Asian Journal of Information Technology 11 (1): 30-35, 2012
ISSN: 1682-3915

Audio_Video Based Segmentation and Classification Using SVM

K. Subashin, S. Palanivel and V. Ramaligam
Department of Computer Science and Engineering, Annamali University,
Chidambaram, Tamil Nadu, India

Abstract: This study presents a method to classify audio-video data into one of seven classes: advertisement, cartoon, news, movie and songs. Automatic audio-video classification is very useful to audio-video indexing, content based audio-video retrieval. Mel frequency cepstral coefficients are used to characterize the audio data. The color histogram features extracted from the images in the video clips are used as visual features. Support vector machine is used for audio and video segmentation and classification. The experiments on different genres illustrate the results of segmentation and classifications are significant and effective. Experimental results of audio classification and video segmentation and classification results are combined using weighted sum rule for audio-video based classification. The method classifies the audio-video clips with an accuracy of 95.79%.

Key words: Support vector machines, Mel frequency cepstral coefficients, color histogram, audio classification, video classification, audio-video classification

INTRODUCTION

To retrieve the user required information in huge multimedia data stream an automatic classification of the audio-video content plays major role. Audio-video clips can be classified and stored in a well organized database system which can produce good results for fast and accurate recovery of audio-video clips. The approach has two major issues: feature selection and classification based on selected features. Recent years have seen an increasing interest in the use of SVM for audio and video classification.

RELATED WORKS

Audio classification: Recent study shows that the approach to automatic audio classification uses several features. To classify s peech/music element in audio data stream plays an important role in automatic audio classification. The method described by Dhanalakshmi *et al.* (2008) uses SVM and Mel frequency cepstral coefficients, to accomplish multi group audio classification and categorization. The method gives by Rajapakse and Wyse (2005) uses audio classification algorithm that is based on conventional and widely accepted approach namely signal parameters by MFCC followed by GMM classification. In (Kiranyaz *et al.*, 2006) a generic audio classification and segmentation approach for multimedia indexing and retrieval is described. Musical classification of audio signal in cultural style like timber, rhythm and wavelet confident based musicology feature is explained by Liu and Xie (2010). An approach given by Jiang *et al.* (2005) uses Support Vector Machine (SVM) for audio scene classification which classifies audio clips into one of five classes: pure speech, non pure speech, music, environment sound and silence.

Video classification: Automatic video retrieval requires video classification. In Brezeale and Cook (2008), surveys of automatic video classification features like text, visual and large variety of combinations of features have been explored. Video database communication widely uses low-level features such as color histogram, motion and texture. In many existing video data base management systems content-based queries uses low-level features. At the highest level of hierarchy, video database can be categorized into different genres such as cartoon, sports, commercials, news and music and are discussed by Kaabneh *et al.* (2006), Vakkalanka *et al.* (2004) and Jaser *et al.* (2004). Video data stream can be classified into various sub categories cartoon, sports, commercial; news and serial are analysis by Geetha *et al.* (2007, 2008), Brezeale and Cook (2008) and Suresh *et al.* (2004). The problems of video genre classification for five classes with a set of visual feature and SVM is used for classification is discussed by Suresh *et al.* (2004).

Corresponding Author: K. Subashin, Department of Computer Science and Engineering, Annamali University, Chidambaram, Tamil Nadu, India

FEATURE EXTRACTION

Acoustic feature: Acoustic features representing the audio information can be extracted from the speech signal at the segmental level. The segmental features are the features extracted from short (0-5 min) segments of the speech signal. These features represent the short-time spectrum of the speech signal. The short-time spectrum envelope of the speech signal is attributed primarily to the shape of the vocal tract. Mel-Frequency Cepstral Coefficients (MFCC) have been commonly used in speech processing. The computation of MFCC can be divided into 5 steps:

- Audio signal is divided into frames
- Coefficients are obtained from the Fourier transform
- Logarithm is applied to the Fourier coefficients
- Fourier coefficients are converted into a perceptually based spectrum
- Discrete cosine transform is performed

In the experiments Fourier transformation uses a hamming window and the signal should have first order pre-emphasis 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 features, the first and second derivatives are appended to the original feature vector to form a 39 dimensional feature.

Visual feature: Color histogram is used to compare images in many applications. In this research, RGB (888) color space is quantized into 64 dimensional feature vector, only the dominant top 16 values 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 (Suresh *et al*., 2004)
n = Total number of pixels in the image
$p(r_k)$ = Gives the probability of occurrence of gray level r_k

MODELING TECHNIQUE FOR AUDIO AND VIDEO CLASSIFICATION

Support Vector Machine (SVM) for classification: The basic idea is to map the input space to the higher dimensional feature space as shown in Fig. 1.

