International Journal of Computer Applications and Technology
Volume 1- Issue 2, 2012, 53-56

International Journal of Computer Applications and Technology
journal homepage: www.ijcat.com

Audio-Video Based Segmentation and Classification using AANN

K. Subashini
Dept of Comp Sci. and Engg.
Annamalai University
Chidambaram, India

S. Palanivel
Dept of Comp Sci. and Engg.
Annamalai University
Chidambaram, India

V. Ramaligam
Dept of Comp Sci. and Engg.
Annamalai University
Chidambaram, India

Abstract: This paper 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. Auto associative neural network 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 effective and efficient results obtained.

Keywords: Mel frequency cepstral coefficients, color histogram, Auto associative neural network, audio segmentation, video segmentation, audio classification, video classification, audio-video Classification and weighted sum rule.

1. 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 above approach has two major issues: feature selection and classification based on selected features. Recent years have seen an increasing interest in the use of AANN for audio and video classification.

2. Outline of the work

This work presents a method for audio-video segmentation and classification. The paper is organized as follows. The acoustic and visual feature extractions are presented in section 4. Modeling techniques for audio and video segmentation and classification are described in section 5. Experimental results of audio- video segmentation and classification are reported in section 6. Conclusion is given in section 7.

3. Related Work

Recent study shows that the approach to automatic audio classification uses several features. To classify speech/music element in audio data stream plays an important role in automatic audio classification. The method described in [1] uses SVM and Mel frequency cepstral coefficients, to accomplish multi group audio classification and categorization. The method gives in [11] uses audio classification algorithm that is based on conventional and widely accepted approach namely signal parameters by MFCC followed by GMM classification. In [6] 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 in [5]. An approach given in [8] 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.

Automatic video retrieval requires video classification. In [7], 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 in [13], [14], and [15]. Video data stream can be classified into various sub categories cartoon, sports, commercial, news and serial are analysis in [2], [3], [7] and [16]. The problems of video genre classification for five classes with a set of visual feature and SVM is used for classification is discussed in [16].

4. Feature Extraction

4.1 Acoustic Feature

The computation of MFCC can be divided into five steps.

1. Audio signal is divided into frames.
2. Coefficients are obtained from the Fourier transform.
3. Logarithm is applied to the Fourier coefficients.
4. Fourier coefficients are converted into a perceptually based spectrum.
5. Discrete cosine transform is performed.

In our 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 ms, and the window size is 20 ms. 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.

4.2 Visual Feature

Color histogram is used to compare images in many applications. In this work, 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.

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

5. Autoassociate Neural Network (AANN)

Autoassociative neural network models are feedforward neural networks performing an identity mapping. The modality would be the ablity to solve the scaling problem. The AANN is used to capture the distribution of the input data and learning rule . Let us consider the five layer AANN model shown in Fig1, 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. As the error between the actual and the desired output vectors is minimized, the cluster of points in the input space determines the shape of the hypersurface obtained by the projection onto the lower dimensional space. A five layer autoassociative neural network 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. The non-linear output function for each unit is tanh{s}, Where s is the activation value of the unit. The standard backpropagation learning algorithm is used to adjust the weights of the network to minimize the mean square error for each feature vector. The AANN captures the distribution of the input data depending on the constraints imposed by the structure of the network, just as the number of mixtures and Gaussian functions do in the case of Gaussian mixture model.

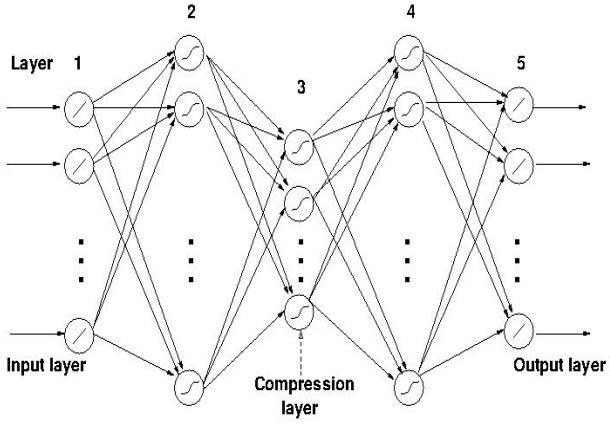

Fig.1. A Five Layer AANN model

6. Experimental Results

Performance of the proposed audio-video segementation and classification system is evaluated using the Television broadcast audio database collected from various channels and various 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. 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 Section 4.1. A non-linear support vector classifier is used to discriminate the various categories. The training data is segmented into fixed-length and overlapping frames (in our experiments we used 20 ms frames with 10 ms frame shift.)

