results.

(4) Vote on the results and see if there are two classes with the highest and identical votes.

(5) If no, choose the class with the highest votes as the final prediction result.

(6) If yes, select the specified descriptors according to the classes and construct another SVM.

(7) Train and do a second prediction between the

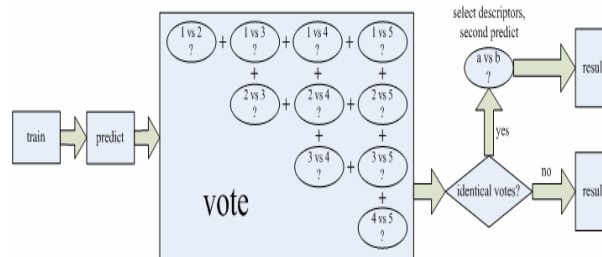


Figure 2. Second-Prediction Strategy

two classes, then take this result as the final prediction result.

In the (6) step, we select different descriptors according to the two classes to train and predict. For example, we use the texture descriptor to distinguish cartoons and other classes because cartoons have very different texture. And we use the motion descriptor to distinguish news and other classes, since news videos are most stationary, which is very different with commercials and music. The reason we use only part of the descriptors is that we can not only save the computation time but also discard the noise factor, which is helpful to the whole classification accuracy. The details of the descriptors we select are listed in Table II. The single descriptor comparison experiment will demonstrate that these are the most effective descriptors to distinguish the two particular video genres.

Since there are 5 different classes, there are 10 possible cases in the second prediction between the two classes. In our implementation, we train the 10 different SVM models at first and predict between the two classes immediately when the second-prediction is needed. Thus we enhance the second-prediction's efficiency by saving its possible training time.

We will demonstrate in Section IV's experiments that, the 1-1 method with the second-prediction strategy has higher accuracy than the original one.

TABLE III. CLASSIFICATION ACCURACY ON EACH SINGLE DESCRIPTOR

Descriptor	DCD	CLD	CSD	GoP	HTD	EHD	RSD	CCD	MID
Cartoons	44.0	63.3	78.7	76.0	79.7	84.0	58.7	70.0	30.7
Commerci	36.4	62.1	60.5	73.3	60.5	67.3	53.0	58.3	23.5
Music	73.3	68.7	81.3	85.3	70.3	75.0	68.7	72.0	77.7
News	71.7	70.1	82.5	87.3	77.9	82.7	59.4	72.3	82.7
Sports	82.5	82.5	91.6	87.3	82.2	82.9	73.0	83.9	54.6
Overall	61.6	69.3	78.9	81.8	74.1	78.4	62.6	71.3	53.8

IV. EXPERIMENTAL RESULTS

We download from Internet five genres of videos, including cartoons, commercials, music, news and sports. They are all MPEG-4 videos with different resolutions from 320*240 to 640*480. Most videos' frame rate is 25 fps or 30 fps, while some cartoons have a frame rate of 15 fps. The total length of videos is over 6 hours, including 1 hour 15 minutes of cartoons, 1 hour of commercials, 1 hour 15 minutes of music, 1 hour 30 minutes of news, and 1 hour 13 minutes of sports. There are totally 7658 samples in our experiment, including 1500 samples of cartoons, 1321 samples of commercials, 1500 samples of music, 1860 samples of news, and 1477 samples of sports. We randomly select half of the samples as the training data and the other half as the test data. DAGSVM, original 1-1 SVM and 1-1 SVM with the second-prediction strategy are all implemented based on libsvm^[11]. We use RBF as the kernel function, and select the best parameters c and γ from cross validation.

The software tools used in the experiment include ffmpeg, MPEG-7 eXperimentation Model (XM), MATLAB, and libsvm. We use ffmpeg to convert videos, and XM to extract features from videos. We also use MATLAB to implement some self-defined feature extraction algorithms. Libsvm is an open source SVM project and we use it to implement our classification algorithms.

A. Accuracy Comparison on Each Single Descriptor

We have compared the classification accuracy on each single descriptor. These accuracies reflect each descriptor's effect and contribution to the video classification. And we have also selected the specified descriptors based on them in the second-prediction strategy. We use original 1-1 SVM in this comparison.

