Audio Features Analysis and Classification Established on Sound Effect

KhinMyo Chit[#1], K Zin Lin[#2]

University of Technology (Yatanarpon Cyber City), ICT Department,

PyinOoLwin, Myanmar

Hardware Department, University of Computer Studies, Yangon (Bahan Campus)

Myanmar

[1]miss.khinmyochit@gmail.com

[2]kzinlin78@gmail.com

***Abstract*—With the rapid growing amount of multimedia, content-based information retrieval has become more and more important. As such, the increasing amount of multimedia information available highlights the need to develop systems able to automatically describe this information for more efficient filtering, retrieval and, in general, management. The goal is to allow that the huge amount of multimedia content available can be filtered, searched for, managed and consumed in a thoughtful, flexible, fast and efficient way. There is an increasing need to summarize and personalize audio-visual content.The main challenge is to index information retained in video in order to make them searchable and thus (re-) usable. This requires the multimedia content to be annotated, which can either be done manually or automatically. In this proposed system, SVM is combined with HMM based on audio features which are used to classify sound effect types including gunshot, scream, Aircraft, Car, Speech with Crowd, Music, Horror, Battle, Comic, Background Crowd, Train, Truck and Vehicle.**

***Keywords*— Support Vector Machine, Hidden Markov Model**

I. INTRODUCTION

Most of the researchers have been doing audio classification which can use in application area such as audio indexing, audio information retrieval (AIR), video highlighting, scene segmentation and event detection. Also audio feature analysis is important things and need to reduce feature dimension. Sound effect are used in most of the movies especially action movies. They did not use the real sound and they make sound effect such as waterfall, train, car, fighting scene and explosion. 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.

This is important that incorrectly labelled training samples can significantly reduce classification performance. This approach is to achieve high accuracy in classifying of mixed types of audio by combining two types of classifiers that are Hidden Markov Model (HMM) and Support Vector Machine (SVM). Before performing the actual data mining process, for the sake of accuracy and efficiency, a pre-filtering process is needed to clean data. 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.

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 in [1]. Linear predictive coefficients, linear predictive cepstral coefficients and mel-frequency cepstral coefficients audio features are calculated as features to characterize audio content.S. Jain and R.S. Jadon [2] focused on neural net learning based method for characterization of movies using audio information. They characterized the movie clips into action and non-action. In [3] 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.

Wan et al. [4] proposed an interesting system to automatically insert virtual content (advertisement, logo, etc) into an existing sports video. Most current researches only involve limited classes of sounds e.g. discrimination between music and speech or classification among silence, music, speech and noise etc, [5]. Several techniques have been employed for the purpose of classifying an unknown sound. The most common approaches are GMM based methods [6], HMM [7], Nearest Neighbor methods [5], Neural Network (NN) variants [8], Vector Quantization (VQ) [6, 8] and Support Vector Machine (SVM) [5].

S. Gao Ma and W. Wang [9] presented the discriminating fighting shots in Action Movies by using the camera motion and SVM classifier. V. Elaiyaraja and P. Meenakshi [10] presented audio classification system by using audio features and a frame-based multiclass support vector machine.Speech recognition and analysis has also a long tradition and is a

matured research area that focuses in the identification and recognition of input speech signals. For further discussion on classification of general sound, refer to [11].

III. BACKGROUND

A. Feature Extraction

One of the most important parts of automated audio classification is the choice of features or properties. 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. The mathematical representations of these features are described as equation (1) to (4).

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

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

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

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 n^{th} frame is calculated, by the following formula:

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

Where $S_n(i)$ is the i^{th} sample in the n^{th} 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

$$VDR = [MAX(V) - MIN(V) / MAX(V)] \quad (4)$$

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.

Consider the problem of separating a set of training vectors belonging to two separate classes, $(x_1; y_1), \ldots, (x_l; y_l)$, where $x_i \in R_n$ is a feature vector and $y_i \in \{-1, +1\}$ is avclass label, with a separating hyper-plane of equation $w \cdot x + bv = 0$; 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 x_i is called as Support Vector with α_i 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. 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 x_j from x_i(one of these will be the support vector and the other will be the testing data point). The support vector will be the centre of the RBF and σ will determine the area of influence this support vector has over the data space.

IV. PROPOSED SYSTEM DESIGN

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 Movie 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 pre-processing 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 thirteen classes.

