An Audio Classification Method for Structuring Video Scenes

K Zin Lin, Moe Pwint
University of Computer Studies, Yangon
kzinlin78@gmail.com

Abstract

Video is a rich source of information, with visual, audio, and textual content. The process of building audio classifiers for high-level content descriptors, especially in large datasets, is not trivial. An important and critical aspect of this process is obtaining ground truth annotations for training the classifiers. Building audio classifiers for a large real world collection is significantly harder than working with a well labeled small data set as is the case with the majority of existing work. In this approach we combine the SVM with decision tree into an algorithm called SVM-DT to assist this annotation process. We use the SVM as a preprocess of decision tree to select strong instances to generate rules; it is known as decision support vector machine (DSVM). In this proposed system eight audio classes are considered: silence, non-silence, announcer's excited speech, speech over crowd, whistle, applause, crowd only and cheering over crowd. This classification and analysis is intended to analyze the structure of the sports video. Soccer videos are experimented in this system.

1. 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.

A long sports video can be divided into parts and only a few of these parts contain certain highlights which are interesting to human. For example, in a soccer video, there are events such as goal, corner kick, free kick, etc. It is not difficult for human beings to understand the video by using cognitive stills. However, it is still challenging task to develop an automatic system to fully understand the video content, although several feature sets and machine learning algorithms have been tested, providing choices of speed and performance for a target system.

Sports video highlight detection is a popular topic. The sports video retrieval task requires analyzing for labeling, 1 hour and 45 minutes of soccer video that correspond to approx. 1 gigabyte of audio data (mono, 22050 Hz sampling rate). In this proposed system, the algorithm SVM-DT employs the SVM as a preprocess of decision trees for multi class audio classification and the obtained training set is used to train a decision tree learning system and to extract the corresponding rule sets. This is important that incorrectly labeled training samples can significantly reduce classification performance. Since the SVM usually has strong generalization ability and using the SVM as the

inputs to DT, the noise, may be reduced by the process of SVMs, and some weak cases may be sieved by SVMs. This approach is to achieve high accuracy in classifying of mixed types of audio by combining two types of classifiers.

The rest of this paper is organized as follows. In Section 2, we present the related work and Section 3 describes how an audio clip is represented by low level perceptual and cepstral feature and gives an overview of linear, kernel SVM and decision tree. In Section 4, a method for multi-class classification is proposed and in section 5, experimental study is presented. Finally, in Section 6, we conclude for the proposed system.

2. Related Work

There has been a lot of work in various types of audio classification. This classification is to determine the category of audio file automatically according to the features under given classification system. A. Rabauoi [1] consists of illustrating the potential of SVMs on recognizing impulsive audio signals belonging to a complex real world dataset. They presented to apply optimized one-class support vector machines (1-SVMs) to tackle both sound detection and classification tasks in the sound recognition process.

The issue of mixed type audio data based on Support Vector Machine (SVM) is addressed in [2]. In order to capture characteristics of different types of audio data, besides selecting audio features, they designed four different representation formats for audio features such as ZCR, silence ratio, harmonic ratio and sub-band energy. Their SVM-based audio classifier can classify audio data into five types: music, speech, environment sound, speech mixed with music, and music mixed with environment sound.

The work presented in L. Bai [3] used a scheme for indexing and segmentation of video by analyzing the audio track using support vector machine. This analysis is then applied to structuring the sports video. They defined three audio classes in sports video, namely Play-audio, Advertisement-audio and Studio-audio based on the attributes of sports video. They used Support vector machine (SVM) for audio classification and then applied smoothing rules in final segmentation of an audio sequence. The results show the performance of SVM on audio classification is satisfying.

R. Shantha Selva Kumari [4] showed an improved feature vector formation technique for audio classification and categorization. This technique makes use of wavelets to extract the features of audio data. Wavelets are first applied to decompose the signal and to extract acoustical features such as sub-band power, brightness and band-width and pitch information. The additional features, such as frequency cepstral coefficients also extracted to accomplish audio classification and use a bottom-up Support Vector Machine over these acoustical features and additional features. The bottom-up Support Vector Machine categorization strategy uses an iterative procedure to match a given audio to progressively larger subsets or categories of classes.

An unsupervised clustering method is proposed [5], based on one-class support vector machines (OCSVM) and inspired by the classical K-means algorithm, which effectively classifies speech/music signals. First, relevant features are extracted from audio files. Then in an iterative K-means like algorithm, after initializing centers, each cluster is refined using a one-class support vector machine. The experimental results show that the clustering method, which can be easily implemented, performs better than other methods implemented on the same database.

3. Background

3.1 Audio Feature Extraction

The foundation of any type of audio analysis algorithm is the extraction of numerical feature vectors that characterize the audio content. In order to obtain high accuracy for audio classification, it is critical to select good features that can capture the temporal and spectral characteristics of audio signal and are robust for circumstance changing.

