EAI Endorsed Transactions
on Creative Technologies
Research Article EAI.EU

An Enhanced ANN-HMM based classification of video recordings with the aid of audio-visual feature extraction

Pooja Mehta[1,*], Sahil Kaswan[2] and Jaspreet Kaur[3]

[1]JCDM College of Engineering, Sirsa Haryana 125055, India

Abstract

INTRODUCTION: As an essential part of life, the use of the Internet has increased exponentially. This rising Internet bandwidth speed has made video data transmission a more popular and modern form of information exchange. For classification of video date files there is a requirement of human efforts.Also for reducing the rate of clutter in video data on Internet, a suitable automatic video classification method is required.
OBJECTIVES: In this work, we tried to find a successful model for video classification.
METHODS: To make a successful model we use different schemes of visual and audio data analysis. On the other hand we choose some music, traffic and sports videos for different analysis. The model is based on Hidden Markov model (HMM) and Artificial neural network (ANN) classifiers.In order to gather the final results, we developed an "enhanced ANN-HMM based" model.
RESULTS: Our approach attained an average of 90% success rate among all three classification classes.
CONCLUSION: In aim of this work is to categorize and caption the videos automatically.Here we proposed an enhanced HMM- ANN based classification of video recordings with the aid of audio visual feature extraction.

Keywords: Hidden Markov model, Artificial neural network, Zero Forcing Equaliser, Mel- frequency cepstral coefficient, Neural Network.

Received on 22 March 2021, accepted on 30 March 2021, published on 31 March 2021

Copyright © 2021 Pooja Mehta *et al.*, licensed to EAI. This is an open access article distributed under the terms of the Creative Commons Attribution license, which permits unlimited use, distribution and reproduction in any medium so long as the original work is properly cited.

doi: 10.4108/eai.31-3-2021.169172

*Corresponding author. Email: poojamehta0193@gmail.com

1. Introduction

Sheer number of Internet users has increased exponentially in the last few years. With an increasing number of users and modes of communications, users are spending a vast amount of time on videos. It is quite difficult, however, for humans to categorize and caption too many videos. Additionally, people need to quickly recognize what kinds of videos they are going to watch, which requires advanced techniques to do video classification and captioning. Also, in some cases, we need to block some contents. For example, parents do not want their kids to watch violent or abusive content on the Internet; for this purpose, we need such an advanced video classification technique which can identify and block that unwanted content. In this work, we have explored methods and architectures to process videos. The automatic categorization and caption are helpful for users to have better experiences when they are watching videos. The purpose of this work is to understand how to categorize and caption the video automatically. In this work, an enhanced technique of video classification has been proposed which is based on Hidden Markov model (HMM) and artificial neural network (ANN techniques) as well as the techniques known as: "ANN-HMM based classification of video recordings with the aid of audio-visual feature extraction"

2. BASIC OF ANN-HMM BASED MODEL

2.1 Artificial neural network

These are the statistical learning algorithms which are derived from the biological neural networks. The

applications of the ANN are simple and used in classification problems to speech recognition. The nodes are used as the processing elements, and the connection between the different nodes have numerical values. The nodes I the network considers different inputs and calculates the single output based on the inputs and the weights. The output is given to another neuron. The layers which lie in between the two layers, or we can say the layers between output and input are known as hidden layers. The hidden layers act as individual feature detectors and helps in recognizing more and more patterns which is complex. These algorithms give the suitable solutions for the problems which are generally characterized high dimensionality noisy and imperfect sensor data. The main advantage of the neural network is that the model of the system can be built from the available data.

2.2 Hidden Markov Model

A hidden markov model is defined as doubly stochastic process with an underlying stochastic process which is not observable but can be observed only through another set of stochastic process which produces the sequence of observed symbols. The mechanism and the elements of the HMM is that,

- There are finite numbers of states say N in the model where the signal in a state have some quantifiable properties.
- At each clock time t, a new state is entered based on the probability of the transition which depends on the previous state.
- An observation output symbol is produced according to probability distribution which depends on the current state. There are N such probability distribution which represents the random variables.

