International Journal of System Signal Control and Engineering Application 4 (4-6): 69-73, 2011
ISSN: 1997-5422

Combining Audio-Video Based Segmentation and Classification Using SVM

K. Subashini, S. Palanivel and V. Ramaligam
Department of Computer Science and Engineering, Annamalai University,
608002 Chidambaram, India

Abstract: The objective in any pattern recognition problem is to capture the characteristics common to each class from features of the segmented data. Audio-video segmentation and classification can provide useful information for multimedia indexing and retrieval. In this study, researchers present a approach to segment and categorize the audio-video classification and highlighted detection. Researchers investigate the performance of Mel-frequency cepstral coefficients and color histogram in a support vector machines frame work and compare it to traditional audio-video features. Researchers achieve a correct identification closed to 96.23% on proposed method. Thus, the new technology for audio-video segmentation and classification obtained effective and efficient results compared to individual results.

Key words: Support vector machines, color histogram, audio segmentation, video segmentation, audio-video segmentation, weighted sum rule, audio classification, video classification

INTRODUCTION

With the continuous development of network and how to retrieve the requisite content of audio-video from the vast multimedia information has become an important research direction of signal/image processing. In recent era, there are many researches on the audio, video automatic classification and it generally adopts the different audio-video characteristics combined with the various classification algorithms. Although, there are many methods for the multi-classification, research shows that the classification algorithm based on Support Vector Machines (SVM) has the best performance of the several classification algorithm.

There for SVM is used to segmentation and classify different audio-video in the research. Experimental results show a good classification performance, compared with the traditional one-to-one and other algorithms. Researchers obtained automatic detection and classification technique.

Related work: Among all the models technique and features MFCC and histogram feature and SVM produced good results. Initially, a survey (Xu *et al*., 1992) of audio based music classification and annotation algorithm is obtained.

Then, Fu *et al*. (2011)'s survey on visual content based video indexing and retrieval shows huge information on video. The method described by Dhanalakshmi *et al*. (2008) uses SVM and Mel-frequency cepstral coefficients to accomplish multi group audio classification and categorization. Unsupervised speaker segmentation with residual phase and MFCC feature is given by Jothilaskmi *et al*. (2009). The technique described in Paralici *et al*. (2008) to develop a reference platform for generic audio classification. As Suresh *et al*. (2004), the researchers address the problem of video genres classification for the five classes with a set of visual features and SVM is used for classification. Huge literature reports can be obtained for automatic video classification by Suresh *et al*. (2004).

Combining the evidence obtained from complementary classifiers can improve performance based on the literature (Geetha *et al*., 2008; Xu and Li, 2003; Kittler *et al*., 1998).

MATERIALS AND METHODS

This study presents a method for audio-video classification. Figure 1 shows the block diagram of combining audio-video segmentation and classification.

Feature extraction

Audio feature: Researchers selected the MFCC features in the study because the audio segmentation and classification systems show better performance (Dhanalakshmi *et al*., 2008). MFCC feature extraction of acoustic data performance can be represented by block diagram in Fig. 2. Computation of MFCC features for audio signal is described as follows:

Corresponding Author: K. Subashini, Department of Computer Science and Engineering, Annamalai University, 608002 Chidambaram, India

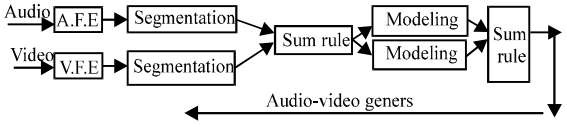

Fig. 1: Block diagram of combining audio-video segmentation and classification

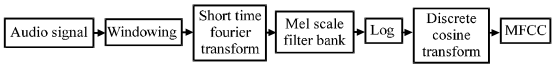

Fig. 2: Extraction of MFCC from audio signal

- Audio is represented in discrete and go through high-frequency pre-emphasis process
- The speech waveform is first windowed with analysis window and the discrete Short Time Fourier Transform (STFT) is computed
- For each frame, coefficients are obtained from fourier transform and logarithm taken
- Fourier coefficients are converted into perceptually based spectrum
- Finally Discrete Cosine Transform (DCT) is applied to the filter bank output to produce the cepstral coefficients

In this experiment, the STFT should use a hamming window and the audio signal should have first order pre-emphasis applied using a coefficient of 0.97%. The frame period is 10 msec and the window size is 20 msec to represent the dynamic information of the feature, researchers compute the 1st and 2nd derivatives and append them to original feature vector to form a 39 dimensional features are computed.

