Content-Based Video Classification Using Support Vector Machines

Vakkalanka Suresh, C. Krishna Mohan,
R. Kumara Swamy, and B. Yegnanarayana

Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai-600 036, India
{suresh,ckm,kswamy,yegna}@cs.iitm.ernet.in

Abstract. In this paper, we investigate the problem of video classification into predefined genre. The approach adopted is based on spatial and temporal descriptors derived from short video sequences (20 seconds). By using support vector machines (SVMs), we propose an optimized multi-class classification method. Five popular TV broadcast genre namely cartoon, commercials, cricket, football and tennis are studied. We tested our scheme on more than 2 hours of video data and achieved an accuracy of 92.5%.

1 Introduction

Due to significant improvement in processing technologies, network subsystems, and availability of large storage systems, the amount of video data has grown enormously in recent years. In order to make efficient use of this data it should be labeled or indexed in some manner. Also, with the advent of digital TV broadcasts of several hundred channels and the availability of large digital video libraries, it is desirable to classify and categorize video content automatically so that end users can search, choose or verify a desired program based on the semantic content thereof.

There are many approaches to content-based classification of video data. At the highest hierarchy level, video collections can be categorized into different program genres such as cartoon, sports, commercials, news and music. In a recent approach [1], Li et al used PCA to reduce the dimensionality of the features (low-level audio and visual) of video and used Gaussian Mixture Models (GMMs) to model the video classes. In an another approach [2], Truong et al used semantic aspects of a video genre such as editing, motion and color features and C4.5 decision tree algorithm to build the classifier.

At the next level of hierarchy, domain videos such as sports can be classified into different sub-categories. In [3], Xavier et al classify sports video into four sub categories (ice hockey, basketball, football and soccer) by using motion and color features and HMM based models for the video classes. In an another approach [4], by using the statistical analysis of camera motion patterns such as fix, pan, zoom and shake, sports videos are categorized into sumo, tennis, baseball, soccer and football.

N.R. Pal et al. (Eds.): ICONIP 2004, LNCS 3316, pp. 726–731, 2004.

At a finer level, a video sequence itself can be segmented and each segment can then be classified according to its semantic content. In [5], sports video segments are first segmented into shots and each shot is then classified into playing field, player, graphic, audience and studio shot categories. Parsing and indexing of news video [6] and semantic classification of basketball segments into goal, foul and coroud categories [7] by using edge-based features are some of the other works carried out at this level.

In this paper, we address the problem of video genre classification for five classes: cartoon, commercials, cricket, football and tennis. In particular, we examine a set of features that would be useful in distinguishing between the classes. We concentrate on features that can be extracted only from the visual content of a video. Although features from multiple modalities (audio, visual and text) have been reported recently in the literature, individual modalities still need to be explored. Support vector machines (SVMs) have been used to perform the classification task.

The rest of this paper is organized as follows: In Section 2, the extraction of spatial and temporal visual features inherent in a video class is described. In Section 3, the system modules for video content modeling and classification are discussed. Section 4 describes experiments on video classification on five TV genre and discusses the performance of the system. Section 5 summarizes/concludes the study.

2 Feature Extraction

A feature is defined as a descriptive parameter that is extracted from an image or a video stream. The effectiveness of any classification scheme depends on the effectiveness of attributes in content representation. We extract set of visual features such as color (features, 1 to 3), shape (features, 4 to 11), motion (feature, 12) and two other visual features (features, 13-14), that provide the discriminatory information useful for high-level classification. The definitions of these features and their intuitive meanings are discussed in the following subsections.

