Audio Annotation on Myanmar Traditional Boxing Video by Enhancing DT

K. Zin Lin and Khin Htar Nwe
Faculty of Computer Systems and Technologies, University of Information Technology, Yangon, Myanmar
Email: {kzinlin, khinhtarnwe}@uit.edu.mm

***Abstract*—Nowadays labelling video data and analysis the resource information from video is very popular but it is difficult for storage and time consuming. Instead of video, audio is pretty one and, audio signal processing and audio annotation is at the heart of highlighting, recording, storing and transferring. The big data processing and training is challenging in research work and how to reduce data size and speedy highlight is the majority job in real world. In this proposed system, the process of audio annotation is focused on boxing video to entertain without time exhausting. This system is intended to provide boxing lover to highlight, review and replay their desired part by removing unwanted video parts. The decision tree is used for classifying audio and support vector machine is combined to enhance the classification accuracy and to promote the weakness of decision tree.**

***Index Terms*—signal processing, support vector machine, decision tree**

I. Introduction

Since the early on 2000s, there are many ways to develop video information retrieval for social science researchers. They must study dynamically changes in social life and they need to provide with the powerful information. Most of the people are using the internet for sharing experiences and they are making many of these data easily accessible. These sharing data are easily taken by camcorders, drones and mobile phone cameras and posted on facebook and uploaded to youtube in every minute. In previous year, the researcher are attempting and concentrating on indexing and retrieval of video, audio and images in digital signal processing.

Retrieving the audio from video is used for automatic speech recognition, speaker identification for security purpose and reclaiming the image is for face recognition, emotion detection and character recognition. On the other hand, the information of audio signal provides the useful information in the application area of IoT, speech security, the speaker recognition, speech and music discrimination and audio highlighting, labelling and annotation. The descriptions and the corresponding content can be interoperated to illustrate the viewer, as it is essential to identify a pattern in multimedia narratives by indicating the syntax and semantics.

Most of the video takes around two to three hours long and the scene is shot from beginning to end. But the time taken for sport videos are from thirty minutes to one and half hour depend on the kind of videos. For example, there are events such as penalty, free kick, goal and corner kick in football match that are more interested from viewer. If the viewer watches the whole match, it is time consuming and some parts make them lazy. They would like to skip uninterested parts and some of the events are attracting them and they want to watch. At that time, sports highlighting is challenging the researcher and there are the tasks to develop an automatic highlight, label and annotation.

In this proposed system, Myanmar traditional boxing video is used for audio classification to provide video highlight detection area. There are two main tasks in this system: one is preprocessing task to enhance the results of decision tree by using Support Vector Machine (SVM) and next is classifying the audio class for audio annotation. The training samples are very important to achieve high classification rate and they provide to reduce the error rates. The classification performance can be reduced by wrong labelling training samples. Since the kernel SVM gives strong generalization ability and it can overcome the weakness of decision tree. The tree style is a hierarchical structure. When the level goes down to step by step, classification accuracy of mixed types of audio can be decreased because the training data is defined to wrong class. In this study, the two classifiers are supported to each other and the outcomes is satisfactory with high correctness in mixed types of audio classification.

The remaining parts are structured as follows. The correlated work with this study is discussed in Section II and in Section III, the low level perceptual and cepstral feature of audio clip are represented and describes the overview of SVM kernel and decision tree. In Section IV, multiple classification is proposed for different audio classes and experimental study is presented in Section V. The conclude words are presented for the proposed system in the last Section VI.

II. Related Work

Even though there are many tools and research work, users still need a large amount of time to annotate properly. In this part, some related papers are discussed and reviewed because alternative methods are considered in annotation, retrieval and classification and different softwares are used for this purpose in this research field.

Enhance feature vector formation technique is presented by R. S. S. Kumari *et al.* [1] for audio

Manuscript received August 11, 2020; revised February 7, 2021.

doi: 10.12720/jait.12.2.107-112

categorization and classification. Acoustic features such as bandwidth, sub-band power, pitch information and brightness are extracted by using wavelet in this system and frequency cepstral coefficients are also extracted to accomplish audio grouping.