As we can see from Table III, GoP reaches an overall accuracy of 81.8%, which is the highest in color descriptors. This indicates that using GoP is an effective way in video classification. For other color descriptors, CSD has the highest accuracy on sports, which is 91.6%, and CCD also performs well on sports. This is probably because sports videos have a relatively fixed color distribution.

For texture descriptors, HTD and EHD both have a high accuracy on cartoons, which indicates that using texture is the easiest way to distinguish cartoons. Since cartoons are made artificially, they have a smoother texture than most of other videos.

For shape descriptors, RSD has the highest accuracy on sports. This is maybe because in sports video the field and stands have regular shapes.

For motion descriptors, MID can distinguish music and news best, with accuracy of 77.7% and 82.7%. MID reflects the motion intensity of a video. Music, especially vocal concerts, often have flash lights and jumping background, so they have a larger MID. In the opposite, most news videos contain nearly static scenes, or the movements of scenes are fairly slow, so they have a smaller MID than other videos.

Figure 3 shows the overall accuracy on each single descriptor. We can see that the result is not so good as we have expected. They have the lowest accuracy of 53.8%, and the highest accuracy of 81.8%. This indicates that single descriptors are not enough for video classification. We should use all these descriptors and combine them to achieve a better result.

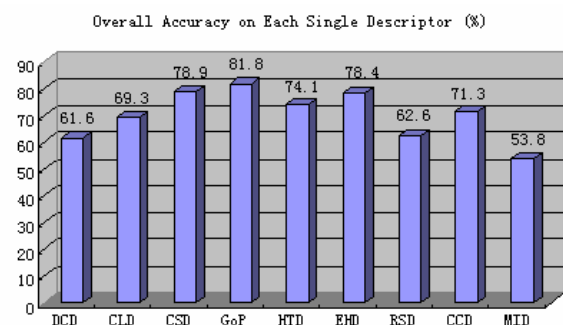


Figure 3. Overall Accuracy on Each Single Descriptor

B. Accuracy Comparison on Multi-class Classification Method

We use the Python programming language to implement three SVM multiclass classification methods mentioned in Section III and compare on them. The result is showed in Table IV and Figure 4.

We have totally tested 3829 samples. DAGSVM has 267 incorrect predictions, original 1-1 SVM has 105

incorrect predictions, and 1-1 SVM with the second-prediction strategy has 41 incorrect predictions. The last one has the highest accuracy of 98.93%, which is 5.90% and 1.67% higher than the accuracy of DAGSVM and original 1-1 SVM. Our analysis shows that in 3829 test samples, there are 170 samples which have identical votes. Among the 170 samples, 31 of them do not belong to the genre with identical votes, so it is impossible to predict them correctly. For the other 139 samples, 1-1 SVM with the second-prediction strategy predicts 129 samples correctly, while the original 1-1 SVM predicts 65 samples correctly. So the improved method does 64 less incorrect predictions than the original one, and it helps to enhance the accuracy.

TABLE IV. ACCURACY ON DIFFERENT MULTI-CLASS CLASSIFICATION METHODS

Methods	DAGSVM	Original 1-1SVM	Second-prediction strategy
Incorrect	267/3829	105/3829	41/3829
Accuracy (%)	93.03	97.26	98.93

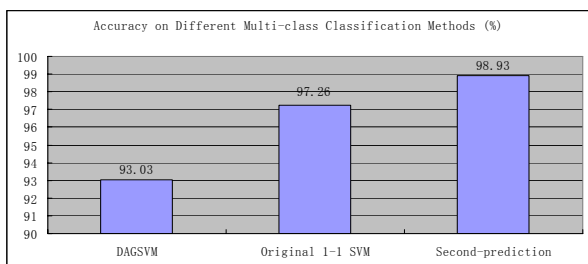


Figure 4. Comparison of Different Classification Methods

C. Overall Classification Accuracy Comparison

We have compared our overall classification accuracy with the results in [5] and [2]. First lines in Table V represent our overall accuracy, second lines with brackets represent the accuracy in [5], and third lines with brackets represent the accuracy in [2]. We use the second-prediction strategy in this comparison.

