Audio-Based Action Scene Classification Using HMM-SVM Algorithm

Khin Myo Chit, K Zin Lin

Abstract— **Nowadays, there are many kind of video such as educational movies, multimedia movies, action movies and scientific movies etc. The careful analysis and classification of movies are very important. For example, movies containing violent or profanity by putting a separate class as they are not suitable for children can cut and avoid from movies for watching. This system is proposed to provide indexing and retrieving the most interesting and important events of the action movie to attract the viewer by using the classified audio categories. The propose system is extracted audio features and make the model by using audio feature vector and classify the audio class to detect and recognize video scenes. In this proposed system, SVM is combined with HMM based on audio features which often directly reflects while image information may not yield useful "action" indexes, to be in detection scene by labeling happy scene, miserable scene and action scene etc. Sound event types are classified including gunshot, scream, car-breaking, people talking, laughter, fighting, shouting and crowd background.**

Index Terms—— **Audio Indexing, Feature Extraction, Hidden Markov Model (HMM), Support Vector Machine (SVM).**

I. Introduction

Due to the advances of information technology, more and more digital audio, images and video are being captured, produced and stored. There are strong research and development interests in multimedia database in order to effectively and efficiently use the information stored in these media types. The research effort of the past few years has been mainly focused on indexing and retrieval of digital images and video.

In video retrieval, the most common use of audio information is for automatic speech recognition and the subsequent use of the generated transcript for text retrieval. However, the audio information can also be used, more directly, to provide additional information such as the gender of the speaker, music and speech separation and audio textures such as fast speaking sports announcers. So that the applications describing the content and the applications using the corresponding descriptions can interoperate, it is necessary to define a standard that specifies the syntax and semantics of these multimedia descriptions.

Moreover, classification of video clips will help to solve the problem of managing and accessing huge amount of video data in these days. A video sequence is a rich information source, containing audio, and text and image objects. Although the human being can understand the semantic content by fusing the information from different modalities, computer understanding of video sequence is still in a quite poor state. In the past several years most of the research has focused on the use of speech and image information. These include the use of speech recognition and language understanding techniques to produce keywords for each video frame or a group of frames.

However, visual features may cause semantically unrelated shots to be clustered into one unit only because they may be similar. Recently several researchers have started to classify video scene by using audio signal because audio-based analysis requires significantly less computation and the audio information contain a number of clues about semantics of the video data.

In this paper, a system is proposed to provide the video scene annotation by classifying the audio classes in action movies. The best audio features are selected for audio classes by modeling HMM and SVM classifier is used to categorize the mixed types of audio. This approach is to provide scene recognition based on separating audio class and to get the better performance of video scene detection. In addition it is to obtain better accuracy for sound classification.

The rest of this paper is organized as follows. In Section II, we present the related work and Section III describes how an audio clip is represented by low level perceptual and cepstral feature and gives and overviews of linear, kernel SVM and HMM. In Section IV, proposed algorithm is explained for classification and in Section V, experimental evaluation is presented. Finally, in Section VI, we conclude for the proposed system.

II. RELATED WORK

The classification of audio signals using SVM and RBFNN was proposed by P. Dhanalakshmi and S. Palanivel [1]. Linear predictive coefficients, linear predictive cepstral coefficients and mel-frequency cepstral coefficients audio features are calculated as features to characterize audio content. Support vector machines are applied to classify audio into their respective classes by learning from training data. Then the method extends the application of neural network (RBFNN) for the classification of audio. S. Jain and R.S.

Manuscript received April, 2013.

