International Journal of Computer Applications (0975 - 8887)
Volume 53 - No. 18, September 2012

Audio-video based Segmentation and Classification using SVM and AANN

K. Subashini
Research Scholar
Department of Computer Science and Engineering
Annamalai University,Annamalainagar-608 002, India

S. Palanivel
Professor
Department of Computer Science and Engineering
Annamalai University,Annamalainagar-608 002, India

V. Ramaligam
Professor
Department of Computer Science and Engineering
Annamalai University,Annamalainagar-608 002, India

ABSTRACT

In this paper, we propose a method for combining audio and video for segmentation and classification. The objective of segmentation is to detect the category change point such news to advertisement. The classification system classify the audio-video data into one of the predefined categories such as news, advertisement, sports, serial and movies. Mel frequency cepstral coefficients(MFCC) are used as acoustic features and color histogram is used as visual features for segmentation and classification. Support vector machine(SVM) and autoassociative neural network(AANN) models are used for segmentation and classification. The evidence from audio and video are combined using weighted sum rule for both segmentation and classifications.

General Terms:

Audio-video segmentation, Audio-video classification

Keywords:

Support vector machines(SVM),Auto associative neural network(AANN),Mel frequency cepstral coefficients,Color histogram,Audio and video segmentation,Audio and video classification,Weighted sum rule

1. INTRODUCTION

In this era of growing information technology, the information is flooding in the form of audio, video, text and audiovisual. Real time broadcasters as well as commercial broadcasters are enabled with devices to easily broadcast and store multimedia contents. This data, Once broadcast and stored, are not changed for any case. Manual handling of this data is impractical for real-time campaigning applications because of its increasingly large volume. Hence, it is important to have a method of automatically index multimedia data for targeting and commercial broad casting application based on multimedia contents. Segmentation and classification of data into different categories is one important step for building such systems. Our main objective in this paper is combining individual results audio-video segmentation and classification. Audio and video detection and categorization are emerging research area.

1.1 Related work

Last few decades, there have been many studies on automatic audio and video classification and segmentation using several features and techniques. In [13], a generic audio classification approach for multimedia indexing and retrivel method is described. An unsupervised speaker segmentation with residual phase and MFCC features is given in[10]. The method described in[17] uses content-based audio classification and segmentation by using support vector machines. The work in[2] speech/music segmentation using entropy and dynamism features in a HMM classification framework. The technique described in[9] to developed a referance platform for generic audio classification. In[20] audio classification system is proposed using SVM and RBFNN. The preceptual approach is used for automatic music genre classification based on spectral and cepstral features in[15].A hierarchy based approach for video classification using a tree-based RBF network is in[8]. In[11] a method is proposed for video classification using normalized information distance. Visual database can be preceptual and categorized into different genres in[7]. The technique described in[23] uses combining multiple evidences for video classification. In[22] 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[4]. Several audio-visual features have been described for characterzing semantic content in multimedia in [25].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, badmiton, soccer, basket ball, tennis and figure skating techniques as explained in [30]. A feature , called motion texture, is derived from motion field between video frames, either in optical flow field or in motion vector field in[18]. In [28] 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 1 hour of records per genre, consisting of continous sequences of 5 minutes each and 40 second decision window. Combining the evidence obtained from several complementary classifiers can improve performance based on the literature shown in[14] and in [27].Initially, in [6] a survey of audio based music classifica-