In [2], classical K-means algorithm are inspired with unsupervised clustering method to classify speech and music signals effectively. The first step is extracting the relevant audio features from audio files and then each cluster is refined based on OCSVM, one-class support vector machine, after initializing centers in an iterative K-means algorithm. Better experimental results are proofed that this system is more efficient than other methods on the same database.

Detection of semantic events on sport events has been focused to ease access, summarization and browsing in research work. Only visual data are analyzed in some of the works but the classification of audio events is also important for video highlighting. Not only using the description of video content can create more but also can help the enhancement of the discovery of highlights. So both of the data are very important in signal processing. [3].

To detect generic sport highlights, a system is presented by Divakaran *et al.* [4]. In this system, only audio features are used to perform real-time classification and classified the different audio types such as cheering, speech with exciting and clapping etc. Gaussian Mixture Models with low complexity is applied to perform classification and the MDCT coefficients are used for audio features in this work. These coefficients are from the AC-3 encoding used in MPEG-2 streams.

In [5], audiovisual annotation is expressed for multi-view field recordings. Annotating multi-view field recordings is presented by manually by low-cost procedure. Users are allowed to perform various kinds of requests and ToCaDa dataset [6] is also utilized. The ground truth is very important for training case so a ground truth is formed by valuable approximation data in the resulting annotations.

R. Cohendet *et al.* [7] submitted memorability for long term with family activities by annotating, understanding and predicting the video data. To build a dataset of 660 videos, 104 people are participated from weeks to years in their systems and that dataset is available in the research community field. The computational models are also proposed for VM prediction where the use of various audio and visual features for the task was investigated.

For multimedia information retrieval, model with video [8] or audio events [9] are challenging in the research field and there are multiple campaigns in this area. Moreover, the annotators are usually making an agreement between them because annotations are hardly objective to meet the requirements [10].

In addition, the dataset is very important role in information retrieval and annotation. Very large audio dataset extracted from videos was supplied with the Audio Set dataset [11], but YouTube users tagged the annotations on the audiovisual content. The sound of the video is not taken into accounts in the AVA dataset [12]. The detailed spatio-temporal annotations of persons leading actions are offered in this dataset for the vision and motion area. The datasets are reviewed with various works from the multiple point of different model and several structures [13].

III. Research Environment

A. Audio Feature Extraction

In any type of audio, image and video signal processing, the extraction of feature vectors is the basic process. The features included useful information and the training dataset can provide to obtain high accuracy for audio classification and recognition.

Audio clip-level elements are calculated based on the frame-level features in this proposed system. A clip is used as the categorization unit. Short-Time Energy (STE) and Spectral Flux (SF) is computed as audio features and the means of all frames are used in each clip which has been proven for operation in distinguishing music, speech and speech with background sound [1], [14]. The mathematical representation of these features is described as Eqs. (1) and (2).

$$E(m) = \sum_{m} (x(n)W(n-m))\hat{}2 \tag{1}$$

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

Spectrum Flux (SF) is identified by the average and mean value of spectrum between the successive two frames, there are adjacent in each clip.

$$SF_t = \sum_{n=1}^{N} (N_t[n] - N_{t-1}[n])\hat{}2 \tag{2}$$

where the magnitude of the Fourier transform by normalization are $Nt\text{-}1[n]$ *and* $Nt[n]$ and $t\text{-}1$ is the previous frame and t is the current frame, respectively. The amount of local spectral changes is measured in this SF. Based on the previous work [15], SF has very useful information to classify environment sound, speech and music.

Noise Frame Ratio (NFR) is defined by the ratio of noise frames in each audio clip. A frame is measured as a noise frame when the highest local peak of its normalized correlation function is smaller than a preset threshold.

MFCC is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency and it is very useful feature for extraction in audio classification.