K. M. Chit *is with the University of Technology, Yatanarpon Cyber City, PyinOoLwin, Myanmar.*

K. Z. Lin *was with the University of Technology, Yatanarpon Cyber City, PyinOoLwin, Myanmar. She is now with the Department of Hardware Technology, University of Computer Studies, Yangon, Bahan Campus, Myanmar.*

Jadon focused on neural net learning based method for characterization of movies using audio information [3]. They

characterized the movie clips into action and non-action. In [4] they performed an empirical feature analysis for audio environment characterization and proposed to use the matching pursuit (MP) algorithm to obtain effective time–frequency features. The MP-based method utilizes a dictionary of atoms for feature selection, resulting in a flexible, intuitive and physically interpretable set of features. S. Gao Ma and W. Wang presented the discriminating fighting shots in Action Movies by using the camera motion and SVM classifier. Fighting shots are the highlights of action movies and it is useful for many applications [5]. Zhang and Kuo introduced a method for automatic segmentation and classification of audiovisual data based on audio features [6]. Classification is performed for discriminating between basic types of sounds, including speech with or without music background, music, song, environmental sound with or without music background, silence, etc. This model does not rely on the learning algorithm, but rather uses rule-based heuristics to segment and classify the audio signals. They utilize a “back-to-back” window scheme to analyze and compare audio features such as energy, zero-crossing and fundamental frequency for segmentation and classification of the segmented events.

V. Elaiyaraja and P. Meenakshi presented audio classification system by using audio features and a frame-based multiclass support vector machine [7]. In feature selection, this study transforms the log powers of the critical-band filters based on independent component analysis (ICA). This audio feature is combined with linear prediction coefficients (LPC) – derived cepstrum coefficients (LPCC), Mel Frequency Cepstral coefficients (MFCCs) perceptual features to form an audio feature set.

III. Theoretical Background

A. Feature Extraction

One of the most important parts of automated audio classification is the choice of features or properties. Features serve as the input to pattern recognition systems and are the basis upon which classifications are made. Most audio classification systems combine two processing stages: feature extraction followed by classification. The following audio features, described in detail below, are based on time domain and frequency domain.

In this system, audio clip-level features are computed based on the frame-level features and used a clip as the classification unit. For features such as zero-crossing rate (ZCR), short-time energy (STE), volume root mean square (VRMS) and volume dynamic range (VDR), means of all frames in a given clip is computed as basic clip-level features which are proved to be effective for distinguishing speech, music and crowd background [2]. The mathematical representations of these features are described as (1) to (4).

$$ZCR = \frac{1}{2(N-1)}\sum_{m=1}^{N-1}\left|\operatorname{sgn}[x(m+1)] - \operatorname{sgn}[x(m)]\right| \qquad (1)$$

Where sgn[.]is a sign function and x(m) is the discrete audio signal, m = 1…..N.

$$STE(m) = \sum_{m}(x(n)W(n-m))\hat{}\,2 \qquad (2)$$

where m is the time index of the short time energy, x(n) is the discrete time audio signal, W(n) is the window (audio frame) of length N where n = 0,1,2,…..,N-1.

Volume Root Mean Square (VRMS): VRMS of the nth frame is calculated, by the following formula:

$$VRMS(n) = \sqrt{\frac{1}{N}\sum S_n^2(i)} \qquad (3)$$

where Sn(i) is the ith sample in the nth frame audio signal, and N is the total number of samples in the frame.

In audios with action in the background, volume of the frame does not change much, while in non-action audios, there are silent period between the speeches, and hence VDR is expected to be higher. VDR is calculated as

VDR = [(MAX(v)-MIN(v))/MAX(v)] (4)

Where MIN(v) and MAX(v) represent the minimum and maximum volume within a clip respectively.

MFCC is one of the most popular feature extraction techniques used in audio classification, whereby it is based on the frequency domain of Mel scale for human ear scale.

B. Hidden Markov Model

HMM has shown to be powerful statistical tool in speech processing. The features extracted from the test’s video are considered to be a sequence of events and then used as the input for the HMM. It can automatically find the temporal pattern of video scene streams. It represents a set of states and the probabilities of making a transition from one state to another state. The typical usage in video classification is to train one HMM for each class.