As we can see from Table V and Figure 5, our overall accuracy is higher than the accuracy in [5] and [2] by combining MPEG-7 descriptors and using the second-prediction strategy. Our accuracy for cartoons, commercials, music, news, sports is 97.7%, 97.4%, 99.5%, 99.8%, 99.4%, respectively, and the accuracy over all five genres of videos reaches 98.93%. Some accuracies are missed in Table V because sports videos are not included in [5].

As we have expected, the largest part of misclassified videos comes from cartoons and commercials. Cartoons misclassified as commercials and commercials misclassified as cartoons account for 1.33% and 1.36%, respectively. This is because some commercials, especially commercials for children’s products, are made in the same way as cartoons, and it

makes these cartoon-like commercials difficult to classify. Commercials misclassified as news also account for 1.06%. This is probably because some static commercials, which contain a lot of text, are similar to news videos. In addition, some cartoons are misclassified as music and some music are classified as commercials.

We can also see from Table V that commercials have the lowest accuracy among the five genres. The reason is that commercials come from all kinds of sources and do not have a regular form. In commercials there are often music and sport scenes, sometimes cartoons. All these elements are mixed together and the features of commercials are similar to other videos. It makes commercials difficult to be distinguished from others. It is a big problem in automatic video classification.

TABLE V. OVERALL ACCURACY COMPARISON (%)

Genres	Cart.	Comm.	Music	News	Sports
Cart.	97.7 (94.9) (79.5)	1.33 (0) (17.1)	0.73 (5.58) (3.4)	0.20 (0) (0)	0 (-) (0)
Comm.	1.36 (0) (6.8)	97.4 (92.5) (89.7)	0.15 (0) (0)	1.06 (7.47) (3.5)	0 (-) (0)
Music	0.07 (0) (0)	0.40 (0) (15.8)	99.5 (100) (81.6)	0 (0) (2.6)	0 (-) (0)
News	0.11 (0) (4.6)	0.05 (2.44) (4.5)	0 (0) (3.4)	99.8 (97.6) (87.5)	0 (-) (0)
Sports	0.20 (-) (0)	0.34 (-) (0)	0 (-) (0)	0.07 (-) (3.3)	99.4 (-) (96.7)

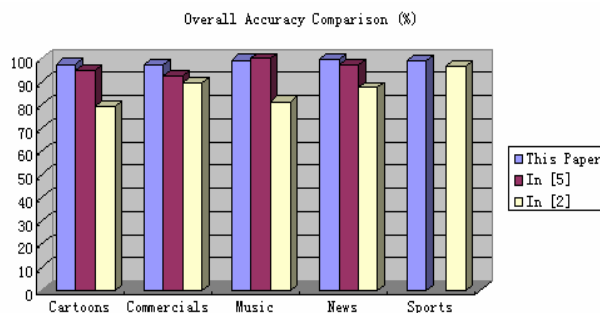


Figure 5. Overall Accuracy Comparison

V. CONCLUSION

To deal with the increasing video resources on the Internet, an effective scheme for automatic video content understanding is presented in this paper. The videos are classified into five different genres, including cartoons, commercials, music, news, and sports. First we extract MPEG-7 descriptors from the raw video data, and combine them to represent the whole video feature. Then

we put the feature into the SVM, which we use as the classifier. We add the second-prediction strategy to the traditional 1-1 SVM to improve the classification accuracy, which is demonstrated to be effective. And our final classification result is competitive to some results in the recent papers.

With the development of the Internet technology, the content security of the multimedia has been a critical problem. It is necessary to prevent some unauthorized video from spreading on the Internet. This paper provides an approach to classify the video content, which is demonstrated to be of high accuracy. However, due to the complexity of the video contents, it is a difficult task to distinguish between some genres of videos. So future work will be devoted to extract more effective features to improve the classify accuracy of those videos which are difficult to distinguish. We will also add audio and text information into our approach and extend the current approach to other areas of the multimedia content security.

The future prospect of the approach suggested in this paper is promising, and can be further improved in many aspects. Such algorithms will be real time when dealing with video streams, and of high accuracy when distinguishing the video content. Further more, the approach in this paper will be applied to the video content understanding and video searching technology.

