Audio-video Geners Classification Using AANN

Dr. K. Subashini, Vijayanand.S
Professor, Department of Computer Science and Engineering,
Rajarajeswari college of Engineering, Bangalore-74, India.
Associate Professor ,Department of Computer Science and Engineering,
ACS college of Engineering, Bangalore-74, India.
subashinikrishnaswamy@gmail.com, kgsvanand@gmail.com

Abstract—**In this work presents a method for automatic genres classification into one of sixty-four predefine advertisements of various channels classes used in South Indian TV Broadcasting. Hierarchical Approaches are obtained effective results audio-video classification is very useful to all multimedia retrieval. Features used for categorizing audio are Mel frequency cepstral coefficients. Visual features are extracted using color histogram features in the video clips. The experiments on different Sixty-Four genres illustrate the results of classification done automatically are significant and effective. The results of audio and video confidence score are combined at each level(all six levels) using weighted sum rule for automatic audio-video based genres classification. This method genres classification using AANN systems constructed for sixty four advertisement and results obtained with an accuracy of 98 %.**

Keywords-*Autoassociative neural network (AANN), Mel frequency cepstral coefficients, Color histogram, Audio and video Classification, Audio and video Classification, Weighted sum rul*

I. INTRODUCTION

Last few decays growth of information technology and multimedia information are flooding in the form of audio, video, text and audiovisual. All advertisement broadcasters as well as commercial advertisement broadcasters are enabled with devices to easily broadcast and store multimedia data contents. This data, once huge advertisements are broadcast and stored, are not changed for any case. Manual handling of various advertisement videos are impractical for real time campaigning applications because of its increasingly huge volume of multimedia data.

Hence, it is important to have a method of automatically genres classification of multimedia data for various advertisement videos from the broad casting contents. Categories of audio and video is one important step for automatic indexing and retrieval systems. Our main objective in this paper is to confidence score audio and video classification at all the six levels are combined using weighted sum rule.

II. BACKGROUND

A. Related work

Last few decades, there have been many studies on automatic audio and video classification and segmentation using several features and techniques. In [1], a generic audio classification approach for multimedia classification and retrieval method is described. Unsupervised speaker segmentation with residual phase and MFCC features is given in [2]. The method described in [3] uses content-based audio classification and segmentation by using support vector machines. The work in [4] speech/music segmentation using entropy and dynamism features in a HMM classification framework. The technique described in [5] developed a reference platform for generic audio classification. In [6] audio classification system is proposed using SVM and RBFNN. The perceptual approach is used for automatic music genre classification based on spectral and cepstral features in [7]. A hierarchy based approach for video classification using a tree-based RBF network is described in [8]. In [9] a method is proposed for video classification using normalized information distance. Visual database can be perceptual and categorized into different genres in [10].

The technique described in [11] uses combining multiple evidences for video classification. In [12] the authors 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 in [13]. Several audio-visual features have been described in [14] for characterizing semantic content in multimedia. The edge based feature, namely, the percentage of edge pixels, is extracted from each key frame for classifying a given sports video into one of the five categories, namely, badminton, soccer, basket ball, tennis and figure skating techniques in [15]. A feature, called motion texture, is derived from motion field between video frames, either in optical flow field or in motion vector field in [16]. In [17] GMM is used to model low level audio/video feature for the classification of five different categories namely, sports, cartoon, news, commercial, and music. An average correct classification rate of 86.5% is achieved with one hour of records per genre, consisting of continuous sequences of five minutes each and 40 second decision window. Combining the evidence obtained from several complementary classifiers can improve performance based on the literature shown in [18] and [19]. In [20] a survey of audio based music classification and annotation is described. Then, in [21] a survey on visual content based video indexing and retrieval shows huge information on video. A effective algorithm for unsupervised speaker segmentation using AANN is described in [2]. In [22] a robust speaker change detection algorithm is proposed. Evaluation of indexing techniques for audio indexing is described in [23]. In [24] a hybrid approach is presented for audio

segmentation. Acoustic, strategies for automatic segmentation are described in [25]. In [26] unsupervised speaker change detection using SVM misclassification rate is described. Automatic segmentation, classification and clustering of broadcast news audio is given in [27].