C. Support Vector Machine

SVM models the boundary between the classes instead of modelling the probability density of each class (Gaussian Mixture, Hidden Markov Models). SVM algorithm is a classification algorithm that provides state-of-the-art performance in a wide variety of application domains. There are two main reasons for using the SVM in audio classification.

First, many audio classification problems involve high dimensional, noisy data. The SVM is known to behave well with these data compared to other statistical or machine learning methods.

Second, the feature distribution of audio data is so complicated that different classes may have overlapping or interwoven areas. However, a kernel based SVM is well suited to handle such as linearly non-separable different audio classes. The classifier with the largest margin will give lower expected risk, i.e. better generalization

SVM transforms the input space to a higher dimension feature space through a nonlinear mapping function. Construct the separating hyper plane with maximum distance from the closest points of the training set.

Consider the problem of separating a set of training vectors belonging to two separate classes, (x1; y1), . . . , (xl;yl), where xi Є Rn is a feature vector and yi Є {−1, +1} is a class label, with a separating hyper-plane of equation w·x+bv = 0; of all the boundaries determined by w and b. On the basis of this rule, the final optimal hyper-plane classifier can be represented by the following equation:

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{l} \overline{\alpha} y_i x_i x + \overline{b}) \quad (5)$$

where α and b are parameters for the classifier; the solution vector xi is called as Support Vector with αi being non-zero. In the linearly non-separable but non-linearly separable case, the SVM replaces the inner product by a kernel function K(x,y), and then constructs an optimal separating hyper-plane in the mapped space.

According to the Mercer theorem, the kernel function implicitly maps the input vectors into a high dimensional feature space in which the mapped data is linearly separable. In this method, the Gaussian Radial Basis kernel will be used:

$$K(x_i, x_j) = \exp(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2} \quad (6)$$

where the output of the kernel is dependent on the Euclidean distance of xj from xi (one of these will be the support vector and the other will be the testing data point). The support vector will be the center of the RBF and σ will determine the area of influence this support vector has over the data space.

IV. Proposed System

Recent advances in multimedia compression technology, the significant increase in computer performance and the growth of Internet, have led to the widespread use and availability of digital video. The availability of audio-visual data in the digital format is increasing day by day. Currently several web sites host movies and provide users with the facility to browse and watch online movies. Movies constitute a large portion of the entertainment industry and they are always preceded by previews (trailers) and promotional videos because of the commercial nature of movie productions. So, several researchers have started to investigate the potential of analyzing audio signal for video scene classification.

In this proposed system, only audio information is used to separate the different audio class because audio-based analysis requires significantly less computation, it can be used in a preprocessing stage before more comprehensive analysis involving visual information. To provide automatic scene extraction models, audio data is considered in order to make the extraction results closer to human understanding and audio in fact tells a lot about mood of the clip, the music component, the noise, fast or slowness of the pace and the human brain too can classify just on the base of audio.

The motivation of combining the HMM and SVM is to combine the strong generalization ability of the SVM and the output probability of HMM for reducing audio features dimension by modeling to select the best features for each class.

In this work, an audio clip is classified into one of eight classes. The first step of the classification is to extract audio track from video file for feature calculation. To convey the information on extracted audio file, the audio quality is put with the sampling frequency of 22 kHz, bit rate of 128 kbps and mono channel.

The purpose of the second stage is to extract features and analyze in two levels: frame-level and clip-level by using audio features ZCR, STE, VRMS, VDR and MFCC which are proved to be effective for distinguishing audio classes. In this stage, the audio stream is analyzed into 1sec audio clips with 0.5 sec overlap and each clip is divided into frames of 20ms with non-overlapping.