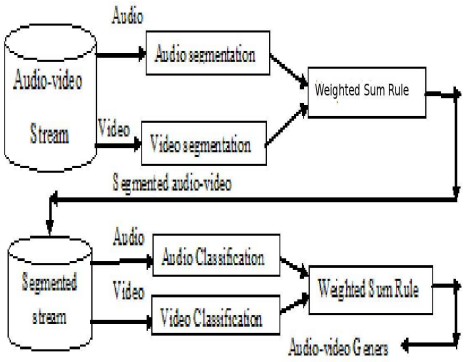


Fig. 1. Combining audio and video segmentation and classification

tion and annotation algorithm is obtained. Then, in[26] a survey on visual content based video indexing and retrieval shows huge information on video. In [31] a high-accurancy audio classification algorithm is proposed based on SVM-UBM using MFCCs as classification features. A effective algorithm for unsupervised speaker segmentation using AANN is described in [10]. In [1] a robust speaker change detection algorithm is proposed. Evaluation of classification techniques for audio indexing is described in[3]. In [5] a hybride approach is presented for audio segmentation. Acoustic, strategies for automatic segmentation are described in[12]. In [16] unsupervised speaker change detection using SVM missclassification rate is described. Automatic segmentation, classification and clustering of broadcast news audio is given in [21].

1.2 Outline of the work

In this paper, audio and video are combined for segmentation and classification. Fig. 1., Shows the block diagram of audio and video segmentation and classification. The paper is organized as follows: Acoustic feature extraction and Visual feature extraction are described in Section 2. Modeling techniques used for segmentation and classification are described in Section 3. Segmentation and classification methods are in Section 4., and 5, respectively. Experimental results are explained in section 6. Finally, conclusions is given in section 7.

2. FEATURE EXTRACTION FOR SEGMENTATION AND CLASSIFICATION

2.1 Acoustic Feature Extraction

MFCC is perceptually motivated representation defined as the cepstrum of a windowed short-time signal. A non-linear mel-frequency scale is used which approximates the behaviour 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 as shown in Fig. 2. Whose centre frequencies are spaced according to melscale. The mel-cepstrum exploits auditory principles as well as the decorrelating property of the cepstrum [10]. Fig. 3. 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 signals and in modeling the subjective pitch and frequency content of audio signals. Psychophysical studies have found the phenomena of the mel pitch

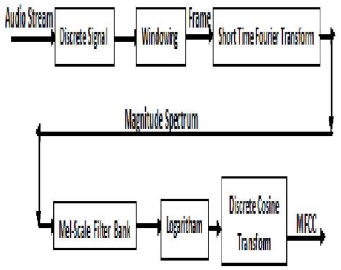


Fig. 2. Mel scale filter bank

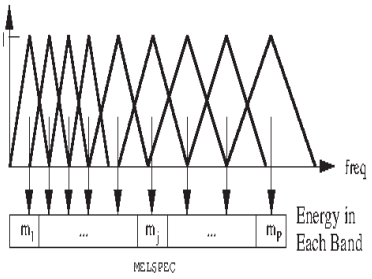


Fig. 3. Extraction of MFCC from audio signal

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

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

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 bandpass 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 $(k = 1, 2, \cdots, K)$, the MFCCs are calculated as

$$c_n = \sqrt{\frac{2}{K}} \sum_{k=1}^{K} (\log S_k) \cos\left[n\,(k - 0.5)\,\frac{\pi}{K}\right], n = 1, 2 \cdots, L \quad (2)$$

In order to evaluate the relative performance of the proposed work, we compared it with the well-known MFCC features. MFCCs are short-term spectral features as described in above and are widely used in the area of audio and speech processing. To obtain MFCCs [2], the audio signals were segmented and windowed into short frames of 256 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.

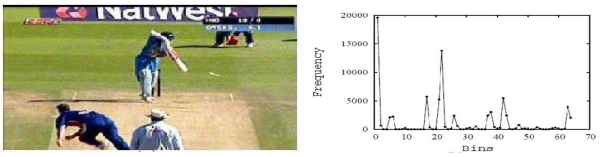
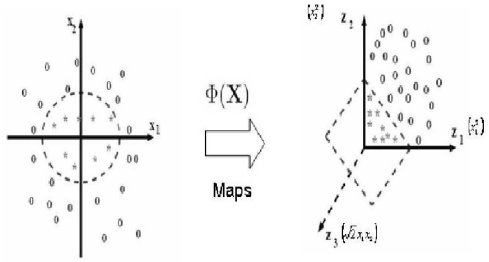


Fig. 4. An example of a Feature Extraction. (a) Input Image. (b) Color Histogram.



$$\mathbf{Z} = \Phi(\mathbf{X}) = \{x_1^2, x_2^2, \sqrt{2}x_1x_2\}$$

Fig. 5. Principle of Support Vector Machine.

