A Hybrid Model used for Audio Video Classification

Puneet Thapar
DAVIET, Jalandhar City, India.
puneet.thapar90@gmail.com

Abstract

In this paper we present a system to categorize audio-video files into one of five modules: news, movie, advertisement, cartoon, and songs. Spontaneous audio-video classification is very useful to audio-video indexing, content based audio-video retrieval. MFCCs are used to distinguish the audio data. The color histogram features mined from the images in the video clips are used as graphic features. SVM (Support Vector Machine) is used for audio and video segmentation. ANN (Artificial Neural Network) is used for audio and video classification. The trials on different fields illustrate the results of segmentation and classifications are significant and effective. Trial results of audio and video segmentation or classification results are combined using weighted sum (WS) rule for audio-video based classification. Combining the features of SVM and ANN techniques, system classifies the audio-video clips with effective and efficient manner to obtain accurate result.

Key Terms— MFCC, SVM, ANN, segmentation, classification, weighted sum (WS) rule.

1. Introduction

To retrieve the user requisite information in vast audio video data stream a spontaneous classification of the audio-video content shows key role. Audio-video data can be classified and stored in a well-ordered database system, which can yield good results for fast and accurate recovery of audio-video data. Recent years have seen an increasing interest in the use of SVM and ANN for audio and video segmentation and classification.

Huge volume of digital videos is available online to the crowds. People are concerned in searching videos comprising specific material i.e. Content Based Video Retrieval (CBVR) requires video segmentation and classification [1]. There is a Lack of tools to categorize and retrieve video content. So there is a need to develop a system which classifies videos according to its content. The proposed system will classify and segmented videos based on the content of the videos. For example if a video includes of several scenes out of which few scenes are from news and few are from cartoon then while playing the video, the system will display the type of the scene being played. This will be done by first manually labeling set of video frames by user as training set by use of SVM and matching the played video features with the training set using ANN. The aim of audio-video classification system is also to integrate the audio features along with visual feature for more robust classification.

In some systems video are classified into shots and scenes using shot boundary detection technique. There are number of methods used to detect the shot boundary such as Cut, Fading, dissolves, Wavelet Transform. Main frames will be mined from shots by setting the time pause for image for instances in a shot the frame appearing after every 5 seconds will be taken as a main frame. Thus there are number of methods used for main frames extraction but very less systems use the wave variance between two frames to extract main frames as used in video classification system. Most of the systems use only graphic features such as color, text but these features some time may offer unfitting results therefore there is a need to extract more features and integrate them for robust classification. In this case if ANN and SVM is used then classification will be done efficiently and it will give correct results if number of modules are more.

2. Audio-Video Classification System

In this paper, audio and video are combined for classification using WS rule Fig. 1 shows the block diagram of classification using hybrid SVM and ANN modeling techniques.

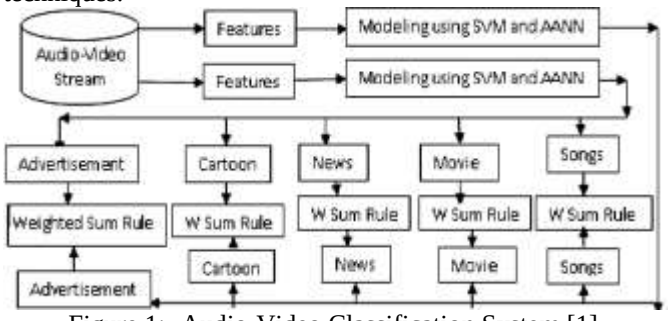

Figure 1:- Audio-Video Classification System [1]

The basic Audio-Video classification system is shown in Figure 2. This system consists of three main steps: audio-video segmentation, feature extraction and processing,

Audio-Video class classification. Here SVM used to audio-video segmentation, Feature extraction and ANN is used to classify according to the training data generated by SVM.

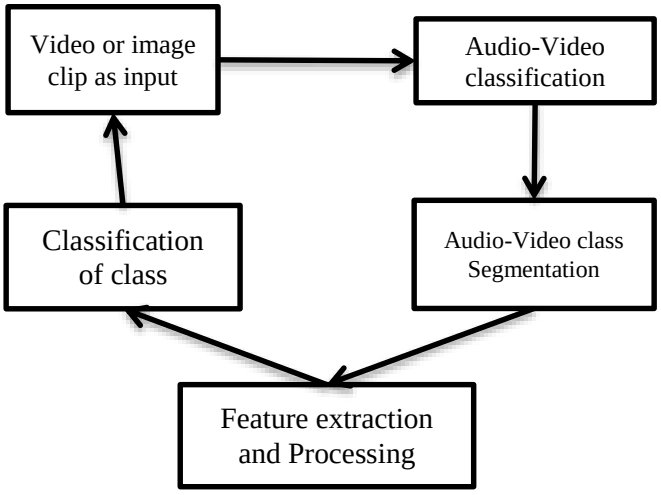

Figure 2: Audio-Video classification System

The first step in this system is to segment the incoming audio/video signal into the meaningful units that can serve as emotional classification units, such as utterance.