Video feature: Color histogram is used to compare images in many applications. In this study, RGB (888) color space is quantized into 64 dimensional feature vectors are used as features. The image/video histogram is a simply bar graph of pixel intensities. The pixels are plotted along the x-axis and the number of occurrences for each intensity represent the y-axis:

$$p(r_k) = n_k/n,\ 0 \le k \le L-1 \quad (1)$$

Where:

r_k = kth gray level
n_k = Number of pixels in the image with that gray level
L = Number of levels (16)
n = Total number of pixels in the image
$p(r_k)$ = Gives the probability of occurrence of gray level r_k

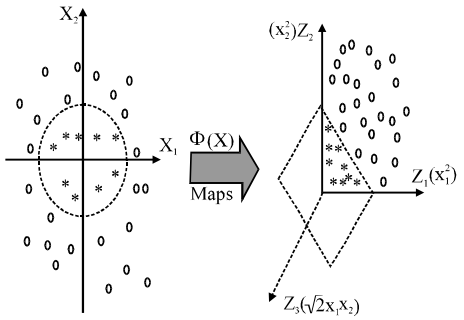

Fig. 3: Principle of Support Vector Machine (SVM)

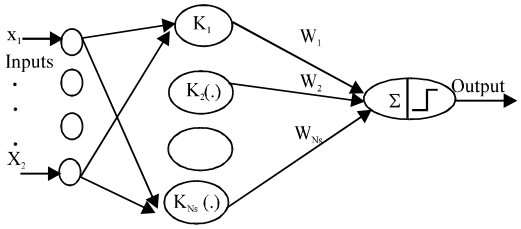

Fig. 4: Architecture of the SVM (Ns is the No. of support vectors)

Modeling technique for audio and video segmentation and classification: The basic idea is to map the input space to the higher dimensional feature space as shown in Support Vector Machine (SVM) has been used for classifying the obtained data (Burges, 1998). SVM is a supervised learning method used for classification and regression (Fig. 3). They belong to a family of generalized linear classifiers. Let us denote a feature vector (termed as pattern) by $x = x_1, x_2, \ldots\ldots x_n$ and its class label by y such that $y = \{+1,-1\}$. Therefore, consider the problem of separating the set of n-training patterns belonging to two classes:

$$(x_i, y_i),\ x_i \in R^n,\ y = \{+1, -1\};\ i = 1, 2\ldots..n \quad (2)$$

A decision function g (x) that can correctly classify an input pattern x that is not necessarily from the training set (Fig. 4).

SVM for linearly separable data: A linear SVM is used to classify data sets which are linearly separable. The SVM linear classifier tries to maximize the margin between the separating hyperplane. The patterns lying on the maximal margins are called support vectors. Such a hyperplane with maximum margin is called maximum margin hyperplane (Xu *et al*., 1992). In case of linear SVM, the discriminant function is of the form:

$$g(x) = w^t x + b \quad (3)$$

Such that $g(x_i) = 0$ for $y_i = +1$ and $g(x_i) < 0$ for $y_i = -1$. In other words, training samples from the two different classes are separated by the hyperplane $g(x) = w^t x+b = 0$. SVM finds the hyperplane that causes the largest separation between the decision function values from the two classes. Now, the total width between two margins is $j(w) = 2/w^tw$ which is to be maximized. Mathematically, this hyperplane can be found by minimizing the following cost function; subject to separability constraints:

$$J(w) = \frac{1}{2} w^t w \quad (4)$$

$$g(x_i) \geq +1 \text{ for } y_i = +1$$

or;

$$g(x_i) \leq -1 \text{ for } y_i = -1 \quad (5)$$

Equivalently, these constraints can be re-written more compactly as:

$$y_i (w^t x_i + b) \geq 1, \; i = 1, 2 \ldots . n \quad (6)$$

For the linearly separable case, the decision rules defined by an optimal hyperplane separating the binary decision classes are given in the following equation in terms of the support vectors:

$$Y = \text{sign}\left(\sum_{i=1}^{i=Ns} y_i \alpha_i (x x_i) + b \right) \quad (7)$$

Where:

Y = The outcome

y_i = The class value of the training example x and represents the inner product

The vector corresponds to an input and the vectors x_i, i = 1.... Ns are the support vectors. In Eq. 6, b and α_i are parameters that determine the hyperplane.