2.2 Visual Feature Extraction

A color histogram is a representation about distribution of colors in a representation about distibution 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. 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. 64 bin histogram extracted from an image is shown Fig. 4.

3. MODELING TECHNIQUES USED FOR SEGMENTATION AND CLASSIFICATION

3.1 Support Vector Machine (SVM)

The SVM method is based on structural risk minimization principle and finds the best balance between the model complexity and learning ability according to the limited sample of information. The basic idea is to find the optimal separable hyper-plane that not only spearates the two classes without error, but makes the largest interval between them. SVM transforms input vectors into a high-dimensional feature space using a non-linear transformation ϕ, and then to do a linear separation in feature space as shown in Fig. 5. Support Vector Machine (SVM) can be used for classifying the obtained data (Burges, 1998). SVMs are a set of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classifiers. Let us denote a feature vector (termed as pattern) by $\boldsymbol{x}$=$(x_1, x_2, \cdots, x_n)$ 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

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

A decision function $g(x)$ that can correctly classify an input pattern x that is not necessarilly from the training set. 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 [24]. In case of linear SVM, the discriminant function is of the form:

$$g(x) = w^t x + b \tag{3}$$

such that $g(x_i) \geq 0$ 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 $\frac{2}{w^t w}$, which is to be maximized. Mathematically, this hyperplane can be found by minimizing the following cost function:

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

Subject to separability constraints

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

Equivalently, these constraints can be rewritten more compactly as

$$y_i \left(w^t x_i + b \right) \geq 1; \quad i = 1, 2, \cdots, n \tag{5}$$

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:

$$Y = sign \left(\sum_{i=1}^{i=N_s} y_i \alpha_i (x x_i) + b \right) \tag{6}$$

where Y is the outcome, y_i is the class value of the training example and x_i represents the inner product. The vector corresponds to an input and the vectors $x_i, i = 1, \ldots, N_s$, are the support vectors. In Eq. 6, b and α_i are parameters that determine the hyperplane. A non-linear support vector classifier implementing the optimal separating hyperplane in the feature space with a kernel function K(x_i, x_{new}) is given by

$$f(x_{new}) = sgn \left(\sum_{i=1}^{SV} \alpha_i y_i K(x_i, x_{new}) + b \right) \tag{7}$$

The SVM has two layers. During the learning process, the first layer selects the basis $K(x_i, x_{new})$, $i = 1, 2, \cdots, N$, from the given set of bases defined by the kernel; the second layer constructs a linear function in this space. This is completely equivalent to constructing the optimal hyperplane in the corresponding feature space. The SVM algorithm can construct a variety of learning machines by use of different kernel functions. Three kinds of kernel functions are usually used. They are

(1) Polynomial kernel of degree d

$$K(x_1, x_2) = (\gamma \langle x_1, x_2 \rangle + c_0)^d \tag{8}$$

(2) Gaussian radial basis function (RBF)

$$K(x_1, x_2) = exp \left(-\gamma \left\| x_1 - x_2 \right\|^2 \right) \tag{9}$$

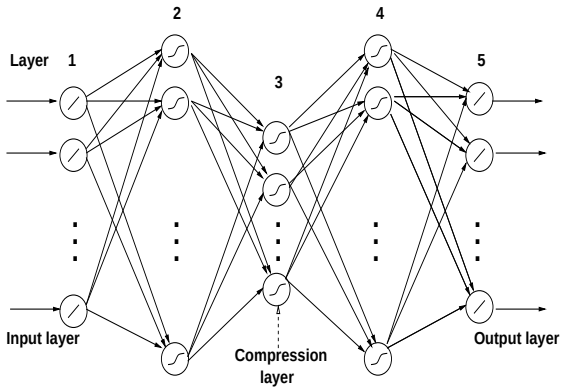


Fig. 6. A Five Layer AANN Model.