To convey theinformation on audio file, the audio quality is putwith the sampling frequency of 44 kHz, bit rate of 128 kbpsand mono channel.The first thing what to do 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 modelled 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.

HMM-SVM Algorithm

1. Support that there are k training set expressed as $L=\{L_1,\ldots..,L_2\}$.
2. Each training set corresponds to an HMM set $\lambda=\{\lambda_1,\ldots\ldots,\lambda_k\}$, when $\lambda_i=\{\lambda_1,\ldots\ldots,\lambda_{ki}\}$.
3. Calculate the probability $P(O \mid \lambda)$, the maximum probability of a class L_j in this set is $\max_{1\le j\le Ni} P(O \mid \lambda j)$.
4. From HMM algorithm, get new training data set.
5. Run SVM algorithm by using new training data sets from HMM.

Fig. 1. HMM-SVM algorithm

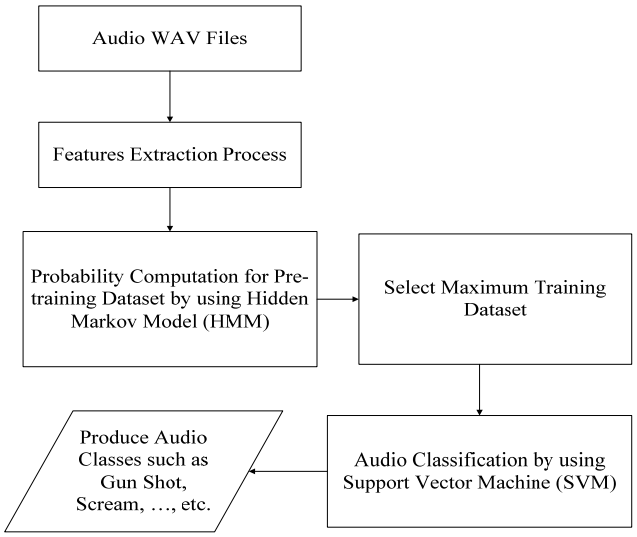

Fig.2. Proposed system design

The overall procedure of the proposed system is followed:

1. Collect the audio files with various sound effects in movie.
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. EXPERIMENT

A. Data Format

Audio features and HMM-SVM algorithm are applied to classify the various sound effects which are used in most of the movies. In this system thirteen kinds of 2000 audio files consisting of two gigabytes of data with WAV format are used. As the preliminary investigation for audio feature analysis, five hundred audio files are used as the training data and the other one thousand and five hundred audio sound files are used for testing. The effect of sound files are labelled and determined with classes to obtain the ground truth.

To convey the information on extracted audio file, the audio quality is put with the sampling frequency of 44 kHz, bit rate of 128 kbps and mono channel. To extract features and

analyse in two levels: frame-level and clip-level by using audio features ZCR, STE, VRMS, VDR and MFCC. The audio stream is analysed into 1sec audio clips with 0.5 sec overlap and each clip is divided into frames of 20ms with non-overlapping. The same feature dimension is used in this experiment for feature extraction.

B. Experimental Studies

Sound effects for some events are quite different in some cases and some are similar. Train and truck sound are similar in most of the time and background crowd and battle sound is frequently similar. On the other hand, music, people talking and gunshot sounds are quite different. All these sound are related with events so these audio classes can be used in scene segmentation and event detection. For example, gunshot, car and battle sound can be used for determining the action scene in movies and also the miserable scene can be detected by using screaming and horror sound. Moreover, comic and music is related to define the happy scene.

The good accuracy rate can get by using the clean training data set. So the best features are selected to get the clean training data set and speed up the process performance. The resulting accuracy percentage from HMM model's probability is used to select the best features for classifying the sound file.

The following tables I, II and III shows the result of testing the audio features by using the possible combination. Some examples are described. By using three features (1, 5, 11), the maximum classification rate 90.7 for Gunshot is established from Table I.

Table II shows the accuracy test result for scream and three feature combinations are used to separate this class from other classes. In this experiment the training data sample is 5000 and the testing data sample is 10978. Length of a clip is 1 second. By using three features (2, 9, 11), the maximum classification rate for scream is 81.31.