SVM for linearly non-separable data: For non-linearly separable data, it maps the data in the input space into a high dimension space, $x \epsilon R^I \rightarrow \Phi(x) \epsilon R^H$ with kernel function to find the separating hyperplane. A high-dimensional version of Eq. 6 is given as follows:

$$Y = \text{sign}\left(\sum_{i=1}^{i=N} y_i \alpha_i \; K(x, x_i) + b \right) \quad (8)$$

Experimental results: For conducting experiments, audio and video data are recorded using a TV tuner card from various television channels at different timings to ensure quality and quantity of data stream. The training data test includes 2-4 sec of audio stream for each mixture of dual genres, 2-4 sec of video stream for each mixture of dual genres. Audio stream is recorded at 8 kHz with mono channel and 16 bits per sample. Video clips are recorded with a frame resolution of 320×240 pixels and frame rate of 25 frames sec^{-1}.

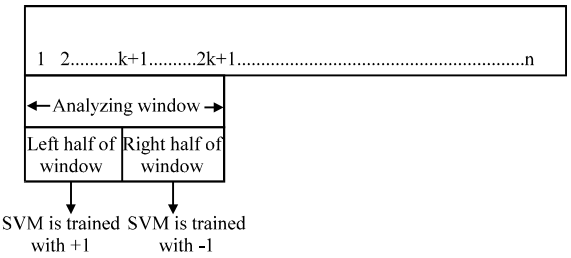

Fig. 5: Proposed alogrothm for segmentation

Combining audio-video based segmentation: The proposed audio (video) segmentation uses a sliding window of about 1 sec assuming the category change point occurs in the middle of the window. The sliding window is initially placed at the left end of the audio (video) signal. The SVM is trained to classify the feature vectors in the left half of the window and the feature vectors in the right half of the window and it is shown in Fig. 5.

The SVM is tested with all these feature vectors. A low misclassification or a high correct classification indicated a category change point such as news to advertisement because the SVM is able to discriminate the two classes. The above process is repeated by moving the window with a shift of the about 100 msec until it reaches the right end of the audio (video) signal. The audio detection and video detection results are combining using sum rule and good results obtained that is show in Fig. 6.

Audio-video segmentation: The performance of audio-video segmentation is assessed in terms of two types of error related to category (audio, video and audio-video) change detection namely false alarms Fig. 5. Proposed alogrithm for segmentation and missed detections. A false alarm namely (α) of category change detection occurs when detected category change is not a true one. A missed detection (β) occurs when true category change cannot be detected. The false alarm rate is called as precision (p) and missed detection rate is called as recall (r) are defined as:

$$p = \frac{\text{No. of correct found category changes}}{\text{Total No. of changes found}} \quad (9)$$

$$r = \frac{\text{No. of correctly found category changes}}{\text{No. of actual category changes}} \quad (10)$$

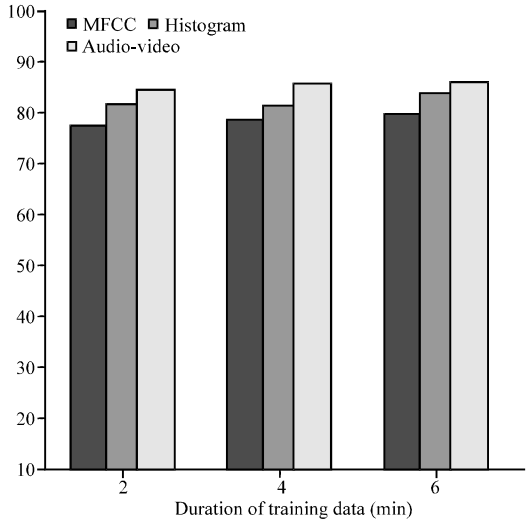

Fig. 6: Performance of SVM for audio-video based classification

In order to compare the performance of different systems, the f-measure is often used and is given by:

$$f = 2\frac{pr}{p+r} \tag{11}$$

The f measure varies from 0-1 with a higher f measure indication better performance. In the literature, the false alarms are treated as less cumbersome when compared to missing detections. Over segmentation caused by a high number of false alarms is easier to remedy then under segmentation caused by high number of missed detection. This means that the segmentation algorithms should yield a lower number of miss detections when compared to the false alarms.