A HMM Model is defined by some set of states S = {s1, s2, . . .,sNs }, (analogous to the above described three possible weather conditions), and also a set of parameters Θ = {π, A, B}: 34
The prior probabilities πi = P(q1 = si) are the probabilities of si being the first state of a state sequence. Collected in a vector π. (The prior probabilities were assumed equi-probable in the last example, πi = 1/Ns.)
The transition probabilities are the probabilities of switching from state i to state j: ai,j = P(qn+1 = sj |qn= si). They are arranged in the matrix A.
The emission probabilities distinguished the possibility of a definite observation x, if the model is in state si . Depending on the kind of observation x we have:

- for discrete observations, xn∈ {v1, . . .,vK}: bi,k = P(xn = vk|qn = si), the probabilities to observe vk if the current state is qn = si . The numbers bi, k can be arranged in a matrix B. (For the case of weather model, with K = 2 attainable observations v1 = and v2 = .)
- for continuous observations, e.g., xn∈ R D: A set of functions bi(xn) = p(xn|qn = si) represent the probability densities (probability density functions, pdfs) over the observation space for the system being in state si . Assembled in the vector B(x) of functions. Emission pdfs are mostly configured, e.g., by mixtures of Gaussians.

The implementation of the video recordings with the help of the audio-visual feature extraction process is explained where the audio feature extraction is performed by the zero forcing equalizer integrated with Mel-Frequency Cepstral Coefficients (MFCC). The audio features which are extracted using the method is co-efficient, Delta and Delta data. The three types of video signal are considered such as music, traffic and sports and the extraction are performed for all the signals. The video signal which is separated from the input video is converted into frames and the Segmentation histogram technique is applied. The feature extraction technique used for the video signals is Grey-level co-occurrence Matrix. The features extracted which includes energy, variance, sum entropy, homogeneity, correlation, contrast, sum variance etc. and is performed for rest of the two video signals also. The audio feature signals are trained for n number of predefined inputs. After extraction of both audio features and video feature the fusion of two is combined using feature Concatenation technique. The concatenation of audio and video features is performed and the classification technique should be applied. The classification model is trained and will generate the references and the features of the test input will be compared in a similarity measures to find that the test input belongs to the class of trained model. The Artificial neural 35 network classifier is used for the fusion data where the output of the ANN is given to Hidden Markov model. The flow diagram of the model is shown below.

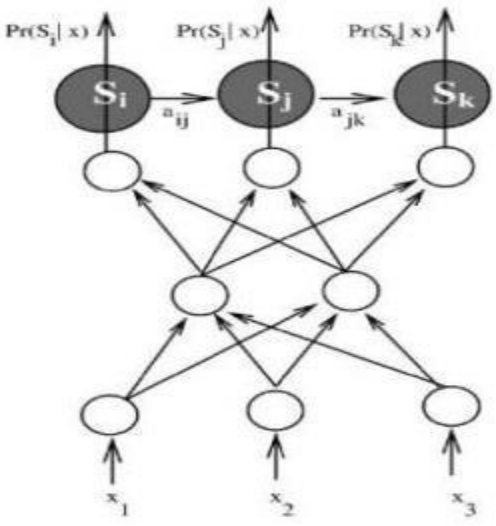

Figure 1. ANN-HMM model

The implementation of the video recordings using the ANN-HMM model is the best classification technique performed. The HMM model is evolved by using a spoken broadcast digit corpus compiled by Bell core. There are totally 9 states in HMM. For every digit, 120 exemplars are used in training and 38 exemplars reused for trials. The neural network

model considered has 16 input nodes with 10 nodes in hidden layers and 3 nodes in output layer. In case of hidden layer, different number of nodes are used for experiments.

There are two types of neural networks where the speech signals are considered such as the time delay network and length sequence which is processed at every step. The method followed is specified the inputs are shifted right for every unit delay line which links each set of inputs u units to the right adjacent and then the next input pattern is fed to the left position.perceptron for time sequence process that helps in changing the spatial sequence over corresponding units. The input layer will settled as man the recurrent neural networks. The time delay neural networks represents an effort to train a static multi-layer.

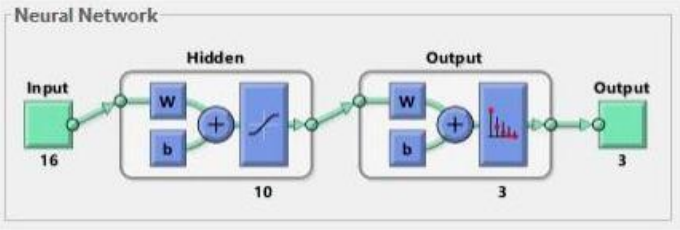

Figure 2. Neural network model

The recurrent neural networks help in providing a powerful extension of feed forward models by inserting the connections between the arbitrary pairs of units independently. The recurrent connections are reliable for an evolution of time of the internal state of the network.