TABLE I
ACCURACY TEST RESULTS FOR GUNSHOT

3F	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9
3	68.5							
4	72.3	74.1						
5	72.4	72.8	70.9					
6	73.8	75	73.5	73.8				
7	71.9	71.4	71.7	77.9	81.3			
8	77.2	76.9	76.5	72.9	69.6	74.4		
9	71.8	75.5	73.0	82.2	78.6	75.2	75.3	
10	70.8	74.3	57.2	70.3	71.1	69.8	78.3	70.1
11	76.6	85.5	86.1	90.7	88.1	86.3	88.5	86.5
12	72.6	72.7	69.6	74.6	79.3	71.7	78.6	70.7
13	69.2	73.3	54.5	67.7	73.3	71.8	76.6	67.0
14	71.7	73.1	73.7	71.7	82.5	73.1	73.6	70.1
15	71.4	74.7	59.3	70.4	77.5	72.5	75.5	70.2
16	70.6	71.5	56.6	62.7	68.5	68.5	77.1	74.8
17	72.1	75.6	79.8	87.6	84.0	77.8	81.0	81.5
	77.2	85.5	86.1	90.7	88.1	86.3	88.5	86.5

TABLE II
ACCURACY TEST RESULTS FOR SCREAM

3F	2,2	2,3	2,4	2,5	2,6	2,7	2,8	2,9
4	73.2							
5	70.9	69.8	70.7					
6	71.0	69.7	71.0	72.5				
7	68.9	67.9	68.9	71.0	74.2			
8	76.5	72.6	76.5	73.0	74.3	76.8		
9	73.5	70.7	74.0	71.5	72.6	72.6	76.2	
10	70	68.3	70.0	73.0	70.3	69.2	76.4	75.3
11	76.8	75.9	76.6	76.0	76.8	75.7	76.5	81.3
12	72.6	70.3	73.2	72.0	73.3	68.7	76.0	73.8
13	69.5	70.8	69.6	74.0	69	66.8	74.8	70.1
14	72.6	69.5	72.6	71.3	73.4	69.3	75	71.8
15	73.6	70.7	73.6	71.8	70.2	67.3	74.1	72.4
16	71.7	68.5	71.7	68.9	69.7	68.3	77.2	74.3
17	72.1	73.5	72.4	72.1	74.5	68.3	77.6	75.1
	76.8	75.9	76.6	76.0	76.8	76.8	77.6	81.3

TABLE III
ACCURACY TEST RESULTS FOR AIRCRAFT

2F	1	2	3	4	5	6	7	8
2	52.2							
3	54.7	81.7						
4	62.9	67.4	66.7					
5	53.4	85.4	85.5	65.7				
6	69.3	66.8	68.3	65.8	69.9			
7	74.5	78.9	80.1	80.5	82.5	77.6		
8	85.8	89.3	90.6	88.7	90.5	84.4	93.6	
9	45.9	37.3	55.5	55.5	69.4	65.6	79.4	89.3
10	37.0	40.6	53.7	45.1	43.8	62.8	78.6	91.1
11	70.1	69.7	73.2	74.9	80.7	79.5	91.7	89.9
12	57.1	75.1	71.6	69.5	68.4	70.8	86.8	89.3
13	50.4	65.8	66.7	70.7	67.5	67.6	79.3	87.9
14	44.3	42.5	50.7	50	47.6	60.9	78.3	86.5
15	45.6	51.1	64.2	65.8	75	64.5	79.2	86.6
16	60.7	64.4	69.7	71.6	78.7	70.3	84.6	92.2
17	42.6	63.5	64.9	59	68.0	64.3	77.4	90.8
	85.8	89.3	90.6	88.7	90.5	84.4	93.6	92.2

TABLE IV
SELECTED FEATURES SUMMARY

Classes/Features	ZCR	STE	VRMS	VDR	MFCC
Gunshot	√	X	X	X	√√ (5,11)
Scream	X	√	X	X	√√ (9,11)
Aircraft	X	X	X	X	√√ (7,8)
Car	X	√	X	X	√√ (5,8)
Speech with Crowd	√	X	√	√	X
Music	√	X	X	√	√
Horror	X	√	√	X	X
Battle	X	√	X	√	√
Comic	√	X	√	X	√
Background Crowd	√	√	X	X	√
Train	X	X	√	X	√√ (6,16)
Truck	X	X	X	√	√√ (2,4)
Vehicle	X	√	√	√	X

The above table IV summarized the selected features for each class respectively after testing with all possible combinations which shows with some examples in the above table. In this table, rows refer to extracted features and

columns are the classes at all classification levels. For example, the features ZCR and MFCC feature's 5 and 11 can be applied to discriminate the target gunshot from other classes. For screaming class, the features STE and MFCC feature's 9 and 11 are used.