SVM makes the boundary between the classes and it is used by changing kernel instead of using the probability density of each class from the model such as (Gaussian Mixture, Hidden Markov Models, etc.). SVM algorithm is a classification algorithm that provides state-of-the-art performance in a wide variety of application domains, state-of-the-art performance is provided by using the SVM classification algorithm. In the following section, it is expressed in two ways why the SVM is used in this study.

One is high dimensional, noisy data are involved in many audio classification problems. For this problem,

SVM behavior is well suited when machine learning methods or additional statistical are compared. Another one is feature selection of audio data. The different classes is so complex that may have intersecting or interconnected regions. The case such as linearly non-separable different audio classes can be handled by SVM because it is based on kernel. The classifier will make the larger scope to reduce estimated risk for clearer generalization.

SVM renovated the input space to make a better feature space via a nonlinear plotting task. From the nearest points of the preparation, the splitting hyper plane with maximum distance is constructed. For example, the challenge of splitting a training dataset vectors belonging to two distinct classes, (*x1; y1*), ..., (*xi; yi*), where *xi* $\in$ *Rn* is a vector of feature and *yi* $\in$ *{−1, +1}* is a labelling the class, with a splitting hyper-plane of equation $w\,x+b=0$; of all the boundaries established by *w* and *b*. Based on this rule, the ultimate best possible hyper-plane classifier can be characterized by the following equation:

$$f(x)=\operatorname{sgn}\left(\sum_{i=1}^{l}\overline{\alpha_t}\,y_i x_i x+\bar{b}\right) \tag{3}$$

where α and *b* are parameters for the classifier; the solution vector *xi* is called as Support Vector with αi being non-zero.

By using the kernel function *K(x,y)*, the inner product is replaced by SVM in some cases such as linearly non-separable and non-linearly separable events. After that an optimal dividing hyper-plane is constructed in the mapped space. In this study, the Gaussian Radial Basis kernel will be used:

$$K\left(x_i, x_j\right)=\exp\left(-\frac{\left\|x_i-x_j\right\|^2}{2\sigma^2}\right) \tag{4}$$

B. Decision Tree

The decision tree is very easy to understand for classification. It is not only very simple for implement but also very economy for computing. A decision tree is a hierarchical categorization process that is determined by a series of inquiries. The option of following questions varies on the response to the present question as soon as the first question has been asked. A directed graph is represented as a tree. The first issue is queried at the 'root top' and the following questions are requested at 'nodes'. The appropriate 'branch' to further nodes will be selected by the answer to a question posed at a node. A 'leaf' is defined at a terminal node. The decision tree is the same style with if-then rules that is more readability for human.

IV. Proposed System

The sport video such as soccer, basketball, swimming, boxing and tennis has included the background noise and it is very difficult to get the pure signal so the researcher has to focus on audio data by concentrating in this area. We should describe goodness of decision tree based SVM for mixed type audio.

In this paper, we discuss about combining Support Vector Machine and decision trees for multi class audio classification. As a decision of binary tree, the SVM is used to select ground truth training data and to create the better rules, the whole training set is not needed to train. At that time, better accuracy and time saving can be expected after classify each audio clip.

An acoustic piece is categorized among one of four classes in this proposed system. As SVM can classify into two-class so three SVMs are used for two level classification.

In the first phase, the audio quality standard is setting with the same properties that is 22 kHz for sampling frequency and the bit rate is 128 kbps and using mono channel for all videos, which are used in this intended scheme. After the standardization of the audio value, the audio features are extracted and calculated in the second phase. In this phase, audio features are analyzed with two level such as frame and clip-level into 1 sec audio clips with overlapping for 0.5 sec. 20 ms frames are non-overlapping within each clip and the effective features: STE, MFCC, NFR and SF are used to distinguish between audio classes.

In classification and recognition, the preprocessing state is the most important to gain the best efficiency and accuracy. Moreover, the training data is needed to clean by making the pre-filtering process. For this purpose, SVM classifier is applied to classify the data before going to decision tree. Based on the output result of each SVM, the new training dataset is used in the decision tree. The original training data may reduce the classification accuracy but the new one can improve the better result. Because the SVM is the higher classifier and it has the advantages by varying the kernel value to classify the data.