3. PROPOSED ALGORITHM AND ITS WORKFLOW

In this research, the proposed a video classification method which uses audio as well as visual features of input video data for classification and the framework starts with the collection of sample video and is processing through pre-processing technique in which video and audio information is separated from the input video. The video data is converted into a set of frames followed by segmentation and extraction of critical features from both audio and video data. Later, feature level fusion process is initiated for concatenating amongst the audio, and video features and classification is achieved through a machine learning algorithm. The block diagram of the system being proposed is shown in Figure 3.

The workflow of the proposed system is illustrated in Fig. 4. Firstly, the video is extracted from the online website and is processed through the audio-video separation stage where the visual images and the audio file is extracted from the video. Initially, the video is converted into individual frames and processed through the key feature extraction stage, which is used to distinguish between the individual images and is processed through dimensionality reduction technique.

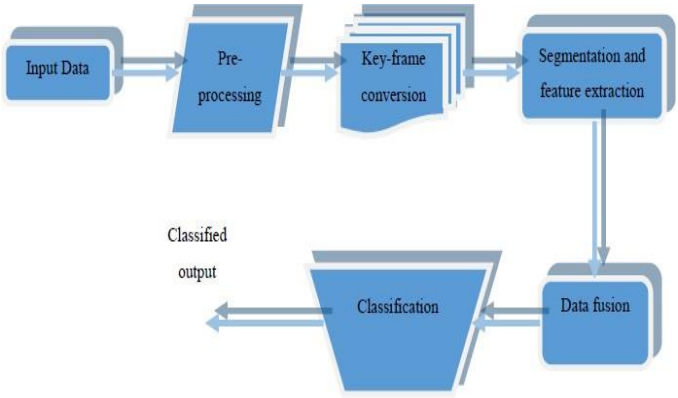

Figure 3. Block diagram of the proposed system

The histogram-based segmentation process is considered to enhance the quality of the image and to obtain the digitalized form.

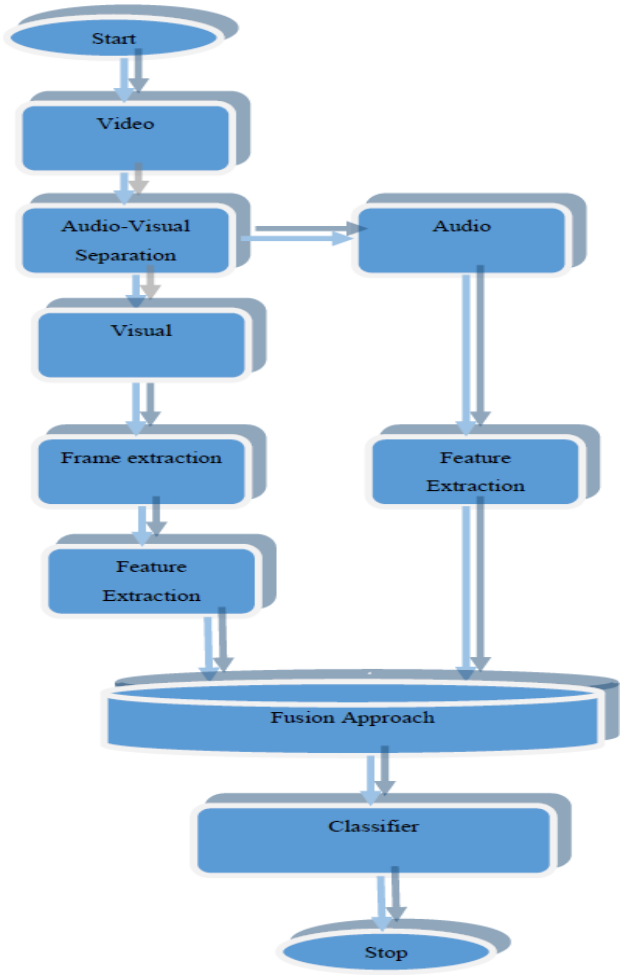

Figure 4. Workflow of the proposed system

Later, GLCM based features extraction technique is employed for extracting the useful features from video and zero forcing equalizer with Mel- Frequency Cepstral Coefficient (MFCC) for extracting the useful features. These features are later combined in the fusion stage and then processed through ANN_HMM stage to assess the performance of system. The in-depth description of the individual algorithm will be explained in the following stages as follows.

- In this research, the input data is processed through the keyframe extraction stage which is done on the basis of feature information. Color is the frequently accepted and essential feature for analyzing the input video.
- Features are stated as a function of individual or more number of measurement units in which quantifiable amount of data about the object is specified along with its significant characteristics of the video.