C. Outcomes

Table V gives information about the classification accuracy and error recognition rate resulted from the tests of 1500 audio files for all classes by using the selected features. From the information, aircraft class is highest in accuracy rate at about 97.23, and lowest in horror and train. Perhaps surprisingly, the speech with crowd class also make far greater rate than other classes. But overall classification accuracy is exceeded average of 87%.

TABLE V
CLASSIFICATION ACCURACY FOR INTRODUCED METHOD

Classes	Accuracy (%)	Error Recognition Rate (%)
Gunshot	95.68	4.32
Scream	92.77	7.32
Aircraft	97.23	2.77
Car	96.56	3.44
Speech with Crowd	97.02	2.98
Music	91.08	8.92
Horror	87.65	12.35
Battle	90.36	9.64
Comic	88.74	11.26
Background Crowd	89.22	10.78
Train	87.95	12.05
Truck	88.82	11.18
Vehicle	95.5	4.5

TABLE VI
COMPARISON OF PROPOSED METHOD AND OTHERS

Classes	SVM (%)	KNN-SVM (%)	Proposed Method (%)
Gunshot	90.44	90.45	95.68
Scream	86.23	89.12	92.77
Aircraft	74.29	90.08	97.23
Car	88.76	90.67	96.56
Speech with Crowd	65.03	80.94	97.02
Music	86.53	89.87	91.08
Horror	75.64	85.43	87.65
Battle	80.76	85.33	90.36
Comic	89.22	89.78	88.74
Background Crowd	88.81	89.2	89.22
Train	75.44	86.54	87.95
Truck	78.91	89.45	88.82
Vehicle	80.76	85.33	95.5

With a sound effect of 2000 files, KNN-SVM method is only slightly higher than SVM method in Table VI and it also has similar figures for accuracy rate of gunshot, background crowd and truck classes. However, on three classification methods, the proposed method is higher than others. Of the thirteen classes, aircraft and speech with crowd appears to have the best accuracy rate overall. In addition, train, truck and vehicle classes have the least accuracy rate and the highest levels of error recognition rate. The classification rate, according to the tables, would be that there is a higher accuracy rate in all classes in the sound effect.

VI. CONCLUSIONS

This proposed system use HMM which is fast, simple and multi-class with probability and can handle some degree of overlapping between classes. 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. Not only this system is designed to fulfil the requirement of video viewer and to improve the dealings with video for home user but also it provides the video editor and director.

The advantage of this proposed system is that sound event types such as gunshot, scream, car-breaking, people talking, laughter, fighting, shouting and crowd background can provide to detect the abnormal case in our environment. These classes are used to get the safe place and protect the dangerous place. The disadvantage of the system is that some sound types are difficult to discriminate because of background crowd. A more general classification system can differentiate between speech, music and other environmental sounds. Then, based on the result of classification, different processing, indexing or retrieval techniques will be applied accordingly. More types of sound effect are still need to classify for further study.

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] 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.

[3] 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.

[4] C. Xu, K. Wan, S. Bui, and Q. Tian, "Implanting virtual advertisement into broadcast soccer video,"Proc of IEEE PCM'04, pp. 264-271, 2004.

[5] F. Yan, W. Christmas, and J. Kittler, "A maximum a posteriori probability viterbi data association algorithm for ball tracking in sports video," Proc. of IEEE ICPR'06, pp. 279-282, 2006.

[6] M. Xu, J. Orwell, L. Lowey, and D.Thirde, "Architecture and algorithms for tracking football players with multiple cameras," Proc. of IEE Vision, Image and Signal Processing, pp. 232-241, 8 April 2005.

[7] E. Singer, M. A. Kohler, P. A. Torres and R. J. Greene, "Approaches to language identification using Gaussian mixture models," ICASSP, 2002.

[8] M. Liu and C.Wan, "A study on content-based classification and retrieval of audio database," Proc. of 2001 International Database Engineering and Applications Symposium,2001, pages 339–450.

[9] 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.

[10] 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.

[11] D. Mitrovic, M. Zeppelzauer and C. Breiteneder,"Discrimination and Retrieval of Animal Sounds," Proceedings of the IEEE conference on Multimedia Modeling, 2006. Multimedia Applications (ICCIMA), vol. 2, pp. 544-548, Dec 2007.