Before performing the actual data mining process, for the sake of accuracy and efficiency, a pre-filtering process is needed to clean data. To represent a clip, a total of 17 features are extracted from each clip. Although all these features can be used to distinguish audio, some features may contain more information than others. Using only a small set of the most powerful features will reduce the time for feature extraction and classification. Moreover, the existing research has shown that when the number of training sample is limited, using a large feature set may decrease the generality of a classifier. Therefore, each audio class is modeled by using HMM to get the clean data and effective features. Extracted audio features using HMM are in the same dimension and in the same vector space. So reducing feature dimension by using HMM model can help the SVM to get the best training data and to raise the classification speed and accuracy while SVM classify the audio classes. In summary, the pseudo code of HMM-SVM algorithm is shown in Fig. 1.

```
HMM-SVM Algorithm
1. Support that there are k training set expressed as
   L={L_1,….., L_2}.
2. Each training set corresponds to an HMM set λ={λ_1,……,
   λ_k}, when λ_i={λ_1,……, λ_ki}.
3. Calculate the probability P(O | λ), the maximum
   probability of a class L_j in this set is $\max_{1\le j\le Ni} P(O|\lambda j)$.
4. From HMM algorithm, get new training data set.
5. Run the SVM algorithm by using new training data sets
   from HMM.
```

Fig.1. HMM-SVM algorithm

The overall procedure of the proposed system is followed:

1. Collect the audio files from video files.
2. Extract the features from audio files by using ZCR, STE, VRMS, VDR and MFCC.
3. Select the best features for each class by using HMM model.
4. Get the best training dataset from the above step.
5. Train the data for SVM classifier.
6. Classify the audio classes by using SVM.
7. Calculating the performance and accuracy of proposed system.

The proposed system design is shown in Fig.2.

V. EXPERIMENTAL EVALUATION

A. *Data*

Action movies which are collected from various sources are used to validate the developed HMM-SVM algorithm. In this system 6:30 hours of action movies consisting of four gigabytes of data is used.

In order to train classifiers it is necessary to have data labeled with ground truth by a human. The quantity and quality of this ground truth data is critical to build robust classifiers that have good generalization properties.

As the preliminary investigation, two action movies of 3 hours and 30 minutes is used as the training data containing classes such as gunshot, scream, laughter, speech, fighting with background crowd, shouting and background music…etc. To obtain the ground truth, each of them is hand-labeled into one of eight classes in movies in which mixed audio with background interference.

Four movies are divided into training and testing for datasets. The experiments will be performed with same feature dimensions for each of the feature extraction methods. The total duration of one action movie is approximately 1 hour and 45 minutes.

The occurrences of these events in action movies are quite different in some parts and some are similar. Gunshot and bomb are similar and also screaming and shouting is frequently similar. But people talking, fighting and music with background crowd sounds are quite different. These sounds are related with event. So these audio classes are used in scene segmentation and event detection. For example, gun, bomb and fighting class can use to define action scene and speech and laughter sequence is related to define happy scene. Moreover, miserable scene is defined by using screaming and shouting classes.

The classification of an audio stream can be achieved by classifying each clip into an audio class in sports video. The performance of the result is measured by classification accuracy defined as the number of correctly classified clips over total number of clips in respective class.

B. *Results*

All the experiments have been performed on a real world dataset, consisting in more than 6 hours of action movies. Training and test dataset have been taken according to a 2-fold cross-validation. The proposed system HMM-SVM classification method is compared with KNN-SVM and SVM only that is widely used in the literature as classification method.

Table I summarizes the selected features for each class respectively. The clean training data set gets by selecting the best features. In this table, rows refer to extracted features and columns are the classes at all classification levels. For example, the features STE, ZCR and VDR can be applied to discriminate between gunshots and scream.

TABLE I: SELECTED FEATURES TABLES FOR RESPECTIVE

Features/Classes	ZCR	STE	VRMS	VDR	MFCC
Gunshot	√	√	×	√	×
Scream	×	√	×	×	√
Car breaking	×	×	√	√	×
Speech	√	×	×	×	√
Laughter	√	×	×	√	×
Fighting	×	√	√	√	×
Shouting	×	√	×	√	√
Music with background crowd	√	×	√	√	×

Table II reports the accuracy and error recognition rate (ERR) resulted from the tests for 22798 clips. Length of a clip is 1 second. In this system, eight classes of audio are tested. Note that there is strong agreement of manual and automatic classification. Overall classification accuracy is exceeded average of 85%.