2.1 Color Features

Color is an important attribute for image representation. Color histogram, which represents the color distribution in an image, is one of the most widely used color feature. Since it is not feasible to consider the complete histogram, we have considered the first three moments of the color histogram. From the probability theory we know that a probability distribution is uniquely characterized by its moments. Thus, if we interpret the color distribution of an image as probability distribution, then the color distribution can be characterized by its moments as well [8]. Furthermore, because most of the information is concentrated in the lower-order moments, only the first three moments, mean, variance and skewness are used. Let N be the number of quantized colors and P_i be the number of pixels of the i^{th} color, then the first, second and third moments of the color histogram are given by

$$E = \frac{1}{N}\sum_{i=1}^{N} P_i \tag{1}$$

$$\sigma^2 = \frac{1}{N}\left[\sum_{i=1}^{N}(P_i - E)^2\right] \tag{2}$$

$$S = \frac{1}{N}\left[\sum_{i=1}^{N}\left(\frac{P_i - E}{\sigma}\right)^3\right] \tag{3}$$

The RGB(888) color space is quantized into 64 colors by considering only the 2 most significant bits from each plane. For each frame of the video sequence, color histogram is obtained with 64 bins. Then the mean, variance and skewness of the color histogram are obtained as described above.

2.2 Shape Features

Low-level shape based features can be formed from the edges in the image. A histogram of edge directions is translation invariant and it captures the general shape information in the image. Edge histogram descriptor was one of the recommended descriptors for the MPEG-7 standard content description for an image or video. We approximate the horizontal and vertical derivatives by separately filtering each from with the horizontal and vertical sobel filters. Let $\frac{\partial I(x,y)}{\partial x}$ and $\frac{\partial I(x,y)}{\partial y}$ denote the horizontal and vertical derivatives of frame I at point (x, y). Then the angle of the gradient at (x, y) is given by

$$\theta(x, y) = tan^{-1}\left(\frac{\frac{\partial I(x,y)}{\partial y}}{\frac{\partial I(x,y)}{\partial x}}\right) \tag{4}$$

The domain of the edge directions $(0 - 180)$ can be divided into 8 bins. Finally, the 8-dimensional edge direction histograms are calculated by counting the edge pixels in each direction and normalizing with the total number of pixels.

2.3 Motion Feature

Motion is an important attribute of video. Different video genre present different motion patterns. The motions in a video sequence are caused by two different sources which are camera and object(s). In this paper, we use a simple and effective technique where motion is extracted by pixel-wise differencing of consecutive frames using the equation:

$$M(t) = \frac{1}{w * h}\sum_{x=1}^{w}\sum_{y=1}^{h} P_t(x, y) \tag{5}$$

$$\text{where } P_t(x, y) = \begin{cases} 1 & \text{if } |I_t(x, y) - I_{t-1}(x, y)| > \beta \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where $I_t(x,y)$ and $I_{t-1}(x,y)$ are the pixel values at pixel location (x,y) in t^{th} and $(t-1)^{th}$ frames, respectively. β is the threshold and w and h are width and height of the image respectively.

2.4 Other Visual Features

Along with the color, shape and motion features, we have taken two other features: 1) Ratio of number of pixels with brightness greater than 0.5 to the total number of pixels and 2) ratio of edge pixels to the total number of pixels in an image.

3 Video Content Modeling and Classification

After feature extraction, the next step in video classification task is video content modeling. Many effective modeling techniques have been proposed in the literature. The effectiveness of the classification task depends on the classifier chosen. In this work, we have chosen support vector machines (SVMs) to model the video content wherein, the learning of model involves discrimination of each class against all other classes.

Support vector machines [9] for pattern classification are built by mapping the input patterns into a higher dimensional feature space using a nonlinear transformation (kernel function), and then optimal hyperplanes are built in the feature space as decision surfaces between classes. Nonlinear transformation of input patterns should be such that the pattern classes are linearly separable in the feature space. According to Cover's theorem, nonlinearly separable patterns in a multidimensional space, when transformed into a new feature space are likely to be linearly separable with high probability, provided the transformation is nonlinear, and the dimension of the feature space is high enough [10]. The separation between the hyperplane and the closest data point is called the margin of separation, and the goal of a support vector machine is to find a particular hyperplane for which the margin of separation is maximized. The support vectors constitute a small subset of the training data that lie closest to the decision surface, and are therefore the most difficult to classify.

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. In this work, we have used Gaussian kernel, since it was empirically observed to perform better than the other two. SVMs are 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.