B. Outline of the work

In these work two systems (audio and video) based on two(audio and video) modeling techniques (AANN) approach is used optimal class obtained boundary between the classes by learning from training audio and video data. Results of sixty four advertisement genres are classified obtain using AAN center approximates a cluster of training data vectors (audio and video) such that they are close to each other in calculating Euclidean space. Here a vector is input to the AANN, the centers that are very near to that particular vector become strongly activated, in turn activating certain output nodes used in five layers AANN architecture. Experimental results show that the two systems (audio and visual) genres classification better results obtain using hierarchical approach to get maximum accuracy of through AAN.

The paper is organized as follows: Feature extraction is presented in Section III. Modeling techniques used for classification and hierarchical approach described in Section IV. Experimental results and conclusions described in Section V, and VI, respectively.

III. FEATURE EXTRACTION FOR CLASSIFICATIONS

a. Acoustic Feature Extraction for Classifications

MFCC is perceptuallymotivatedrepresentation defined as the cepstrum of a windowed short-time signal. A non-linearmel-frequencyscaleisusedwhich approximates the behavior of the auditory system. The MFCC is based on the extraction of the signal energy with-in critical frequency bands by means of a series of triangular filters. Whose centre frequencies are spaced according to melscale. The mel-cepstrum exploits auditory principles as well as the de-correlating property of the cepstrum [2][31][32]. Fig. 1, Illustrates the computation of MFCC features for a segment of audio signal which is described as Follows: The mel-frequency cepstrum has proven to be highly effective in recognizing structure of music signalsandinmodelingthesubjectivepitchand frequency content of audio signals. Psychophysical studies have found the phenomena of the mel pitch scale and the critical band, and the frequency scale-warping to the mel scale has led to the cepstrum domain representation.The mel scale is defined as

$$F_{mel} = \frac{c \log\left(1 + \frac{f}{c}\right)}{\log(2)}$$

Where F_{mel} is the logarithmic scale of f normal frequency scale. The mel- cepstral features can be illustrated by the MFCCs, which are computed from the fast Fourier transform (FFT) power coefficients. The power coefficients are filtered by a triangular band pass filter bank. When c in (1) is in the range of 250 - 350, the number of triangular filters that fall in the frequency range 200 - 1200 Hz (i.e., the frequency range of dominant audio information) is higher than the other values of c. Therefore, it is efficient to set the value of c in that range for calculating MFCCs. Denoting the output of the filter bank by S_k (k = 1, 2, · · · ,K), the MFCCs are calculated as

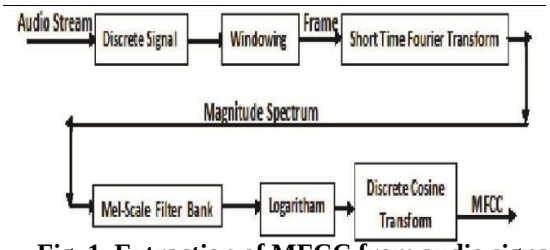

Fig. 1. Extraction of MFCC from audio signal

3

MFCCs are short-term spectral features as described above and are widely used in the area of audio and speech processing. To obtain MFCCs [2],[31],and [32]., the audio signals were segmented and windowed into short framesofn samples.Magnitudespectrumwas 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 fromthelengthofthe parameterized static vector 13, plus the delta coefficients 13, plus the acceleration coefficients

b. Visual Feature Extraction for Classificationtions

A color histogram is a representation about distribution of colors in a representation about distribution of colors in an image, derived by counting the number of pixels in each of the given set of color ranges in a typically two dimensional (2D) color space. A histogram of an image is produced first by dicretization of the colors in the image into a number of bins, and counting the number of image pixels in each bin. This is described in audio-video based segmentation and classification using SVM and AANN [31][32].

The histogram provides a compact summarization of the distribution of data in a image. The color histogram of an image is relatively invariant with transaction and rotation about the viewing axis, and may vary very slowly with the view angle. Further, they are computationally trivial to compute. Moreover, small changes in camera viewpoint has on color histograms. Hence, they are used to compare images in many applications. This work uses color histogram as visual feature. The RGB color space is quantized into 64 bins by n.

Fig.2 Color Histogram

The RGB (888) color space is quantized into 64 colors. For each frame 320*240 size, there are 64-dimensional feature vector are extracted. In order to reduce the dimension of the feature vector, only the dominant top 16 values are taken as features in experiments and is shown in Fig. 2.