The goal of feature extraction and processing is to the extract relevant features from audio/video signals with respect to classes, and to reduce the size of the audio/video feature set to fewer dimensions. The widely used acoustic features indicating audio/video files are prosody features and voice quality features.

3. Feature Extraction

3.1 Audio Features

Various Acoustic features are derived from the computation of MFCC by following five steps [16].

1. Audio signal is distributed into frames.
2. Fourier transform is used to obtain Coefficients.
3. Logarithm is applied on the Fourier Coefficients.
4. Perceptually based spectrums are derived from Fourier Coefficients.
5. Then performed discrete cosine transform (DCT).

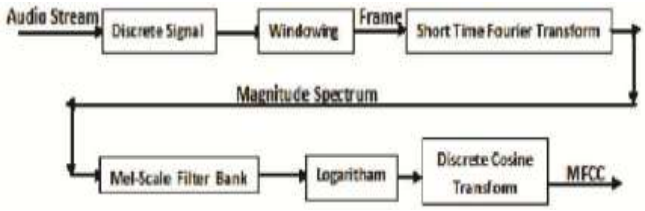

Figure 3: Extraction of MFCC from audio signal

3.2 Video Features

Color histogram is used to relate images in many applications. In this work, RGB color space is quantized into 64 dimensional feature vectors, only the leading top 16 values are used as features. The video histogram is a just bar graph of pixel strengths. The pixels are plotted along the x – axis and the number of incidents for each strength signal is signify the y-axis.[2]

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

where,

r_k – k^{th} 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.

4. SVM Classification

SVM, a binary classifier is a simple and efficient computation of machine learning algorithms, and is widely used for pattern recognition and classification problems, and under the conditions of limited training data, it can have a very good classification performance compared to other classifiers [9]. The idea behind the SVM is to transform the original input set to a high dimensional feature space by using kernel function. Therefore non-linear problems can be solved by doing this transformation.

The SVM is train according to labeled features. The SVM kernel functions are used in the training process of SVM. Binary classification can be viewed as the task of separating classes in feature space.

5. ANN Used as Audio-Video classifier

Since ANN possesses excellent discriminate power and learning capabilities [17], the hybrid classification in this paper takes advantage of a 3 hidden layer and 9 hidden nodes net to classify audio-video classes. The input of the ANN consists of distortions and likelihood probabilities, while the output is the assumed video class.

The ANN is used to capture the distribution of the input data and learning rule. Let us consider the five layers ANN model shown in Fig 4, which has three hidden layers. The processing units in the first and third hidden layers are non-linear, and the units in the second compression/hidden layer can be linear or non-linear. Five layers Artificial Neural Network (ANN) model is used to capture the distribution of the feature vectors. The second and fourth layers of the network have more units than the input layer. The third layer has fewer units than the first or fifth. The activation functions at the second, third and fourth layers are non-linear.

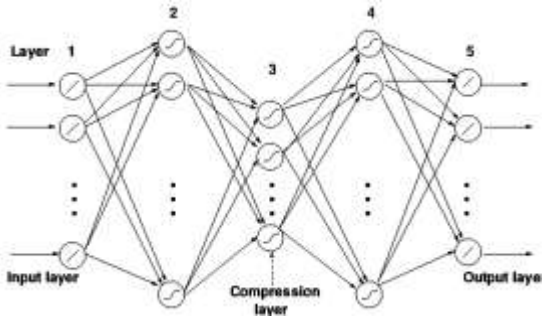

Figure 4: ANN Model

6. Experimental Study

Performance of the proposed audio-video classification system is evaluated using the Television broadcast audio database collected from numerous channels and several genres. Audio samples are of different length, ranging from 2seconds to 6seconds, with a sampling rate of 8 kHz, 16-bits per sample, monophonic and 128 kbps audio bit rate. Silence segments are removed from the audio sequence for further processing 39 MFCC coefficients are extracted. A linear support vector classifier is used to differentiate the various classes or categories. The training data is segmented into fixed-length and overlapping frames.

The distribution of 39 dimensional MFCC feature vectors in the feature space and 64 dimensional feature vectors is capture the dimension of feature vectors of each class. The acoustic feature vectors are given as input to the ANN model.

7. Audio and Video data for Segmentation

The experiments are conducted using the television broadcast audio-video data collected from various channels evaluation database [15]. A total dataset of 50 recorded is used in our studies. This includes 10 datasets for each dual combination of dataset such as news flowed by advertisement, advertisement followed by sports etc. The audio is sampled at 8 kHz and encoded by 16-bit. Video is recorded with resolution 320*240 at 25 fps. 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 to 6 seconds.