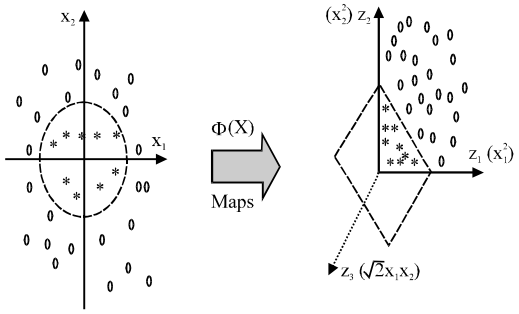

Fig. 1: Principle of support vector machine

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. 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:

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

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

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 (Vapnik, 1995). In case of linear SVM, the discriminant function is of the form:

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

Such that $g(x_i) \ge 0$ 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:

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

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:

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

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:

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

Where:

Y = The outcome

y_i = The class value of the training example x_i 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 $x \in R^L \rightarrow \Phi(x) \in R^H$ with kernel function $\Phi(x)$ to find the separating hyperplane. A high-dimensional version of Eq. 6 is given as follows:

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

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 5 min of audio stream for each genres, 5 min of video stream for each 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 per second.

Audio and video data for segmentation: The experiments are conducted using the television broadcast audio-video data collected from various channels (both Tamil and English) evaluation database. A total dataset of 50 recorded is used in this studies. This includes 10 datasets for each dual combination of dataset such as news flowed by advertisement, advertisement followed by sports ect. The audio is sampled at 8 kHz and encoded by 16 bit. Video is recorded with resolution 320×240 at 25 fbs. The category change points are manually marked. The manual segmentation results are used as the reference for evaluation of the proposed audio-video segmentation method. A total of 1,800 audio segments and 3,600 are marked in the 50 datasets. Excluding the silence periods for audio signal, the segment duration is mostly between 2-6 sec.

Feature Representation: The extraction of MFCC features is based on first pre-emphasising the input speech data using a first order digital filter and then segmenting it into 20 msec frames with an overlap of 50% between adjacent frames using Hamming window. For each frame the first 13 cepstral coefficients other than the zeroth value are used. The color histogram is obtained from the video signal using 16 order analysis. The extraction of color histogram is based on first pre-emphasising the input video data using a first-order digital filter and then segmenting it into 20 msec frames with an overlap of 50% between adjacent frames using Hamming window. The color histogram feature is computed for the entire video signal using the method described in this study. For each frame 16 samples around the highest Hilbert envelope is extracted.

Audio and video segmenatation: The tests in this experimental investigation are conducted using the procedure mentioned in this study. The procedure is applied for MFCC features and color histogram separately in order to locate the category change frames. The 200 frames analysing window size is used in the experiments. For SVM classifier the linear kernel function with the upper bound of the Lagrange multiplier C = 1 has been used. The misclassification threshold of 0.085 has been used for reliable category change detection.

Combining audio-video segmentation: The category change detection algorithm begins with the analysing window of first 200 frames with the assumption that there is a category change at the center of the window. Then the data samples on either side of the center of this window are used to train the SVM classifier and the hyperplane is obtained. Then the same samples are classified using this hyperplane. The misclassification rates are computed and compared with the threshold. The final decision has been taken that is whether the true category change is at the center of the window or not. Then, the window is shifted 80 msec to the right and the algorithm is repeated. The entire procedure is continued until the rightward window has reached the 200 frames before the end of the audio and video stream.

Audio and video classification: Performance of the proposed audio-video classification system is evaluated using the television broadcast audio database collected from various channels and various genres consists of the following contents: 100 clips of advertisement in different languages, 100 clips of cartoon in different languages, 100 clips on news, 100 clips of movie from different languages and 100 clips of songs. Audio samples are of different length, ranging from 1-5 min with a sampling rate of 8 kHz, 16 bits per sample, monophonic and 128 kbps audio bit rate. The waveform audio format is converted into raw values (conversion from binary into ASCII). Silence segments are removed from the audio sequence for further processing 39 MFCC coefficients are extracted for each audio clip as described in study.

Audio and video classification: A non-linear support vector classifier is used to discriminate the various categories. The N-class classification problem can be solved using N SVMs. Each SVM separates a c single class from all the remaining classes (one-vs-rest approach). Support vector machine is trained to distinguish MFCC features of five categories. Support vector machines are created for each category. The training data finds on optimal way to classify audio frames into their respective classes.

The derived support vectors are used to classify audio data. For testing MFCC feature vectors are given as input to SVM model and the distance between each of the feature vector and the hyperplane is obtained. The average distance is calculated for each model. The category of the audio is decided based on the maximum distance. The training data is segmented into fixed-length and overlapping frames (in the experiments we used 20 msec frames (160 samples) with 10 msec (80 samples) frame shift).