IV. MODELING TECHNIQUES USED FOR CLASSIFICATION

a. Autoassociate Neural Network (AANN)

Autoassociative neural network models are feed forward neural networks performing an identity mapping. The AANN is used to capture the distribution of the input data [29], [30] and [31]. Let us consider the five layer AANN model shown in Fig. 6, 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 hyper surface obtained by the projection onto the lower dimensional space. A five layer auto associative 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 back propagation 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.

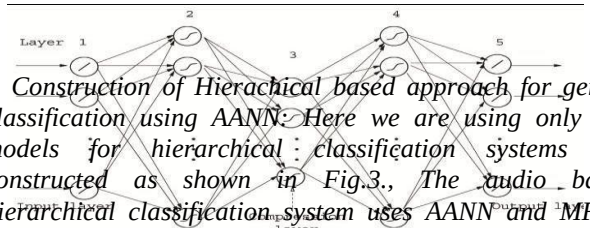

Fig. 3 A five layer AANN Model

Each box in the hierarchical classification system represents an AANN with the structure 39L 68N 14N 68N 39L for audio and 64L 128N 18N 128N 64L for video. In this structure there are 2^i auto associative neural networks in level i. For example in the last level n there are auto associative neural networks. These networks are trained using 2^n different languages sports data. In the n^{th}, , $n\text{-}1^{th}$ level, the features of two n^{th} level AANNs are combined to create an AANN as shown in Fig.4.

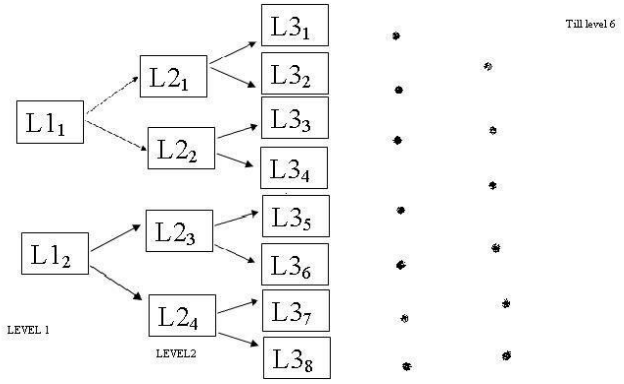

Fig.4 Hierarchical indexing system

b. Audio and Video Classification using AANN

Auto associative neural network is used to capture the distribution of the acoustic and visual feature vectors of a category. Separate AANN model are trained to capture the distribution of acoustic and visual feature vectors of each category. In testing process we used every acoustic and visual feature vector is given as input to each of the models. The output of the model is compared with the input to compute the normalized squared error. The normalized squared error is transformed into a confidence score as described in Section 4.2. Average value of confidence score is calculated for all models. The category is decided based on the maximum confidence score. This is described in audio-video based classification using AANN [31][32].

$$m_j = \frac{w}{n} a_j + \frac{1-w}{p} v_j, \quad 1 \le j \le c,$$

Where $a_j = \sum_{i=1}^{n} x_i^j$ $v_j = \sum_{i=1}^{p} y_i^j$

a

c_i - Category label for i^{th} audio frame.

c_i^v - Category label for i^{th} video frame.

vj - video based score for j^{th} category.

aj - audio based score for j^{th} category.

mj - Combined audio and video based score for j^{th} category.

c - number of categories.

n - number of audio frames.

p - number of video frames.

w - weight

The category is decided based on the maximum mj .

Similarly, the results obtained for audio and video classification by AANN are combined using:

$$s = \frac{w}{n}\sum_{i=1}^{n} s_i^a + \frac{(1-w)}{p}\sum_{i=1}^{p} s_i^v$$

Where n - number of frames in audio signal.

p - Number of frames in video signal.

s_i^a - Confidence score rate of the i[th] audio frame.

s_i^v - Confidence score of the i[th] video frame.

s - Combined audio and video confidence score.

w - Weight.

The category are decided based on the maximum confidence score obtained from the models.

V. EXPERMENTAL RESULTS

In our experiments, 64 advertisements of TV broadcasting video are recorded with a resolution of 320*240 pixels and at 8 KHz with 16-bits per sample. The LPCC, LPC, and MFCC features are extracted as comparing the features results we optimal result obtain in MFCC that is described in section 3.1 and similarly edge feature, motion features and color histogram features are extracted but overall performance of color histogram obtained effectively that is described in section 3.2.