A variety of audio features have been proposed in the literature that can serve the purpose of audio tracks characterization. Generally, audio features can be classified into

two major groups: time-domain and frequency-domain audio features.

In our work, audio clip-level features are computed based on the frame-level features and used a clip as the classification unit. For ZCR, STE and SF we compute its mean of all frames in a given clip respectively as base clip-level features which is proved to be effective for distinguishing speech, music and speech with background sound [4,6].

Zero-crossing rate is proved to be useful in characterizing different audio signals. In general, speech signals are composed of alternating voiced sounds and unvoiced sounds in the syllable rate, while music signals do not have this kind of structure. Hence, for speech signal, its variation of ZCR will be in general greater than that of music. Also voiced sound is decided when the rate of zero crossing is low.

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

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

STE is the audio feature that is widely used and the easiest. It is also called volume. STE is a reliable indicator for silence detection. Normally STE is approximated by the rms (root mean square) of the signal magnitude within each frame. In the ratio of energy method, voiced sound is decided when the energy is high.

$$E(m) = \sum_m (x(n)W(n-m))\text{^}2 \quad (2)$$

where m is the time index of the short-term 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.

Usually non-speech has a low short-term energy but a high zero crossing rate. Combining ZCR and STE to prevent low energy unvoiced speech frames from being classified as silent.

Spectrum flux (SF) is defined as the average variation value of spectrum between the adjacent two frames in a given clip.

$$SF_t = \sum_{n=1}^{N} (N_t[n] - N_{t-1}[n])\text{^}2 \quad (3)$$

where $N_t[n]$ and $N_{t-1}[n]$ are the normalized magnitude of the Fourier transform at the current frame t, and the previous frame t-1, respectively. The spectral flux is a measure of the amount of local spectral change. Music has a higher rate of change, and goes through more drastic frame-to-frame changes than speech does; this value is higher for music than for speech.

For speech signal, the spectrum flux (SF) of environment sounds is among the highest and change more dramatically than those of speech and music. Based on the previous work [7], this feature is especially useful for discriminating some strong periodicity environment sounds such as tone signal, from music signals. SF is a good feature to discriminate among speech, environment sound and music.

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. The frequency bands are decided using the Mel-frequency scale (linear scale below 1kHz and logarithmic scale above 1kHz). MFCCs consist of preprocessing, framing, windowing, DFT, Mel Filterbank, Logarithm and Inverse DFT.

$$C_n = \sqrt{\frac{2}{k}} \sum_{k=1}^{K} (\log S_k) \cos[n(k-0.5)\pi/K] \quad (4)$$

MFCC becomes more robust to noise and speech distortion, once the Fast Fourier Transform (FFT) and Mel scale filter applied. MFCCs use Mel scale filter bank, where higher frequency filter have greater bandwidth than the lower frequency filter, but their temporal resolutions are the same [8].

MFCCs can be estimated using a parametric approach derived from Linear Prediction Coefficients (LPC), or using a non-parametric FFT-based approach. However, FFT-based MFCCs typically encode more information from excitation, while LPC-based MFCCs remove the excitation. FFT-based MFCCs are found to be more dependent on high-pitched speech resulting from loud or angry speaking styles than LPC-based MFCCs, witch were found more sensitive to additive noise in speech recognition tasks. This is so because LPC-based MFCCs ignore the pitch-based harmonic structure seen in FFT-based MFCCs.

Silence Ratio (SR) is defined as the ratio of the amount of silence in an audio piece of the length of the piece to the length of the piece. SR is a useful statistical feature for audio classification; it is usually used to differentiate noise from speech.

Noise Frame Ratio (NFR) is defined as the ratio of noise frames in a given audio clip. A frame is considered as a noise frame if the maximum local peak of its normalized correlation function is lower than a preset threshold. The NFR value of noise-like environment sound is higher than that for music, because there are much more noise frames of the previous class.

3.2 Support vector machine

SVM models the boundary between the classes instead of modeling 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.

3.2.1 Linear support vector machines

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 $x_i \in R_n$ is a feature vector and $y_i \in \{-1, +1\}$ is a class label, with a separating hyper-plane of equation $w \cdot x + b = 0$; 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:

$$f(x) = sgn\left(\sum_{i=1}^{l} \bar{\alpha}_i y_i x_i x + \bar{b}\right) \quad (5)$$

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.

3.2.2 Kernel support vector machines

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 [9], the kernel function implicitly maps the input vectors into a high dimensional feature space in which the mapped data is linearly separable. In our method, we will use the Gaussian Radial Basis kernel:

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

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 center of the RBF and σ will determine the area of influence this support vector has over the data space.