TABLE II: ACCURACY AND ERROR RECOGNITION RATE FOR INTRODUCED METHOD

Classes	Accuracy (%)	ERR (%)
Gunshot	95.68	4.32
Scream	92.77	7.23
Car breaking	96.56	3.44
Speech	97.02	2.98
Laughter	88.74	11.26
Fighting	90.36	9.64
Shouting	87.65	12.35
Music with Background Crowd	91.08	8.92

Table III compares the accuracy of HMM-SVM over KNN-SVM and SVM. For all classes, the recognition rates using HMM-SVM is outperformed than using KNN-SVM and only SVM.

TABLE III: COMPARISON OF PROPOSED METHOD AND OTHERS

Classes	SVM (%)	KNN-SVM (%)	Proposed Method (%)
Gunshot	90.44	92.45	95.68
Scream	86.23	89.12	92.77
Car breaking	74.29	90.08	96.56
Speech	88.76	90.67	97.02
Laughter	65.03	80.94	88.74
Fighting	86.53	89.87	90.36
Shouting	70.64	85.43	87.65
Music with background crowd	89.75	91.11	91.08

VI. CONCLUSION

This proposed system use HMM that can automatically find the temporal pattern of video scene streams. SVMs are used to incorporate with HMM to determine how to partition multiple classes in the system. The whole framework can be predictable more flexibility and the accuracy of auditory feature analysis can be improved to increase the overall event detection accuracy. In addition, the integration of classifiers can reduce the classification errors and it can detect events from other games, or even other type of videos. This system is designed to fulfill the requirement of video viewer and to improve the dealings with video for home user.

Ongoing work of this research is to detect and recognize video scene by using the audio classes. The resulting audio classes are applied to event detection and provide action scene annotation based on separating audio class. Other audio classes such as siren, horn, running water, thundering and bird sound are considered in future to skip commercial shots and locate the shot or story that users are interested in.

ACKNOWLEDGMENT

I am very grateful to Dr. K Zin Lin for fruitful discussion during the preparation of this paper and also specially thank to Rector Dr Aung Win, Professors and all my teachers from Technology University (Yatanarpon Cyber City), Myanmar.

REFERENCES

[1] P. Dhanalakshmi and S. Palanivel, “Classification of audio signals using SVM and RBFNN,” Journal of Expert Systems with Applications, vol. 36, pp. 6069-6075, 2008.

[2] R. S. Selva Kumari, D. Sugumar and V. Sadasivam, “Audio signal classification based on optimal wavelet and support vector machine,” IEEE, International Conference on Computational Intelligence and Multimedia Applications (ICCIMA), vol. 2, pp. 544-548, Dec 2007.

[3] S. Jain and R.S. Jadon, “Audio based movies characterization using neural network,” International Journal of Computer Science and Applications, vol. 1, no. 2, pp. 87-90, August 2008.

[4] S. Chu, S. Narayanan, and C.-C. J. Kuo, “Environmental sound recognition with time-frequency audio features,” IEEE Trans. on Audio, Speech, and Language Processing, vol. 17, no. 6, pp. 1142–1158, August 2009.

[5] S.-G. Ma and W.-Q. Wang, “Effectively discriminating fighting shots in action movies,” Journal of Computer Science and Technology, IEEE, vol. 26, no. 1, pp. 187-194, Jan 2011.

[6] T. Zhang and C.-C. Kuo, “Audio content analysis for online audiovisual data segmentation,” IEEE Trans. On Speech and Audio Processing, vol. 9, no. 4, pp. 441-457, May 2011.