4. RESULTS AND DISCUSSION

The input videos considered are music, Sports and traffic. The input video is divided into 750 frames and those images are extracted using the feature extraction. The software used is MATLAB and following are the results parameters:

4.1 Confusion matrix

The confusion matrix is used in the machine learning to specify the complications of the statistical classification which is also called as error matrix and helps in visualizing the performance of a classification model on a set of test data for which the true values are known. In this the rows of the matrix represents the instances in a class while the columns represent the instances in an actual class. The confusion matrix model for the proposed classification technique is shown on next page as fig 5.

The plot shows the variation of training and testing data. This is a simple technique for the summarization of the performance of the classification algorithm. By calculating the confusion matrix, the idea of right types of errors appearing in the model is identified. The basic terms used in the confusion matrix are:

- True positives (TP): The cases which predicted yes and they do.
- True negatives (TN): The case has predicted no and they don't.
- False positives (FP): The case was predicted yes but they actually don't.
- False negative (FN): The case was predicted no, but they actually do.

From the confusion matrix the list of rates that are predicted are,

- Accuracy
- Misclassification rate
- True positive rate
- False positive rate
- Specificity
- Precision
- Prevalence
- Null error rate
- Positive prediction value
- ROC curve

Figure 5. Confusion matrix for the proposed mode

4.2 ROC curve

ROC curve gives the ability to assess the performance of the classifier over the entire operating range. The neural networks are the classifier where these curves are used as shown in fig 6..
The plotting of the true positive rates against the false positive rates are performed at various threshold values. The above figure shows the graph of the three video signals and is classified as class 1, class 2, class3.The comparison of the two operating characteristics is performed as the criterion changes. The above figure shows that the class 1 has accuracy of almost 100% and the class B has accuracy of 90% whereas class c has accuracy of 99%.

4.3 HMM output

The below figure7 shows Hidden Markov model graph where the parameters such as emission and transition are compared. As the transition rate increases the emission rate increases gradually and shows a sudden increase at some point and remains constant after few transitions.
The table 1.1 shows the complete values which is obtained after the simulation process and gives the accuracy achieved. In the above table, considering Music Video-2, Sports video-2, and Traffic Video-2 we achieved 96.5777 % accuracy. So we conclude that 96.5777 5% data is correctly classified and the Sensitivity is 100%.