3.3 Decision tree

The decision trees are the single most popular data mining tool. It is easy to understand, implement, use and computationally cheap. They do classification that predict a categorical output from categorical and/or real input.

Learned trees can also be re-represented as sets of if-then rules to improve human readability. In a set of records, each record has the same structure, consisting of a number of attribute/value pairs. One of these attributes represents the category of the record. The problem is to determine a decision tree that, on the basis of answers to question about the non-category

attributes, predicts correctly the value of the category attribute.

In the decision tree, each node corresponds to a non-categorical attribute and each arc to a possible value of that attribute. A leaf of the tree specifies the expected value of the categorical attribute for the records described by the path from the root to that leaf.

4. Design of Proposed System

Many efforts in this area focus on audio data that contains some built-in semantic information structure such as in broadcast news, or focus on classification of audio that contains a single type of sound such as clear speech or clear music only. We should describe goodness of decision tree based SVM for mixed type audio.

In this work, an audio clip is classified into eight classes. We should figure out seven SVMs for classification as SVM is a two-class classifier. SVMs are learnt with the training data.

The motivation of combining the SVM and decision tree is to combine the strong generalization ability of the SVM and the strong comprehensibility of rule induction. In this research, we discuss about combining Support Vector Machine and decision trees for multi class audio classification. Since the SVM as a decision of binary tree to select strong instances to generate rules is used, we do not need to train the whole training set when we discriminate each audio clip and better accuracy and time saving can be expected upon the whole architecture of the system.

Firstly, audio file is extracted from video file for feature calculation. To extract this audio file, we put the audio quality with the sampling frequency of 22 kHz, bit rate of 128 kbps and mono channel. Then this audio stream is split 15sec long audio wav file.

In the second stage, the audio stream is analyzed into 1sec audio clips with 0.5 sec overlap and each clip is divided into frames. The frame size is 20ms for sampling frequency of 22 kHz and it is analyzed by non-overlapping.

In the next stage, features are analyzed and extracted in two levels: frame-level and clip-level by using audio features ZCR, STE, SF, MFCC, SR and NFR which is proved to be effective for distinguishing audio classes. But the characteristics of the feature components are so different that it is not appropriate to just put these features into a feature vector. Each feature component should be normalized to make their scale similar. The normalized feature vector is considered as the final representation of an audio clip.

Before performing the actual data mining process, for the sake of accuracy and efficiency, a pre-filtering process is needed to clean data. Machine (SVM) classifiers are used to group these features into predefined groups and each group is labeled with a keyword. At this step, SVM-DT algorithm employs the SVM as a preprocess of decision tree and consists of three major steps. First, this algorithm trains an SVM. Then a new training set can be generated by selecting from the result of the SVM. This new data set for training decision tree will be better than the original data set due to using the advantage of the SVM. Finally, this new training set is used to train a decision tree learning system and to extract the corresponding rule sets. The rules produced by combining the SVM and decision tree are then annotated based on encoding schemes and verified in the test data set according to the biological meaning. From this we can get the rules' accuracy.

Suppose we are given a training data set $S = \{(x_1,y_1), (x_2,y_2),\ldots,(x_m,y_m)\}$, where x_i is the feature vector and y_i is the expected class label or target of the i^{th} training instance. At first, SVMs are trained using N-fold cross validation. That is, for data set S, we divided it into N subsets with similar sizes (k) and similar distribution of classes. We perform the tests for the N runs, each with a different subset as the test set (Te-svm) and with the union of the other N-1 subsets as the training set (Tr-svm). Then, from each test set (Te-svm), based on the result of prediction, we select cases that are correctly predicted by the SVM into new data set (S-svm,). Finally, we use the original test data Te-svm as test data set (Te-svm-dt) and the union of the other N-1 subsets S-svm as the training set (Tr-svm-dt) to train a decision tree and induce the rule sets. In summary, the pseudo code of our SVM-DT algorithm is shown in Figure. 1.

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-svm-dt) from S-svm for decision tree and take Te-svm as Te-svm-dt
5: Run the Decision Tree algorithm for each data set (Tr-svm-dt, Te-svm-dt), get the rule set R

Figure.1. Algorithm: SVM-DT

After that SVM are trained on each class at each level of the tree and the SVM which is more successful in predicting a class at that level is selected as the decision in that node. Thus a tree is constructed with different SVM in each node. And the tree constructed is used for classifying the multi class audio.

SVM1 discriminate between speech and non-speech in the first level. After that those speech clips are classified into pure speech and excited speech by using SVM2 in the second level. At the same level, the non-speech clips are classified into environmental sound and background sound by using SVM3. Then, in the third level, pure speech, excited speech, environmental sound and background sound are classified into silence, non-silence, announcer's excited speech, speech over crowd, whistle, audience clapping, crowd only and cheering over crowd by using SVM4, SVM5, SVM6 and SVM7, respectively.