(3) Sigmoidal kernel

$$K(x_1, x_2) = \tanh\left(\gamma \langle x_1, x_2 \rangle + c_0\right) \tag{10}$$

where kernel parameters

—γ : width of RBF coefficient in polynomial
—d : degree of polynomial
—c_0 : additive constant in polynomial

3.2 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 in[19],[29]. 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 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. Gaussian mixture model.

4. AUDIO VIDEO SEGMENTATION

4.1 Audio and Video Segmentation using SVM

The proposed audi(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 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. 7. The SVM is tested with all these feature vectors. a low missclassification 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. SVM training and testing are repeated by moving the window with a shift of 80 msec until it reaches the right end of audio(video) signal.

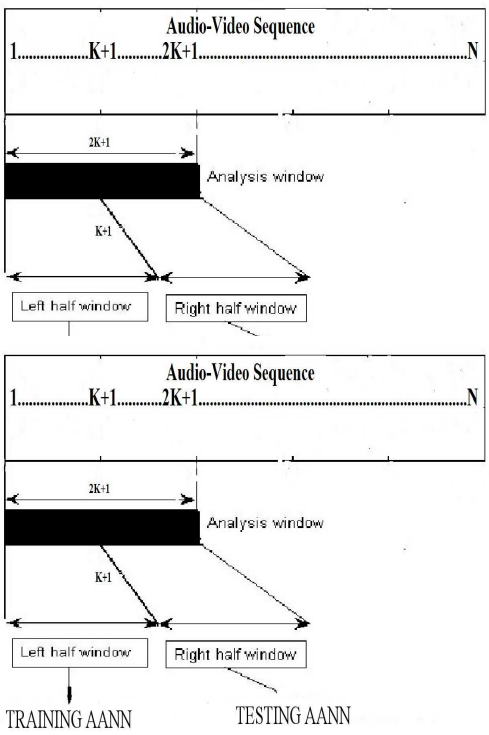


Fig. 8. AANN Based Segmentation Algorithm.

4.2 Audio and Video Segmentation using AANN

The proposed audi(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. 8. 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 ($\boldsymbol{y}_i$) is given by,

$$e_k = \frac{\|\boldsymbol{x}_i - \boldsymbol{o}_i\|^2}{\|\boldsymbol{x}_i\|^2} \tag{11}$$

where $\boldsymbol{o}_i$ is the output vector given by the model. The error e_k is transformed into a confidence score s using

$$s = \exp(-e_k) \tag{12}$$

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.

4.3 Combining Audio-Video Segmentation

The evidence from audio and video segmentation from SVM(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 $\frac{t_1+t_2}{2}$".

5. AUDIO AND VIDEO CLASSIFICATION

5.1 Audio-Video Classification using SVM

Support vector machine is trained to distingush acoustic(visual) features of a category from all other categories. One svm is created for each category. For testing, acoustic(visual) features are given as input to the svm model and the distance between each of the feture vectors and the svm hyperplane is optained. The average distance is calculated for each model. The category of audio is decided based on the maximum distance.

5.2 Audio and Video Classification using AANN

Autoassociative neural network is used to capture the distribution of the acoustic(visual) feature vectors of a category. Separate AANN model are trained to capture the distribution of acoustic(visual) feature vectors of each category. For testing, each acoustic(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 transfered into a confidence score as described in section 6.1. The average confidence score is calculated for each model. The category is described based on highest confidence score.