For conducting experiments, audio and video data are recorded using a TV tuner card from various televisions south Indian language channel at different timings to ensure quality and quantity of data stream. The training data test includes various duration it can be 2, 4, 6-mins of audio stream for each genres duration various such that 2, 4, 6-mins 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. Training data is segmented into fixed overlapping frames (in our experiments we used 160 ms frames with 80ms overlapping). The sample features extraction process for repeated for audio and video data of varying durations.

A six level audio classification system is created using AANN and MFCC features. Similarly, a six level Video classification system is created using AANN and color histogram features. The number of each AANN in each level is 2, 4, 6, 8, 32 and 64 respectively.

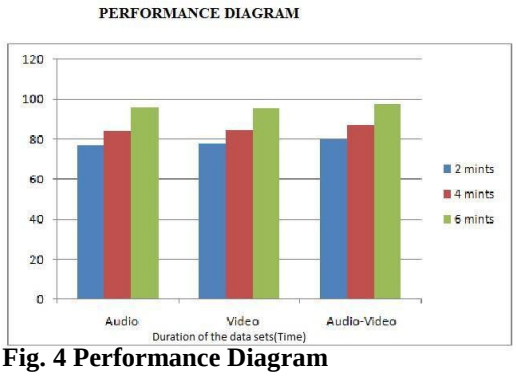

Fig. 4 Performance Diagram

For classifying a advertisement video from a test clip, the MFCC and color histogram features are extracted for the test clip and are given as input to the first level auto associative neural networks. The output of the model (O) is compared with the input to compute the normalized squared error (e_k) for the test feature vector Z is given by [31][32]

$$e_k = \frac{\|z-o\|^2}{\|z\|^2}$$

Where o is the output vector given by the model. The error e_k is transformed into a confidence score s using [31][32].

$S = exp(-e_k)$

The confidence score from the audio system and video system are combined using a weighted sum rule and the highest confidence score at each level is used to find the classification path in the genres classification system. This process is repeated for all the levels and the required various advertisement video is classified in the sixth level. The performance diagram of 100 sample of each advertisement duration various 2 mints, 4 mints and 6 mints are shown in Fig.4. Experimental results show that the proposed method achieves a classification rate of about 98.0%.

VI. CONCLUSIONS

This paper proposed a method for genres classification using AANN. Audio and video based hierarchical classification system was obtained using sixty four advertisements. Audio based hierarchical genres classification system was constructed using MFCC features and video based hierarchical classification system was constructed using color histogram features. Auto associative neural network models are used to capture the distribution of these features. The performance of the system was evaluated using confidence score of audio and video data sets and using the weighted sum rule (Proposed method) achieves a classification rate of about 98.0%.

REFERENCES

[1] S. Kiranyaz, A. F. Qureshi, M. Gabbouj, A generic audio classification and segmentation approach for multimedia classification and retrieval, IEEE Trans. Audio, Speech and Lang Processing 14(3)(2006) 1062–1081.

[2] S. Jothilaskmi, S. Palanivel, V. Ramalingam, Unsupervised speaker segmentation with residual phase and MFCC features, Expert System With Applications 36 (2009) 9799–9804.

[3] L. Lu, H.-J. Zhang, S. Z. Li, Content-based audio classification and segmentation by using support vector machines, Springer-Verlag Multimedia Systems 8 (2003) 482–492.

[4] J. Ajmera, I. McCowan, H. ourlard, Speech /music segmentation using entropy and dynamism features in a HMM classification framework, Speech Communication 40 (3) (2003) 351–363.

[5] R. Jarina, M. Paralici, M. Kuba, J. Olajec, A. Lukan, M. Dzurek, 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) 239–242.

[6] P.Dhanalakshimi, S.Palanivel, V.Ramaligam, Classification of audio signals using SVM and RBFNN, Expert System With Applications 36 (2009) 6069– 6075.

[7] C. Lin, J. Shih, K. Yn, H. Lin, Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features, IEEE Transactions Multimedia 11 (4) (2009) 670–682.

[8] W. Gillespie, D. Nguyen, Video classification sing a tree-based RBF network, IEEE International Conferance on image processing 3 (1) (2005) 465–468.

[9] K. Kaabneh, A. Abdullah, A. Al-Halalemah, Video classification using normalized information distance., In Proceedings of the geometric modeling and imaging-new trends (2005) 34–40.