5. Experimental Study

5.1 Data Set

To validate the effectiveness and robustness of the proposed approach, five soccer videos are collected from various sources. In total we have seen 8 hours of soccer videos consisting of five gigabytes of data. They come from different sources, digitized at different studios, sampled at 22 kHz, and reported by different announcers.

Four of five soccer videos is used as the training data (e.g., announcer's excited speech, speech over crowd, audience clapping, whistle, crowd only, etc). Each of them is hand-labeled into one of 8 classes in soccer video which is composed of mixed audio with background interference such as applause and loud cheering.

The data was divided into training (80% of data) and testing (20%) sets. The experiments will be performed with same feature dimensions for each of the feature extraction methods. The total duration of one soccer video is approximately 1 hour and 45 minutes.

5.2 Experiment Setup

The experimental audio data come from soccer videos each having a total length of 1 hour and 45 minutes. Totally 8 hours of Premier League 2008-2009 soccer video will be used in this experiment. 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.

The occurrences of these events in soccer videos are quite different. In fact, speech over crowd and excited events are the most frequent while, for example, speech only events usually occur only during the break of the match. Silence is always correctly classified, followed by the speech only. This latter class is particularly interesting since it is related to the sequences that show the commentators in the studio, and thus may be used to segment and classify video shots. The speech over crowd is related to ongoing actions while the excited class is related to highlights such as shots on goal, placed kicks near the goal post, penalty kicks, etc. The crowd only class is related to shots showing actions but without the speech of commentators.

To briefly show the efficiency of our proposed system, we combined with the frame-level SVM and Decision tree for audio labeling. Cross-validation results are based to find the optimal SVM parameters such as RBF kernels, variance, margin and cost factor. Training and test

dataset will be taken according to a 5-fold cross-validation. 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.

6. Conclusion

This proposed system focus on developing an effective scheme to apply audio content analysis to improve high-level structures of soccer videos and indexing process. Building general audio classifiers for large "real-world" datasets is challenging. Various factors such as audio annotation, training speed and quality control, that are typically not important in smaller datasets, become crucial for designing and developing effective audio classification algorithms for large datasets.

In this paper, combining SVM and decision tree is presented for effective multi-label audio classification. Audio file will be extracted in two levels from video file. The features set consisting of MFCC, ZCR, STE, SF, SR and NFR will be used. Moreover binary-class approach for multi-label classification is used. Since the SVM as a decision of binary tree to select strong instances to generate rules is used, we do not need to train the whole training set when we discriminate each audio clip and better accuracy and time saving can be expected upon the whole architecture of the system.

References

[1] A. Rabaoui, H. Kadri, Z. Lachiri, and N. Ellouze, "One-Class SVMs Challenges in Audio Detection and Classification Applications", Hindawi Publishing Corporation EURASIP Journal on Advances in Signal Processing, Volume 2008, Article ID 834973.

[2] L. Chen, S. G¨und¨uz, M. Tamer O¨ zsu, "Mixed Type Audio Classification With Support Vector Machine", Proceedings of the 2006 IEEE International Conference on Multimedia and Expo, ICME 2006.

[3] L. Bai, S. Lao, H. Lia0, J. Chen, "Audio Classification And Segmentation For Sports Video Structure Extraction Using Support Vector Machine", IEEE, Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, 13-16 August 2006.

[4] R. Shantha Selva Kumari, D. Sugumar, V. Sadasivam, "Audio Signal Classification Based on Optimal Wavelet and Support Vector Machine", IEEE, International Conference on Computational Intelligence and Multimedia Applications (ICCIMA) 2007.

[5] S. Omid Sadjadi, S.M. Ahadi, O. Hazrati, "Unsupervised Speech/Music Classification Using One-Class Support Vector Machines", IEEE, Information, Communications & Signal Processing, 2007 6th International Conference on Volume , Issue.

[6] L Lu, HJ Zhang, SZ Li, Content-based Audio Classification and Segmentation by using SVM, Multimedia Systems, 2003-Springer Digital Object Identifier (DOI) 10.1007/s00530-002-0065-0, Journal Article (2003)

[7] L. Lu, H. Jiang, H. J. Zhang, A Robust Audio Classification and Segmentation Method. Proc. of the 9th ACM international conference on Multimedia, pp. 203–211, 2001

[8] Z. Razak, N. Jamaliah Ibrahim, E. Mohd Tamil, "Quranic Verse Recitation Feature Extraction Using Mel-Frequency Cepstral Coefficient (MFCC)", Research, 9 April 2008.

[9] R. Duda, P. Hart, and D. Stork, “Pattern Classification”, John Wiley & Sons, New York, 2000