[7] V. Elaiyaraja and P. M. Sundaram, “Audio classification using support vector machines and independent component analysis,” Journal of Computer Applications (JCA), vol. 5, issue. 1, 2012.

K. M. Chit received Bachelor of Engineering from Technological University (Kyaukse), Mandalay in Myanmar. She completed the Master course from Mandalay Technological University since 2010 and especially studied and finished the Thesis by Networking. She is working now in Technological University (Taunggyi) as at Assistant Lecturer under Information Technology Department. Now, She is a Ph.D student in University of Technology (Yatanarpon Cyber City) near PyinOoLwin, Upper Myanmar. Her fields of interest are Multimedia Signal Processing, Audio Indexing and Audio Information Retrieval.

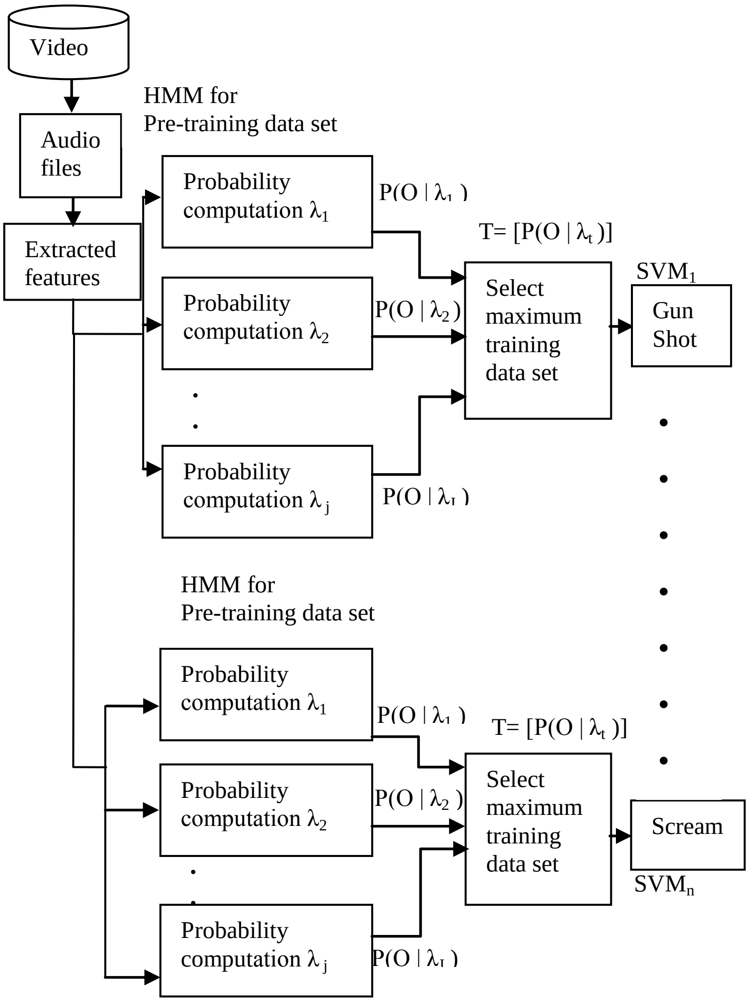
Video
Audio files
Extracted features
HMM for Pre-training data set
Probability computation λ_1
Probability computation λ_2
Probability computation λ_j
$P(O \mid \lambda_1)$
$P(O \mid \lambda_2)$
$P(O \mid \lambda_1)$
T= $[P(O \mid \lambda_t)]$
Select maximum training data set
SVM$_1$
Gun Shot
HMM for Pre-training data set
Probability computation λ_1
Probability computation λ_2
Probability computation λ_j
$P(O \mid \lambda_1)$
$P(O \mid \lambda_2)$
$P(O \mid \lambda_1)$
T= $[P(O \mid \lambda_t)]$
Select maximum training data set
Scream
SVM$_n$

Fig. 2. Proposed System Design