5.3 Combining Audio-video Classification using SVM

The evidence from audio and video classifications are combined using weighted sum rule. The audio and video classification results obtained by SVM are combined using:

$$m_j = \frac{w}{n} a_j + \frac{1-w}{p} v_j, 1 \leq j \leq c, \tag{13}$$

Where

$$a_j = \sum_{n}^{i=1} x_i^j \tag{14}$$

$$v_j = \sum_{p}^{i=1} y_i^j \tag{15}$$

$$x_i^j = \begin{cases} 1, & \text{if } c_i^a = j \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

$$y_i^j = \begin{cases} 1, & \text{if } c_i^v = j \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

c_i^a =Category label for i^{th} audio frame.
c_i^v =Category lable for i^{th} video frame.
v_j= video based score forj^{th} category.
a_j= audio based score for j^{th} category.
m_j= Combined audio and video based score for j^{th} category.
c= number of category.
n= number of audio frames.
p= number of video frames.
w= weights.
The category is decided based on the highest m_j .

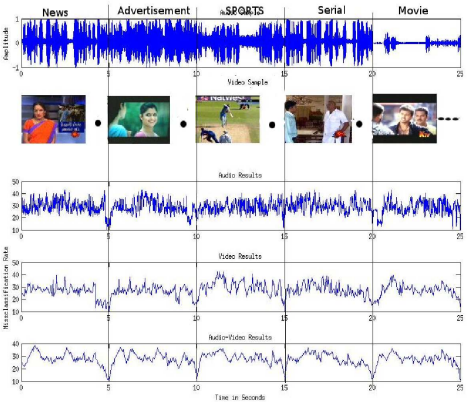


Fig. 9. Audio-Video Based Segmentation Using AANN.

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

$$s = \frac{w}{n} \sum_{i=1}^{n} s_i^a + \frac{(1-w)}{p} \sum_{i=1}^{p} s_i^v \tag{18}$$

n is the number of frames in audio signal.
p is the number of frames in video signal.
s_i^a is the confidence score rate of the i^{th} audio frame.
s_i^v is the confidence score of the i^{th} video frame.
s is the combined audio and video confidence score.
w is weight.
The category is decided based on the highest confidence score obtained from the models.

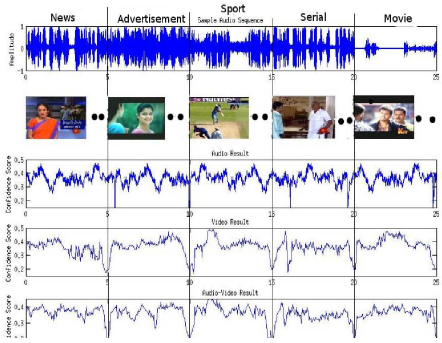
6. EXPERIMENTAL RESULTS

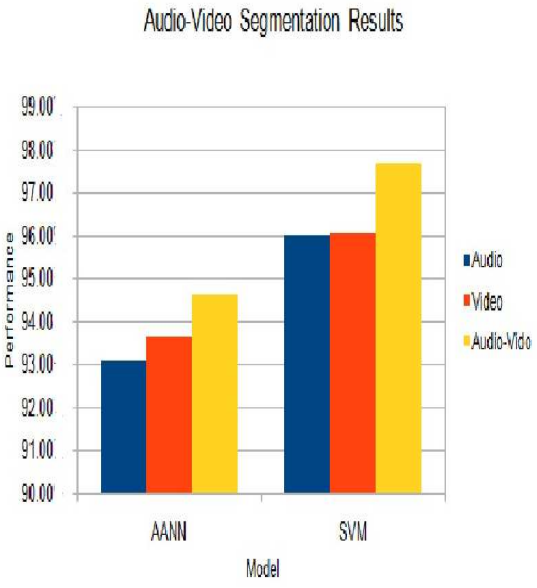
In order to test the performance of the proposed audio-video segmentation and classification system is evaluated using the TV broadcasting programs audio-video collections from various channels, comprising different durations of audio-video ranging from five seconds to one hour.The recording consits news followed by sports etc. In our work, the audio sequence should be cut into short audio segments. Multi-channel audio signals are pre sampled across multiple inputs, they are downsampling rate of 8000khz and 16 bit monophonic PCM formate. For each audio clip, features described above is extracted every 20ms, with a 30 ms overlap. This section analyzes the performance of the proposed audio-video classification in two phases. Initially, the individual segmentation of the audio and video genre are used for classification and the combined results are evaluated. For conducting experiments, video data is recorded using a TV tuner card at 25frames/s,240*320 pixels size of images from various television channels at different timings to ensure variety of data. All the geners are collected from various channels from different regional language channels.The first phase, individual frames