[10] M. Geetha, S.Palanivel, V.Ramaligam, A novel block intensity code for video classification and retrieval, Expert System With Applications 36 (2009) 6415–6420.

[11] V. Suresh, C. K. Mohan, R. Kumaraswamy, B. Yegnanarayana, Combining multiple evidence for video classification, In IEEE international conference intelligent sensing and information processing (2005) 187–192.

[12] V. Suresh, C. K. Mohan, R. Kumaraswamy, B. Yegnanarayana, Content-based video classifica tion using SVM., In International conference on neural information processing.

[13] D. Brezeale, D. J.Cook, Automatic video classification a survey of the literature, IEEE Transaction on System, Man, and cybernetic 38 (3) (2008) 416–430.

[14] Y. Wang, Z. Liu, J. Huang, Multimedia content analysis using both audio and visual clues, IEEE Signal Process. Mag. 17 (2000) 12–36.

[15] Y. Yuan, C. Wan, The application of edge features in automatic sports genre classification., In Proceedings of IEEE Conference on Cybernetics and Intelligent Systems (2004) 1133–1136.

[16] Y.-F. Ma, H.-J. Zhang, Motion pattern based video classification using support vector machines.,In Proceedings of IEEE International Symposium on Circuit and Systems 2 (2002) 69–72.

[17] L. Q. Xu, Y. Li, Video classification using spactial-temporal features and PCA, International Conference on Multimedia and Expo 3 (2003) 345–348.

[18] J. Kittler, M. Hatef, R. Duin, J.Matas, On combining classifier, IEEE Trans. Pattern Anal. Mach. Intell. 20 (3) (1998) 226–239.

[19] L. Xu, A. Krzyzak, C. Suen, Methods of combining multiple classifiers and their applications to handwriting recognition, IEEE Trans. Syst. Man, Cybern. 2 (1992) 418–435.

[20] Z. Fu, G. Lu, K. M. Ting, D. Zhang, A survey of audio-based music classification and annotation, IEEE Transactions Multimedia 13 (2) (2011) 303– 318.

[21] H. V. Weiming, N. xie, L. Li, X. L. Zeng, S. maybank, A survey on visual content-based video classification and retrival, IEEE Transaction on System, Man, and cybernetic part c (2011) 1–23.

[22] J. Ajmera, I. McCowan, H. Bourland, Robust speaker change detection, IEEE Journal of Signal Process Letter 11 (8) (2004) 649–651.

[23] J. A. Arias, J. Pinquier, R. Ande-Obrecht, Evaluation of classification techniques for audio classification, In proc. 13th Eropean conf. Signal Processing.

[24] S. Cheng, H. Wang, Metric SEQDAC: A hybrid approach for audio segmentation, Proc. 8th International conference on spoken language Process. (2004) 1617–1620.

[25] T. Kemp, M. Schmidt, M. Westphal, A. Waibel, Acoustic strategies for automatic segmentation of audio data, Proc. IEEE International conference on Acoust, Speech, Signal Process. (2000) 1423– 1426.

[26] P. Lin, J. Wang, J. Wang, H. Sung, Unsupervised speaker change detection using SVM training misclasssification rate,IEEE Int'l Conf. Acoustics, Speech and Signal Processing 14 (3) (2006) 1062– 1081.

[27] M. Sieglar, U. Jain, B. Raj, R. Stern, Automatic segmentation, classification and clustering of broadcast news audio, Proc. DARPA Speech recognition workshop (1997) 97–99.

[28] V. Vapnik, Statistical Learning Theory, John Wiley and Sons, New York, 1998.

[29] S. Palanivel, Person authentication using speech, face and visual speech, Ph.D thesis, Indian Institute of Technology Madras, Department of Computer Science and Engg (2004).

[30] B. Yegnanarayana, S. Kishore, AANN: An alternative to GMM for pattern recognition, Neural Networks 15.

[31] K.Subashini,S. Palanivel, V. Ramaligam., Audio-video based segmentation and classification using SVM and AANN. International Journal of Computer Applications, **53(18) ,PP. 0975-8887, Sep 2012.**

[32] K.Subashini,S. Palanivel, V. Ramaligam., Audio-video based segmentation and classification using AANN. **International Journal of Computer Application and technology, Volume:1, Issue:2,PP 53-56, Sep-Oct 2012.**