To compute different matrices such as audio, video and audio-video data not exactly defined due to the presence of inter genres or environmental effects.

Combining audio-video based classification: In this study, combining the modalities has been done at the score level. The methods to combine the two levels of information present in the audio signal and video signal have been proposed.

The audio based scores and video based scores are combined for obtaining audio-video based scores as given Eq. 9. It is shown experimentally that the combined system outperforms the individual system, indicating complementary nature. The weight for each modality is decided empirically.

$$m_j = \frac{w}{n}a_j + \frac{(1-w)}{p}v_j \quad 1 \le j \le c \tag{12}$$

$$a_j = \sum_{i=1}^{n} x_i^j \quad 1 \le j \le c$$

$$v_j = \sum_{i=1}^{p} y_i^j \quad 1 \le j \le c$$

$$x_i^j = \begin{cases} 1, & \text{if } c_i^a = j \\ 0, & \text{otherwise} \end{cases} \quad 1 \le i \le n, 1 \le j \le c$$

$$y_i^j = \begin{cases} 1, & \text{if } c_i^v = j \\ 0, & \text{otherwise} \end{cases} \quad 1 \le i \le p, 1 \le j \le c$$

Where:

c_i^a = Class label for ithaudio frame
c_i^v = Class label for ith video frame
v_j = Video based score for jthframe
a_j = Audio based score for jthframe
m_j = Audio-video based score for jth frame
c = Number of classes
n = Number of audio frames
p = Number of video frames
w = Weight

The weight for each of modality is decided by the parameter w is chosen such that the system gives optimal performance for audio-video based classification. Individual fusion results (audio/video) are combined using above weighted sum rule. The effective and efficient results are obtained compared with the individual fusion results. Experimental results are obtained effectively and efficiently. The performance of SVM for audio-video based classification is shown in Fig. 5. This could also be useful for the audio-video indexing and retrieval task.

CONCLUSION

This study proposed a combined audio-video based segmentation and classification using SVM. Mel-frequency cepstral coefficients are used as features to characterize audio content. Color histogram coefficients are used as features to characterize the video content. A non-linear support vector machine learning algorithm is applied to obtain segmentation and the optimal class boundary between the various classes namely advertisement, cartoon, sports, songs by learning from training data. Experimental results show that proposed audio-video classification gives an accuracy of 87.14%.

REFERENCES

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowledge Discovery, 2: 121-167.

Dhanalakshmi, P., S. Palanivel and V. Ramaligam, 2008. Classification of audio signals using SVM and RBFNN. Expert Syst. Applied, 36: 6069-6075.

Fu, Z., G. Lu, K.M. Ting and D. Zhang, 2011. A survey of audio-based music classification and annotation. IEEE Trans. Multimedia, 13: 303-319.

Geetha, M.K., S. Palanivel and V. Ramaligam, 2008. A novel block intensity comparison code for video classification and retrivel. Expert Syst. Applied, 36: 6415-6420.

Jothilaskmi, S., V. Ramalingam and S. Palanivel, 2009. Unsupervised speaker segmentation with residual phase and mfcc features. Expert Syst. Appli., 36: 9799-9804.

Kittler, J., M. Hatef, R.P.W. Duin and J. Matas, 1998. On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell., 20: 226-239.

Paralici, J.R., M. Kuba, J. Olajec, A. Lukan and M. Dzurek, 2008. Development of reference platform for generic audio classification development of reference plat from for generic audio classification. Proceedings of the 9th International Workshop on Image Analysis for Multimedia Interactive Services, May 7-9, 2008, Klagenfurt, Austria, pp: 239-242.

Suresh, V., C.K. Mohan, R. Kumaraswamy and B. Yegnanarayana, 2004. Content-based video classification using SVM. Proceedings of the 11th International Conference on Neural Information Processing, November 22-25, 2004, Kolkata, India, pp: 726-.

Xu, L., A. Krzyzak and C.Y. Suen, 1992. Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Trans. Syst. Man, Cybern., 22: 418-435.

Xu, L.Q. and Y. Li, 2003. Video classification using spactial-temporal featuresandPCA. Int. Conf. Multimedia Expo., 3: 485 -488.