The proposed system algorithm is shown in the following algorithm with pseudo code. In this figure, Tr-svm is the training dataset that is the original data and Te-svm is the testing dataset. By running the SVM algorithm for each dataset, not only which feature pairs can classify more efficiently the audio data but also the predicted training data can be used as a new clean dataset. This new dataset is defined as S-svm and it was used in decision tree training and the amount of training data and the feature vector is reduced in classification for speedy purpose. In addition, Tr-new-dt and Te-new-dt are used for training and testing by making two-fold cross validation and the audio classes are classified accurately. Finally, the tree is constructed to classify the multi-class audio.

Algorithm: SVM-DT

1: Create cross validation data set (Tr-svm, Te-svm)
2: Run the SVM algorithm for each data set (Tr-svm, Te-svm), get the prediction data P-svm
3: FOR each data set P-svm
  {
    IF (prediction is correct)
      Select data into new data set S-svm
  }
4: Create train data (Tr-new-dt) from S-svm for decision tree and take Te-svm as Te-new-dt
5: Run the Decision Tree algorithm for each data set (Tr-new-dt, Te-new-dt), get the rule set R

In the first level of tree, SVM1 discriminates between speech and non-speech. After that those speech clips are classified into commentator's speech and background crowd by using SVM2 in the second level. As a parallel processing, the non-speech clips are classified into bell sound and clapping sound by using SVM3. Our proposed system design is shown in Fig. 1.

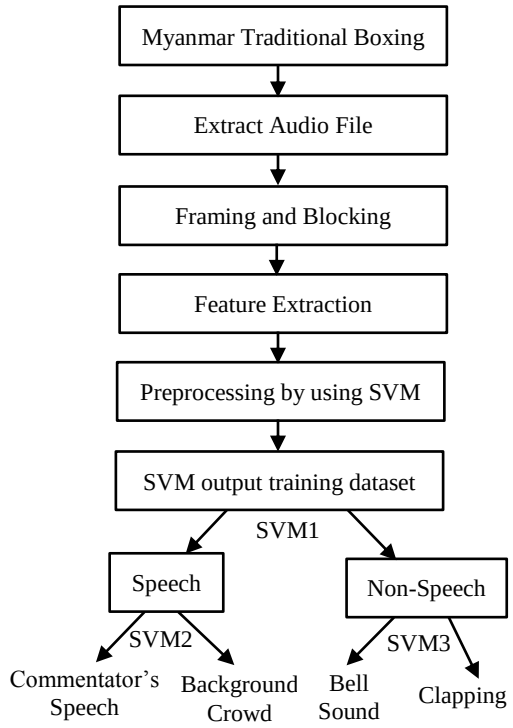

Figure 1. System architecture of proposed system.

V. Evaluation of Experimental Result

A. Data Setup

The used SVM-DT is validated by Myanmar Traditional Boxing (MTB) videos which are collected from different resources. We have seen totally 2.6 hours of MTB videos included over one gigabytes of data. They have different live times, digitized at different studios, sampled at 22 kHz, and reported by different announcers. The information of Five MTB videos are briefly described in Table I and they are used in this study.

An announcer's speech, background crowd, audience clapping, and bell sound classes are included in the training data and all these sounds are involving in 1 hour of MTB video is used as the preliminary investigation. Each sound class are labeled by hand into one of 4 classes in MTB video to obtain the ground truth and background interference such as loud cheering and applause sound are mixed in these boxing video.

Five MTB videos are used to get the datasets for training and testing. Feature dimensions are the same in these experiments. The total duration of one MTB video is different and it is depending on the fighting to win. Totally 2.6 hours of traditional boxing match that are exhibited in 2017, 2018 and 2019 MTB video will be used in this experiment.

The scene and sound of each events in MTB videos are quite different. For example, commentator's speech and excited events are the most frequent while bell sounds usually appear for the duration of the break of the match. The speech with crowd is connected to continuing actions while the excited sound is correlated to focusses such as fight on face, kick near the chin and knee fights into the chest, etc. The clapping class is related to encouragement to winner and introduce about the boxer.