ACKNOWLEDGMENT

I wish to thank Dr. Tanfeng Sun and Mr. Bin Chen, Mr. Rongjie Li, Mr. Bing Feng for their hard work.

This research is funded by NSFC of China under grant number of 60702042/60802057 and the National 863 Hi-Tech Research and Development plan of China under grant number of 2009AA01Z407. This work is partly funded by the Shanghai Research Scholar Plan under grant No.08XD14023.

REFERENCES

- [1] Yi Haoran, Deepu Rajan, Chia Liang-Tien. An Efficient Video Classification System Based On HMM In Compressed Domain[C]. *IEEE Pacific-Rim Conference on Multimedia*, 2003, pp. 1546-1550.
- [2] Li-Qun Xu, Yongmin Li. Video classification using spatial-temporal features and PCA[C]. *Proceedings of the 2003 International Conference on Multimedia and Expo (ICME '03)*, 2003, pp. 485-488.
- [3] Ankush Mittal, Loong-Fah Cheong. Addressing the Problems of Bayesian Network Classification of Video Using High-Dimensional Features[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2004, pp. 230-244.
- [4] W.J. Gillespie, D.T. Nguyen. Video Classification Using a Tree-Based RBF Network[C]. *IEEE International Conference on Image Processing*, 2005, pp. 465-8.
- [5] M. Kalaiselvi Geetha, S. Palanivel. HMM Based Automatic Video Classification Using Static and Dynamic Features[C]. *International Conference on Computational Intelligence and Multimedia Applications*, 2007, pp. 277-281.
- [6] Evaggelos Spyrou, Hervé Le Borgne, Theofilos Mailis, Eddie Cooke, Yannis Avrithis, and Noel O'Connor. Fusing MPEG-7 Visual Descriptors for Image Classification. *ICANN 2005*, LNCS 3697, pp. 847-852.
- [7] Wensheng Zhou, Asha Vellaikal, C.-C. Jay Kuo. Video Analysis and Classification for MPEG-7 Applications[C]. *IEEE International Conference on Consumer Electronics*, 2000, pp. 344-345.
- [8] ISO/IEC JTC1/SC29/WG11 Coding of Moving Pictures and Audio, Overview of the MPEG-7 Standard, *Int'l Organization for Standardisation*, Oct. 2000.
- [9] Lei, Z., L. Fuzong, and Z. Bo, A CBIR method based on color-spatial feature[C], *TENCON 99, Proceedings of the IEEE Region 10 Conference*, 1999, pp. 166-169.
- [10] Cherkassky V, Ma Y. Practical selection of SVM parameters and noise estimation for SVM regression[J]. *Neural Networks*, 2004, pp. 113-126.
- [11] Chih-Chung Chang, Chih-Jen Lin. LIBSVM: a library for support vector machine. <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>.



Prof. Xinghao Jiang was born Zhejiang of China in 1976. He earned PHD degree of electronic science and technology specialty from Zhejiang University in Hangzhou of China in 2003. He earned Professor title in Shanghai Jiao Tong University in 2006. He is a professor in School of Information Security Engineering of Shanghai Jiao Tong

University now. major field of study

He presides at the national nature science foundation and the national 863 hi-tech research plan now. His research includes cyber information security, PMI, security identity authentication, information hiding and watermarking.



Dr. Tanfeng Sun was born Changchun of China in 1975. He earned his PHD degree of information and communication engineering specialty from Jilin University in Jilin of China in 2003. He received lecture title in Shanghai Jiao Tong University in 2005. He is a teacher in School of Information Security Engineering in Shanghai Jiao Tong

University now.

He presides at the national nature science foundation. He participates in the national 863 hi-tech research plan now. His research includes cyber information security, multimedia content security, information hiding and watermarking.



Mr. Bin Chen was born in Ningbo, China, 1984, and received his bachelor degree from Shanghai Jiao Tong University in the major information security in 2007. He is currently studying as a graduate student in the Institute of Information Security Engineering, Shanghai Jiao Tong University.

His research interests are computer vision, scene recognition and machine learning.