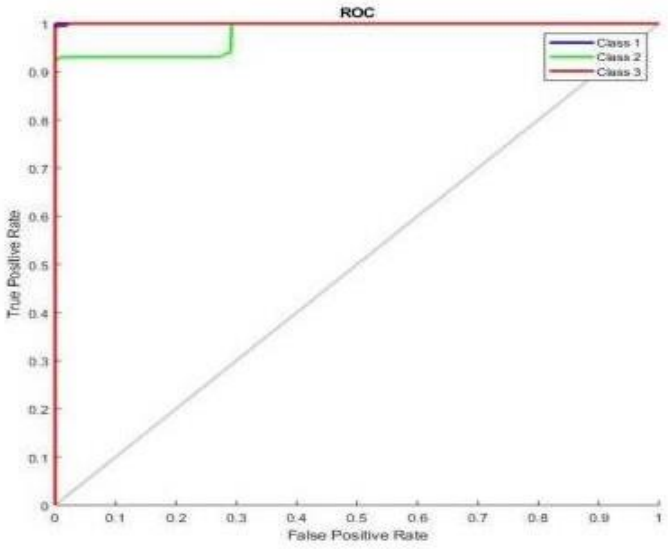

Figure 6. ROC output curve

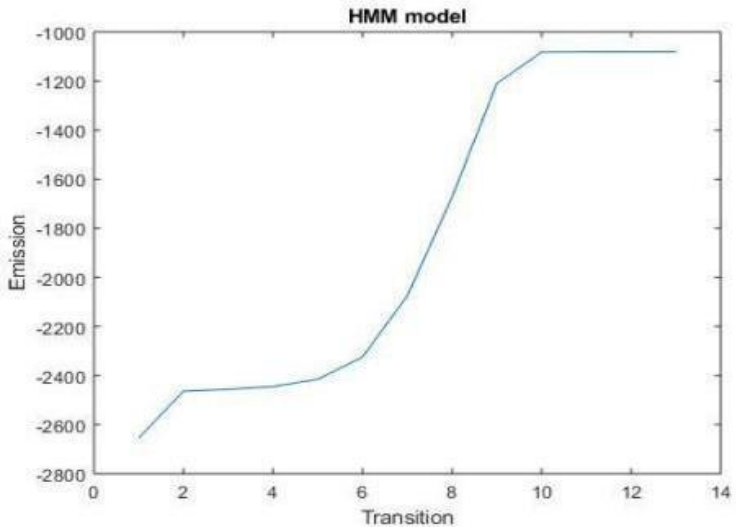

Figure 7. Output of HMM

Table 1. Performance Evaluation Table

Video Name	M2 S2 T2 (music video 2, Sport video2, Traffic Video 2)	M3 S3 T3 (music video 3, Sport video3, Traffic Video 3
Accuracy	96.5777	88.5777
Sensitivity	100	87.6000
Specificity	94.8666	89.0666
Precision	90.6892	80.0243
Recall	100	87.6000
Freq_Measurement	95.1173	83.6409
Gmean	97.3995	88.330

If we consider Music Video-3, Sports video-3, and Traffic Video-3 we can achieve 88.5777 % accuracy. So we can conclude that 88.5777% data is correctly classified. Sensitivity is 87.6000. Thus the results obtained shows that the accuracy obtained in different types of video has maximum accuracy of 97% where the video is classified.

5. CONCLUSION

With an increasing number of users and modes of communications, users are spending an amount of time on videos. However, it is quite difficult for human beings to categorize and caption too many videos. So, videos which we are going to watch require advanced techniques to do video classification and captioning. Also, in a few cases, we need to block some contents. Like, parents don't want their kids to watch violent or abusive content on the internet and for this purpose, we need such an advanced video classification technique which can find and block that unwanted content. In this work, we have explored methods and architectures to understand videos. The automatically categorization and caption are helpful for users to have better experiences when watching videos. The aim of this work is to know how to categorize and caption the video automatically. We have proposed an enhanced HMM- ANN based classification of video recordings with the aid of audio-visual feature extraction. The results evaluated and analysed shows the better accuracy when compared with the traditional. The obtained accuracy of 97% by the ANN-HMM algorithm shows that the classifier used for classifying the type of video based on the categories used.

References

[1] M. Dawood and M. Ghanbari, "Scene content classification from MPEG coded bit stream," in Proc. 1999 IEEE 3rd Workshop on Multimedia Signal Processing, pp. 253– 258, Copenhagen, Denmark, September 1999.

[2] Aizawa, K., Nakamura, Y., & Satoh, S. I. (Eds.). (2004). Advances in Multimedia Information Processing-PCM 2004: 5th Pacific Rim Conference on Multimedia, Tokyo, Japan, November 30-December 3, 2004, Proceedings (Vol. 3332). Springer.

[3] Anjum, N., & Cavallaro, A. (2010). Trajectory clustering for scene context learning and outlier detection. In Video search and mining (pp. 33-51). Springer, Berlin, Heidelberg.

[4] Araújo, C. S., Magno, G., Meira, W., Almeida, V., Hartung, P., & Doneda, D. (2017, September). Characterizing videos, audience, and advertising in Youtube channels for kids. In International Conference on Social Informatics (pp. 341-359). Springer, Cham.

[5] Bahatti, L., Bouattane, O., Echhibat, M. E., & Zaggaf, M. H. (2016). An Efficient Audio Classification Approach Based on Support Vector Machines. international journal of advanced computer science and applications, 7(5), 205-211.

[6] Barbancho, A. M., Tardón, L. J., López-Carrasco, J., Eggink, J., & Barbancho, I. (2015). Automatic classification of personal video recordings based on audiovisual features. Knowledge-Based Systems, 89, 218-227.

[7] Camastra, F., Vinciarelli, A., & Yu, J. (2009). Machine learning for audio, image and video analysis. Journal of Electronic Imaging, 18(2), 029901-029901. 50

[8] Casey, M. A. (2002). Sound• Classification and Similarity. Introduction to MPEG-7: Multimedia Content Description Interface, 309-317.

[9] C. W. Ngo, T. C. Pong, H. J. Zhang, and R. T. Chin, "Motion characterization by temporal slice analysis," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 768–773, HiltonHead Island, SC, USA, June 2000.

[10] Chougule, S. V., & Chavan, M. S. (2014). Channel Robust MFCCs for Continuous Speech Speaker Recognition. In Advances in Signal Processing and Intelligent Recognition Systems (pp. 557-568). Springer, Cham.