8. Conclusion and Future Scope

In this paper proposed a spontaneous audio-video based segmentation and classification using Hybrid of SVM and ANN. Mel frequency cepstral coefficients (MFCC) are used as features to characterize audio content. Color Histogram coefficients are used as features to characterize the video content. A linear support vector machine (SVM) learning algorithm is applied to obtain the optimal class boundary between the various classes namely advertisement, cartoon, sports, songs by learning from training data. An experimental result shows that proposed audio-video segmentation and classification gives effective and efficient results obtained.

REFERENCES

[1] Dhanalakshmi. P.; Palanivel. S.; and Ramaligam. V.; (2008), "Classification of audio signals using SVM and RBFNN*", In Elsevier, Expert systems with application*, Vol. 36, pp. 6069–6075.

[2] Kalaiselvi Geetha. M.; Palanivel. S.; and Ramaligam. V.; (2008), "A novel block intensity comparison code for video classification and retrivel", *In Elsevier, Expert systems with application*, Vol. 36, pp 6415-6420.

[3] Kalaiselvi Geetha, M.; Palanivel, S.; and Ramaligam, V.; (2007), "HMM based video classification using static and dynamic features", *In proceedings of the IEEE international conference on computational intelligence and multimedia applications.*

[4] Palanivel. S.; (2004)., "Person authentication using speech, face and visual speech", *Ph.D thesis, I IT, Madras*.

[5] Jing Liu.; and Lingyun Xie.; "SVM-based Automatic classification of musical instruments", *IEEE Int'l Conf., Intelligent Computation Technology and Automation (2010.),* vol. 3, pp 669–673.

[6] Kiranyaz. S.; Qureshi. A. F.; and Gabbouj. M. ; (2006), "A Generic Audio Classification and Segmentation approach for Multimedia Indexing and Retrieval"., *IEEE Trans. Audi., Speech and Lang Processing,* Vol.14, No.3, pp. 1062–1081.

[7] Darin Brezeale and Diane J. cook, Fellow. IEEE (2008), "Automatic video classification: A Survey of the literature", *IEEE Transactions on systems, man, and cybernetics-part c: application and reviews*, vol. 38, no. 3, pp. 416-430.

[8] Hongchen Jiang. ; Junmei Bai. ; Shuwu .Zhang. ; and BoXu. ; (2005)," SVM - based audio scene classification", *Proceeding of NLP-KE*, pp. 131–136.

[9] V. Vapnik.;"Statistical Learning Theory", *John Wiley and Sons*, New York, 1995. Aastha Joshi and Rajneet Kaur, "A Study of Speech Emotion Recognition Methods", International Journal of Computer Science and Mobile Computing (IJCSMC), Vol. 2, Issue. 4, April 2013, pp.28 – 31.

[10] J.C. Burges Christophe.; "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, No. 2, pp. 121–167, 1998.

[11] Rajapakse. M .; and Wyse. L.; (2005), "Generic audio classification using a hybrid model based on GMMs and HMMs ",*In Proceedings of the IEEE ,pp-1550-1555*.

[12] Jarina. R.; Paralici. M.; Kuba. M.; Olajec. J.; Lukan. A.; and Dzurek. M.; "Development of reference platform for generic audio classification development of reference plat from for generic audio classification", *IEEE Computer society, Work shop on Image Analysis for Multimedia Interactive (2008),* pp-239–242.

[13] Kaabneh,K. ; Abdullah. A.; and Al-Halalemah,A. (2006). , "Video classification using normalized information distance", In *proceedings of the geometric modeling and imaging – new trends* (GMAP06) (pp. 34-40).

[14] Suresh. V.; Krishna Mohan. C.; Kumaraswamy. R.; and Yegnanarayana. B.; (2004).,"Combining multiple evidence for video classification", *In IEEE international conference* Intelligent sensing and information processing (ICISIP-05), India (pp.187–192).

[15] Gillespie. W. J.; and Nguyen, D.T (2005).; "Hierarchical decision making scheme for sports video categorization with temporal post processing", *In Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (CVPR04) (pp. 908 -913).

[16] Suresh. V.; Krishna Mohan. C.; Kumaraswamy. R.; and Yegnanarayana. B.; (2004).,"Content-based video classification using SVM", *In International conference on neural information processing*, Kolkata (pp. 726–731).

[17] Subashini, K.; Palanivel, S.; and Ramaligam, V.; (2007), "Combining audio-video based segmentation and classification using SVM", *In International journal of Signal system control and engineering applications*, Vol.14,Issue.4, pp. 69–73.