Feature representation: To obtain MFCCs, the audio signals were segmented and windowed into short frames of 160 samples. Magnitude spectrum was computed for each of these frames using Fast Fourier Transform (FFT) and converted into a set of Mel scale filter bank outputs. Logarithm was applied to the filter bank outputs followed by discrete cosine transformation to obtain the MFCCs. For each audio signal we arrived at 39 features. This number, 39 is computed from the length of the parameterized static vector (13), plus the delta coefficients (13) plus the acceleration coefficients (13). The classification results for the different features are shown in Fig. 2. From the results, researchers observe that the

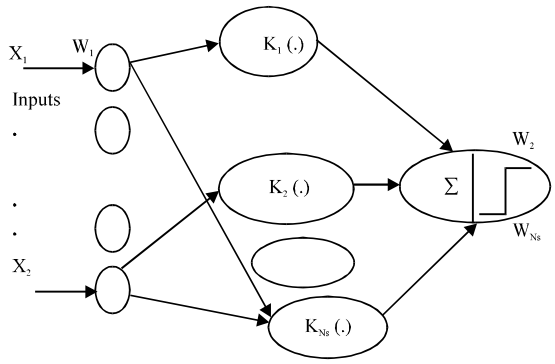

Fig. 2: Architecture of the SVM (Ns is the number of support vectors)

overall classification accuracy is 78.42% using MFCC as feature. Similarly the experiments are conducted using histogram as features in video classification.

In the experiment video streams has been recorded and digitized at a resolution of 320×240 pixels and frame rate of 25 frames per second. From television broadcast video data base 5 min video genres stream are taken for training and testing. Each frame will have 64 dimensional vectors in which top 16 values are used as features. Compared to audio classification, video classification is more complicated. The memory used for video classification is twice that used for audio classification. The classification results are shown in Fig. 2. From the results, we observe that the overall classification accuracy is 78.5% using color histogram as feature.

Combining audio and video classification: In this research, 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:

$$m_j = \frac{w}{n} a_j + \frac{(1-w)}{p} v_j \quad 1 \le j \le c \tag{9}$$

Where:

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

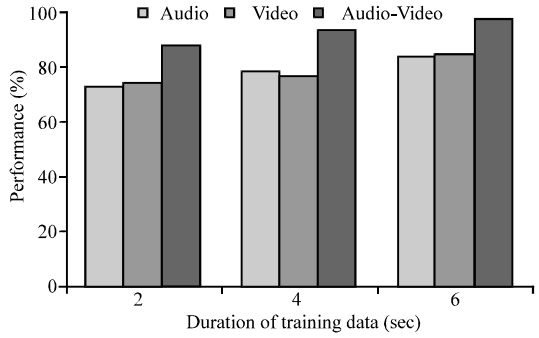

Fig. 3: Performance of SVM for audio-video based classification

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

Otherwise:

$$1 \le i \le n, \qquad 1 \le j \le c$$

Otherwise:

$$1 \le i \le p, \qquad 1 \le j \le c$$

c_i^a = Class label for ith audio frame
c_i^v = Class label for ith video frame
v_j = Video based score for jth frame
a_j = Audio based score for jth frame
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. The performance of SVM for audio-video based classification is shown in Fig. 3. This could also be useful for the audio-video indexing and retrieval task.

CONCLUSION

This study proposed an automatic audio-video based 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 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 95.79%.

REFERENCES

Brezeale, D. and D.J. Cook, 2008. Automatic video classification: A survey of the literature. IEEE Trans. Sys. Man Cybernetics-Part c: Applied Rev., 38: 416-430.

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.

Geetha, M.K., S. Palanivel and V. Ramaligam, 2007. HMM based video classification using static and dynamic features. Proceedings of the IEEE International Conference on Computational Intelligence and Multimedia Applications, December 13-15, 2007, Annamalai, Nagar, pp: 277-281.

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.

Jaser, E., J. Kittler and W. Christmas, 2004. Hierarchical decision making scheme for sports video categorization with temporal post processing. Comput. Vision Pattern Recognit., 2: 908-913.

Jiang, H., J. Bai, S. Zhang and B. Xu, 2005. SVM-based audio scene classification. Proceedings of the Natural Language Processing and Knowledge Engineering, October 30-November 1, 2005, Beijing, China, pp: 131-136.

Kaabneh, K., A. Abdullah and A. Al-Halalemah, 2006. Video classification using normalized information distance. Proceedings of the Conference on Geometric Modeling and Imaging-New Trends, August 16-18, 1993, London, England, pp: 34-40.

Kiranyaz, S., A.F. Qureshi and M. Gabbouj, 2006. A generic audio classification and segmentation approach for multimedia indexing and retrieval. IEEE Trans. Speech Audio Proces., 14: 1062-1081.

Liu, J. and L. Xie., 2010. SVM-based automatic classification of musical instruments. Proceedings of the International Conference on Intelligent Computation Technology and Automation, May 11-12, 2010, Changsha, pp: 669-673.

Rajapakse, M. and L. Wyse, 2005. Generic audio classification using a hybrid model based on GMMs and HMMs. Proceedings of the 11th International, January 12-14, 2005, Institute for Infocomm Research, pp: 53-58.

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.

Vakkalanka, S., C.K. Mohan, R. Kumaraswamy and B. Yegnanarayana, 2004. Combining multiple evidence for video classification. Proceedings of the International Conference on Intelligent Sensing and Information Processing, January 4-7, 2005, India, pp: 187-192.

Vapnik, V.N., 1995. Statistical Learning Theory. 1st Edn., John Wiley and Sons, New York, ISBN-13: 978-0471030034, Pages: 736.