of the audio and video genre are used for the experiments. For segmenting visual and acoustic are obtained for all the training and testing frames. Then, weighted sum rule matching is used to find the distance between two segmentations as described in Section 4. The sample segmentation using SVM and AANN are shown in 9,. and 10., respectively. The overall segmentation performance is reported in 11. For combining the audio,video

and audio-video optained shifting the frames 4:1 ratio of audio and video frames used through out the experiment. In our work the analysing frame is set as 2 sec data set for both audio and video. The performance of the method is compared with audio

and video, and the results are given in Confusion matrix Table 2.,





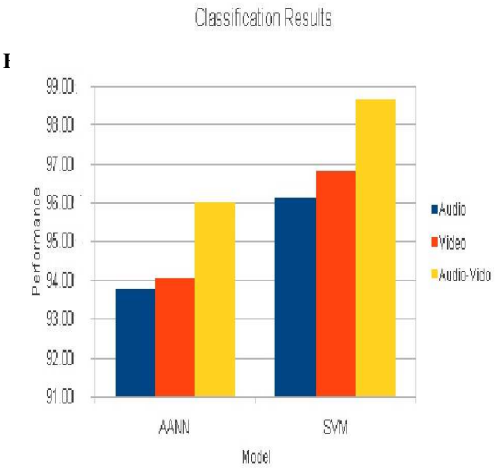


Fig. 12. Performance diagram of Audio-Video Based Classification using SVM and AANN.

and 1. The sample performance for classification is reported in 12.

7. CONCLUSIONS

In this paper, a method was proposed for combining audio and video data for segmentation and classification. In this work mel-frequency cepstral coefficients and color-histogram are used as acoustic and visual features, respectively. The Support vector machine (SVM) and autoassociative neural network (AANN) models are used for modeling the features. The evidence from acoustic and visual features are combined using weighted sum rule and it is used for both segmentation and classifications.

Table 1. Performance of proposed algorithm for audio-video classification using AANN(%)

Classifier	Modality	News	Advt.	Sports	Serial	Movie
News	Audio	94.75	0.96	1.78	1.63	0.78
	Video	94.09	1.73	1.03	1.48	1.57
	Audio-Video	96.07	1.64	0.56	0.48	1.25
Advt.	Audio	0.99	94.14	1.34	1.62	1.95
	Video	1.14	94.19	1.51	1.64	1.52
	Audio-Video	0.32	96.78	1.43	1.16	0.31
Sports	Audio	2.11	1.76	93.23	1.57	1.23
	Video	1.39	1.23	94.44	1.57	1.52
	Audio-Video	1.17	0.51	96.17	1.32	1.52
Serial	Audio	1.48	1.46	1.54	94.08	1.44
	Video	1.12	1.71	1.24	94.41	1.47
	Audio-Vide	1.62	1.24	0.82	95.37	0.97
Movie	Audio	1.18	2.02	2.06	1.73	93.00
	Video	2.12	2.71	1.27	91.78	94.00
	Audio-Video	1.16	0.79	1.49	1.13	95.63

Table 2. Performance of proposed algorithm for audio-video classification using SVM(%)