TABLE I. Information for Five MTB Videos

No	MTB videos	Length	Year
1	Too Too (Myanmar) vs. Dave Leduc (Canada)	25 m 19 s	2016
2	Too Too (Myanmar) vs. Chanajon PK (Thai)	34 m 55 s	2017
3	Shwe Yar Man (Myanmar) vs. Keivan Soleimani (Iran)	30 m 48 s	2019
4	Soe Lin Oo (Myanmar) vs. Reza Ahmadnezhad (Iran)	35 m 50 s	2019
5	Shwe Yar Man (Myanmar) vs. Kohei Tokeshi (Japan)	28 m 00 s	2019

B. Results

The experimental results are evaluated on a real world dataset and it consists of more than two hours for Myanmar Traditional Boxing videos in Myanmar language. 2-fold cross-validation action is taken for training and testing dataset and the proposed enhancing DT with SVM classification technique is compared with SVM and decision tree those are widely used classification method in the literature.

TABLE II. Selected Features for Each SVM

Features	SVM1	SVM2	SVM3
STE	1	0	1
SF	1	1	0
NFR	0	0	1
MFCC1	0	0	0
MFCC2	0	0	0
MFCC3	1	0	0
MFCC4	1	0	0
MFCC5	0	0	0
MFCC6	0	1	0
MFCC7	0	1	0
MFCC8	0	1	1
MFCC9	0	0	1

Table II summarizes features employed for decision tree SVM where extracted features are expressed in rows and columns are the SVM classifier at all classification levels to discriminate the audio classes. For example, at the main level of classification tree, to discriminate between speech and non-speech clips, features such as STE, SF and MFCC3&4 can be applied. Table III reports the test results of all clips and they are presented as the classification accuracy and error recognition rate (ERR) of all experiments. Each clip is 1 second long and four classes of audio are tried in this study. For summary, the average classification accuracy is over 83 % for all classification.

TABLE III. RESULT OF THE INTRODUCED METHOD

Classes	Accuracy (%)	Error Rate (%)
Commentator's Speech	96.05	3.95
Background Crowd	86.33	13.67
Bell Sound	89.23	10.77
Clapping	83.40	16.60

TABLE IV. COMPARISON OF SVM, DT AND PROPOSED METHOD

Classes	SVM (%)	DT (%)	Proposed Method (%)
Commentator's Speech	90.84	88.67	97.05
Background Crowd	75.44	70.72	86.33
Bell Sound	86.93	80.34	89.23
Clapping	78.21	63.06	83.40

In Table IV, the accuracy of SVM-DT is compared with SVM and DT. The proposed method is outperformed in the recognition rates rather than only SVM and decision tree for all classes. Ongoing work is to identify the best feature sets for classifying the other audio classes such as excited speech and environmental sound.

VI. CONCLUSION

This proposed system has two main parts: feature selection and audio classification. This system can develop to provide video indexing, structure parsing process and analyzing the audio content. In this research, two classifiers: SVM and decision tree are presented by combining them for effective multi-label audio classification. STE, SF, NFR and MFCC are used for feature extraction. The SVM as a preprocessing of decision tree to select strong instances to generate rules is used. Also, the SVM as a decision of binary tree to select strong instances to generate rules is used. The performance of the algorithm is measured by accuracy and error rates.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

This research work was mainly conducted by K. Zin Lin. Khin Htar Nwe worked together in analyzing the data and comparing the results. They have done in the same research lab and worked together to get the good results.

REFERENCES

[1] R. S. S. Kumari, V. Sadasivam, and D. Sugumar, "Audio signal classification based on optimal wavelet and support vector machine," in *Proc. International Conference on Computational Intelligence and Multimedia Applications*, 2007, pp. 544-548.

[2] S. O. Sadijadi, S. M. Ahadi, and O. Hazrati, "Unsupervised speech/music classification using one-class support vector machines," in *Proc. 6th Internal Conference on Information, Communications & Signal Processing*, 2007, pp. 1-5.