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 AANN model and the network is trained of 100 epochs. One epoch of training is a single presentation of all training vector. The training takes about 2 mints on a pc with dual core 2.2 GHz CPU.

6.1 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 our 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 to 6 seconds.

6.2 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 ms 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 ms 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 above Section. For each frame 16 samples around the highest Hilbert envelope is extracted.

6.3 Audio and Video Segmentation

The tests in this experimental investigation are conducted using the procedure mentioned in above Section. 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 our experiments.

The proposed audio (video) segmentation uses a sliding window of about 2 sec assuming the category change point occurs in the middle of the window. The sliding window is intially placed at the left end of the audio (video) signal. The AANN model is trained to capture the distribution of the feature vectors in the left half of the window, and the feature vectores in the right half of the window are used for testing as shown in Fig.2.

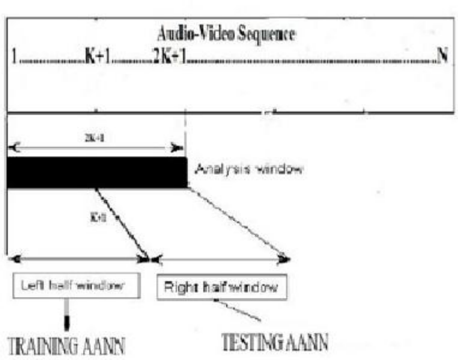

Fig.2. Proposed Segmentation Algorithm Using AANN

The out put of the model is compared with the input to compute the narmalized squared error (e_i) for the i^{th} feature vector ($\mathbf{y}_i$) is given by,

$e_k=|x_i-o_i|^2 \ / \ 2$ (9)

where o_i is the output vector given by the model.

The error e_k is transformed into a confidence score s using

$s = \exp(-e_k)$

Average confidence score is obtained for the right half of the window. A low confidence score indicates that the characteristics of the audio(video) signal in the right half of the window are different from the signal in the left half of the window, and hence, the middle of the window is a category change point. The above process is repeated by moving the window with a above progress is repeated by moving the window with a shift of the about 80msec until it reaches the right end of the audio(video) signal.

6.4 Combining Audio-Video Segmentation

The evidence from audio and video segmentation from AANN are combined using weighted sum rule. The weighted sum rule states that "`If the category change point is detected at t_1 from the audio and at t_2 is within a threshold t then the category change point is fixed at t_1+t_2/2".

6.5 Audio and Video Classification

For evaluating the performance of the system, the feature vector is given as input to each of the model. The output of the input to compute the normalized squared error. The normalized squared error E for the feature vector y is given by $E = \frac{\|y - 0\|^2}{\|y\|^2}$, where 0 is the output vector given by the model. The error E is transformed into a confidence score c using c=exp (-E). Similarly the experiments are conducted using histogram as features in video classification. Compared to audio classification, video classification is more complicated. The memory used for video classification is twice that used for audio classification. From the results, we observe that the overall classification accuracy is beterr using color histogram as feature and Mel-frequency cepstral coefficients

6.6 Combining Audio and Video Classification

In this work, 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 equation (10). It is shown experimentally that the combined system outperforms the individual system, indicating complementary nature. The weight for each modality is decided empirically. The weight for each modality is decided empirically.

$$S = \frac{w}{n}\sum_{i=1}^{n} a_j + \frac{(1-w)}{p}\sum_{i=1}^{p} v_j$$

$$1 \le j \le c \qquad (10)$$

Where

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

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

$$x_i^j = \begin{cases} 1 & \text{if } q_i = j \\ 0 & \text{otherwi} \end{cases}$$

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

$$y_i^j = \begin{cases} 1 & \text{if } q_i = j \\ 0 & \text{otherwise} \end{cases}$$

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

s -is the combinined audio and video confidence score.

s_i^a -is the Confidence score rate of the i^{th} audio frame.

s_i^v -is the Confidence score rate of the i^{th} video frame.

v_j - Video based score for j^{th} frame.

a_j -Audio based score for j^{th} frame.

m_j - Audio-video based score for j^{th} frame.

c -number of classes.

n -number of audio frames.

p -number of video frames.

w -weight.

The category is decided based on the highest confidence score various from 0 to 1. Audio and video frames are combined based on 4:1 ration of frame shifts. 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.

7. CONCLUSION

This paper proposed an automatic audio-video based segmentation and classification using AANN. 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. An experimental result shows that proposed audio-video segmentation and classification gives an effecttive and efficient results obtained.

8. 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.

[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.