4 Experimental Results

The experiments were carried out on more than 2 hours of video data (≈ 400 video clips each of 20 seconds captured at 25 frames per second) comprising of cartoon, commercial, cricket, football and tennis video categories. The data is collected from different TV channels on different dates and at different times to ensure the variety of data. For each video genre, half of the total number of clips were used for training and remaining half were used for testing. A Gaussian kernel based SVM was constructed for each class by using the one against the rest approach.

During testing phase, given a pattern vector to an SVM model, the result will be a measure of the distance of the pattern vector from the hyper plane constructed as a decision boundary between this class and rest of the classes. A positive value represents that the pattern belongs to the target class and vice versa. Based on the outputs from all the SVMs for a given pattern vector, the class label corresponding to the model giving the highest positive value can be assigned to the pattern vector. In order to make the decision at video clip level, two different approaches can be followed: 1) For each model, averaging the values of all pattern vectors belonging to a particular clip and assign the class label to the clip, based on which model gives the highest average positive value, and 2) count the number of positive outputs per model and assign the class label to the clip corresponding to the model with highest count. In this paper, we have used the first method, since it was empirically observed to perform better than the other. The performance of SVM based classifiers is given in Table 1.

Table 1. Confusion matrix of video classification results using SVM.

	Cartoon	**Commercial**	**Cricket**	**Football**	**Tennis**
Cartoon	90.325%	6.45%	0%	0%	3.225%
Commercial	4.1%	95.9%	0%	0%	0%
Cricket	0%	2.56%	92.31%	5.13%	0%
Football	0%	2.63%	5.26%	86.85%	5.26%
Tennis	0%	5%	5%	0%	90%

5 Conclusion

We have presented a novel approach to video classification based on color, shape and motion features using support vector machine models. A video database of TV broadcast program containing five popular genre namely cartoon, commercial, cricket, football and tennis is used for training and testing the models. A correct classification rate of 92.5% percent has been achieved. Experimental results indicate that the considered set of features can provide useful information for semantic content understanding and provide discriminating information among the classes considered. Also, it has been shown that SVM based content modelling is an effective method to bridge the gap between the low level features and the high level semantic conceptions. However, inorder to achive better classification performance, evidence from visual features alone may not be sufficient. Evidence from other modalities in a video like audio and text are to be combined with the visual evidence, which will be our future work.

References

1. Xu, L.Q., Li, Y.: Video classification using spatial-temporal features and PCA. In: Int. Conf. Multimedia and Expo, Baltimore, MD, USA (2003)
2. Truong, B.T., Venkatesh, S., Dorai, C.: Automatic genre identification for content-based video categorization. In: Proc. of Int. conf. Pattern Recognition, Barcelona, Spain (2000)
3. Gibert, X., Li, H., Doermann, D.: Sports video categorizing using HMM. In: Int. Conf. Multimedia and Expo, Baltimore, MD, USA (2003)
4. Takagi, S., Hattori, S., Yokoyama, K., Kodate, A., Tominaga, H.: Sports video categorizing method using camera motion parameters. In: Int. Conf. Multimedia and Expo, Baltimore, MD, USA (2003)
5. Assflag, J., Bertini, M., Colombo, C., Bimbo, A.D.: Semantic annotation of sports videos. IEEE Multimedia **9** (2002) 52–60
6. Wactlar, H.D., Kanade, T., Smith, M.A.: Intelligent access to digital video: Informedia project. IEEE Comput. Mag. **29** (1996) 45–52
7. Lee, M.H., Nepal, S., Srinivasan, U.: Edge-based semantic classification of sports video sequences. In: Int. Conf. Multimedia and Expo, Baltimore, MD, USA (2003)
8. Swain, M.J., Ballard, D.H.: Color indexing. Int. Journal of Computer Vision **7** (1991) 11–32
9. Vapnik, V.: Statistical Learning Theory. John Wiley and Sons, New York (1998)
10. Haykin, S.: Neural Networks - A Comprehensive Foundation. Prentice Hall (1999)