Classifier	Modality	News	Advertisement	Sports	Serial	Movie
Advt.	Audio	97.4	078	0.0	0.0	1.82
	Video	97.96	1.38	0.44	0.0	0.22
	Audio-Video	99.04	0.64	0.0	0.0	0.32
News	Audio	1.42	96.09	0.50	0.50	1.49
	Video	0.78	96.22	0.0	1.15	1.85
	Audio-Video	0.62	98.18	0.0	0.64	0.86
Sports	Audio	2.08	0.96	95.48	0.0	1.48
	Video	1.98	1.09	96.46	0.47	0.0
	Audio-Video	0.58	0.38	99.04	0.0	0.0
Serial	Audio	1.73	2.87	0.0	94.9	0.5
	Video	0.86	1.47	0.0	96.28	1.39
	Audio-Vide	0.54	0.87	0.0	97.99	0.60
Movie	Audio	1.17	1.02	0.0	1.03	96.78
	Video	0.92	0.78	0.86	0.66	96.78
	Audio-Video	0.30	0.70	0.0	0.68	98.52

8. REFERENCES

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

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

[3] J. A. Arias, J. Pinquier, and R. Ande-Obrecht. Evaluation of classification techniques for audio indexing. *In proc. 13th Eropean conf. Signal Processing*, 2005.

[4] Drain Brezeale and Diane J.Cook. Automatic video classification a survey of the literature. *IEEE Transaction on System,Man,and cybernetic*, 38(3):416–430, May 2008.

[5] S. Cheng and H. Wang. Metric SEQDAC: A hybrid approach for audio segmentation. *Proc. 8th International conference on spoken language Process.*, pages 1617–1620, Oct 2004.

[6] Zhouyu Fu, Guojun Lu, Kai Ming Ting, and Dengsheng Zhang. A survey of audio-based music classification and annotation. *IEEE Transactions Multimedia*, 13(2):303–318, April 2011.

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

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

[9] R. Jarina, M. Paralici, M. Kuba, J. Olajec, A. Lukan, and 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,*, pages 239–242, 2008.

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

[11] K. Kaabneh, A. Abdullah, and A. Al-Halalemah. Video classification using normalized information distance. *In Proceedings of the geometric modelling and imaging-new trends*, pages 34–40, 2005.

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

[13] Serkan Kiranyaz, Ahmad Farooq Qureshi, and Moncef Gabbouj. A generic audio classification and segmentation approach for multimedia indexing and retrieval. *IEEE Trans. Audio, Speech and Lang Processing*, 14(3):1062–1081, May 2006.

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

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

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

[17] Lie Lu, Hong-Jiang Zhang, and Stan Z. Li. Content-based audio classification and segmentation by using support vector machines. *Springer-Verlag Multimedia Systems*, 8:482–492, 2003.

[18] Yu-Fei. Ma and Hong-Jiang. Zhang. Motion pattern based video classification using support vector machines. *In Proceedings of IEEE International Symposium on Circuit and Systems*, 2:69–72, 2002.

[19] 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.

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

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

[22] V. Suresh, C. Krishna Mohan, R. Kumaraswamy, and B. Yegnanarayana. Content-based video classification using SVM. *In International conference on neural information processing*, 2004.

[23] V. Suresh, C. Krishna Mohan, R. Kumaraswamy, and B. Yegnanarayana. Combining multiple evidence for video classification. *In IEEE internationalconference intelligent sensing and information processing*, pages 187–192, jan2005 2005.

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

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

[26] H. V. Weiming, Nianhua xie, Li. Li, Xiang Lin Zeng, and Stephen maybank. A survey on visual content-based video indexing and retrival. *IEEE Transaction on System,Man,and cybernetic*, part c:1–23, 2011.

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

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

[29] B. Yegnanarayana and S.P. Kishore. AANN: An alternative to GMM for pattern recognition. *Neural Networks*, 15, 2002.

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

[31] R. Zhang, B. Li, and T. Peng. Audio classification based on SVM-USB. *Proc. Int. Conf. signal Processing*, pages 1586–1589, 2008.