[3] D. A. Sadlier and N. E. O'Connor, "Event detection in field sports video using audio-visual features and a support vector machine," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 10, pp. 1225-1233, 2005.

[4] I. Otsuka, R. Radhakrishnan, A. Divakaran, M. Siracusa, and H. Mishima, "An enhanced video summarization system using audio features for a personal video recorder," *IEEE Transactions on Consumer Electronics*, vol. 52, no. 1, pp. 168-172, 2006.

[5] P. Guyot, T. Malon, G. R. Jimenez, S. Chambon, V. Charvilla, and A. Crouzil, "Audiovisual annotation procedure for multi-view field recordings," in *Proc. International Conference on Multimedia Modeling*, 2019, pp. 399-410.

[6] T. Malon, G. Roman-Jimenex, and C. Senac, "Toulouse campus surveillance dataset: Scenarios, soundtracks, synchronized videos with overlapping and disjoint views," in *Proc. the 9th ACM Multimedia Systems Conference*, 2018, pp. 393-398.

[7] R. Cohendet, K. Yadati, N. Duong, and C. Demarty, "Annotating, understanding, and predicting long-term video memorability," in *Proc. International Conference on Multimedia Retrieval*, 2018, pp. 178-186.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211-252, 2015.

[9] A. Mesaros, T. Heittola, and A. Diment, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proc. International Conference on Detection and Classification of Acoustic Scenes and Events*, November 2017.

[10] D. Turnbull, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Transactions on Audio, Speech and Language Processing*, pp. 467-476, 2018.

[11] J. F. Gemmeke, D. P. W. Ellis, A. Jansen, and D. Freedman, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. Acoustics, Speech and Signal Processing*, 2017, pp. 776-780.

[12] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, and Y. Li, "AVA: A video dataset of spatio-temporally localized atomic visual actions," in *Proc. International Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6047-6056.

[13] A. A. Liu, W. Z. Nie, and N. Xu, "Benchmarking a multimodal and multiview and interactive dataset for human action recognition," *IEEE Transactions on Cybernetics*, vol. 47, no. 7, pp. 1-14, 2016.

[14] L. Lu, H. J. Zhang, and S. Z. Li, "Content-Based audio classification and segmentation by using support vector machines," *Multimedia Systems, Digital Object Identifier*, vol. 8, no. 6, pp. 482-492, 2003.

[15] L. Lu, H. Jiang, and H. J. Zhang, "A robust audio classification and segmentation method," in *Proc. the 9th ACM International Conference on Multimedia*, 2001, pp. 203-211.

Dr. K. Zin Lin is an Associate Professor in Faculty of Computer Systems and Technologies, University of Information Technology. She has completed her M.C.Tech Degree in Computer Technology from University of Computer Studies, Yangon (UCSY) and B.C.Tech. She has already completed her Doctoral Degree (Ph.D. (IT)) in Information Technology from UCSY since 2010. Currently, she is interesting the research area such as digital signal processing, wireless technology and IoT. She is also a member of Digital Signal Processing Research Lab and she is doing the research continuously in her University of Information Technology (UIT). She is teaching the undergraduate, master and Ph.D. students currently and supervises the master and Ph.D. candidates. Her working experience is almost 18 years. She is very happy to do research and to provide the student's research work.

Dr. Khin Htar Nwe is a Professor in Faculty of Computer Systems and Technologies, University of Information Technology. She has graduated with Bachelor of Science (B.Sc.) (Hons:) degree specialized in Physics and completed Master degree from Yangon University. She got Ph.D. degree in 2008 at University of Computer Studies, Yangon (UCSY). Currently, she is interesting the research area such as electronic physics and microcontroller, image processing, digital signal processing, embedded system designs and wireless sensor networks. She is also a leader of Digital Signal Processing Research Lab and she is doing the research continuously in her University of Information Technology (UIT). She is teaching the undergraduate, master and Ph.D. students currently and supervises the master and Ph.D. candidates. Her working experience is almost 20 years.