SELECTING STRONG DECISION RULES FOR VIDEO SCENE DETECTION

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

Abstract: Video is a rich source of information which is composed of multimodal information streams such as visual, auditory, and textual streams. 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. Although most of the classifiers are considered in research areas, the problem of constructing good classification method is always ongoing research issue. 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 and when we construct the binary tree architecture, we use the SVM as a decision; it is known as decision support vector machine (DSVM). This classification and analysis is intended to analyze the structure of the sports video in the context of 8 classes. The experimental results by using audio track features for sports video indicate that good accuracy is achieved with time saving in training phase.

Keywords: SVM, Decision Tree, Content Analysis, Audio

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. Research towards the automatic detection and recognition of events in sport videos data has attracted a lot of attention in recent years.

Design of efficient indexing techniques to retrieve relevant information is another important requirement. Allowing for possible automatic procedures to semantically index audio-video material represents therefore a very important challenge. Such methods should be designed to create indices of the audio-visual material, which characterize the temporal structure of a multimedia document from a semantic point of view.

The actual architecture of a system supporting video annotation and retrieval depends on the application context and, in particular, on end users and their tasks. Although all application contexts demand a reliable annotation of the video stream to effectively support selection of relevant video segments, it's evident that, for instance, service providers (such as broadcasters and editors) and consumers accessing a video-on demand service have different needs [1].

For example, videos add the temporal dimension, requiring object dynamics. Furthermore, although people often think of a video as just a sequence of images, it's actually a compound medium, integrating diverse media such as realistic images, graphics, text, and audio [2]. Also, application contexts for videos are different than those for images and therefore call for different approaches to help users annotate, query, and exploit archived video data. Researchers have also reported on concrete video retrieval applications by high-level semantics in specific contexts such as movies, news, and commercials [3, 4]. Due to their enormous commercial appeal, sports videos represent an important application domain to produce programs that contain one day's best sports actions.

We should be able to automatically annotate video material, which is typically captured live, because detailed manual annotation is mostly impractical. To achieve an effective annotation, we should have a clear insight into the current practice and established standards in the domain of professional sports videos, particularly concerning the nature and structure of their content. We analyze specificity of data and provide an overview on the

rationale underlying how we selected relevant features.

Soccer video analysis and events/highlights extraction are probably the most popular topics in this research area. Soccer is a very popular sport, the whole game is quite long, often there are several games being played on the same day or at the same time and viewer may not be able to watch all of them. Users desire the capability to watch the programs time-shifted (on-demand) and/or desire to watch only the highlights such as goal events in order to save time. Recently, audio contents become more and more important clues for detecting events relating to highlights across different sports, because different sounds can indicate different important events. There have been many works on integrating visual and audio information to automatically generate highlights for sport programs.

We focus on detecting highlights using audio track features alone without using video track features. Visual information processing is computation intensive; it has limited available computational power. An important role plays the choice of appropriate representation of the only audio information and the selection of suitable features that can be extracted from the audio track of the video stream in order to perform a successful detection. We believe that audio information carries out by itself a rich level of semantic significance, and this paper focuses on this issue. The application area of this classification work could have a variety of applications, such as indexing, retrieval and browsing of archives of broadcast news program or sports videos.

In this paper, we discuss about combining Support Vector Machine and decision trees for multi class audio classification. It is specifically focused on extraction of information from the audio stream of soccer video. Since the SVM as a decision of binary tree to select strong instances to generate rules is used, it is not need to train the whole training set when it is to discriminate each audio clip and better accuracy and time saving can be expected.

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 an overview of linear, kernel SVM and decision tree. In Section IV, a method for multi-class classification is proposed and in section V, experimental evaluation is presented. Finally, in Section VI, 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 [5] illustrated the potential of SVMs on recognizing impulsive audio signals belonging to a complex real world dataset. They presented a method to apply optimized one-class support vector machines (1-SVMs) to tackle both sound detection and classification tasks in the sound recognition process.

Kim et al. [6] fused visual analysis that classifies pitching and close-up shots with audio events related to cheering, to detect scoring highlights in baseball videos. The audio features used are based on MDCT coefficients of AC-3, and classification is done using SVMs. Xu et al.

Xu et al. [7] have proposed a system that recognizes several generic sport audio concepts (e.g. whistling, excited speech) and domain specific; feature vectors, composed by a combination of different audio descriptors (Mel-frequency and linear prediction cepstral coefficients, etc.), are processed by SVMs for feature selection and classification.

3. Background

3.1. Audio Feature Extraction

A variety of audio features have been proposed in the literature that can serve the purpose of audio track characterization [8]. Generally they can be divided into two categories: physical features and perceptual features. The perceptual feature describes the perception of sounds by human beings. Loudness, pitch, brightness and timbre are examples of these features. The physical features such as zero crossing rate, MFCC, energy and spectral centroid are further grouped into spectral features and temporal features according to the domain in which they are calculated [9].

To detect main highlight events, only a few of these features are useful. The Mel-scale Frequency Cepstral Coefficients (MFCCs) approach was for a long time widely used on the area of speech recognition. It is also one of the most used features for audio classification. In the soccer videos, the sound track mainly includes the foreground commentary and the background crowd noise. Based on the observation and prior knowledge, we assume exciting segments are highly correlated with

announcers' excited speech and the audience ambient noise can also be very useful, as audience viscerally react to exciting situations.

The audio signals recorded from raw soccer video sequences are divided into 1 second length clip with a 50% overlap between the adjacent clip. To apply the feature extraction, each clip is further divided into 20ms frames without overlapping. For features such as zero-crossing rate (ZCR), short-time energy (STE) and spectral flux (SF), 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 speech with background sound [10,11].

3.2 SVM

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.

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

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. Proposed System

In this paper, we present an automatic multi labeled audio classification in soccer videos by using audio track features alone without relying on expensive-to-compute video track features. The classified audio classes can be used for high-level indexing and selective browsing of soccer videos. We should describe goodness of decision tree based SVM for mixed type audio.

To evaluate our system we collected three soccer videos from various sources. In total we have five hours of soccer games consisting of five gigabytes of data. Two of three soccer videos is used as the training data (e.g., announcers' excited speech, audience crowd, audience clapping, cheering over crowd).

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 paper, a system is proposed employing the algorithm SVM-DT where 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. It 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.

In this work, an audio clip is classified into one of 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 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, SF, MFCC, SR and NFR 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. 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 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
where Tr-svm is train-data-set and Te-svm is test-data-set.

Figure 1. SVM-DT algorithm

After SVMs are trained on each class at all levels of the tree and the SVM becomes more successful in predicting a class at that level is selected as the decision in that node. Next, a tree is constructed with different SVMs in each node. For discriminating 14 audio classes seven SVMs are used in this approach. The tree constructed is finally used for classifying the multi-class audio. Our proposed system design is illustrated in Figure. 2.

5. Experimental Evaluation

5.1 Data

As the preliminary investigation, two soccer videos are used as the training data containing classes such as announcer's excited speech, speech over crowd, audience clapping, whistle, crowd only … etc. To obtain the ground truth, each of them is hand-labeled into one of 8 classes in soccer video in which mixed audio with background interference such as applause and loud cheering are also contained. The data was divided into training (70 % of data) and testing (30 %) sets. The experiments are 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. Totally 5 hours of Champion 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. 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 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 are taken according to a holdout 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 its respective class.

5.2 Results

All the experiments have been performed on a real world dataset, consisting of more than two hours of soccer videos in Thai language. Training and test

dataset have been taken according to a 2-fold cross-validation. We compare the proposed SVM-DT classification method to SVM that is widely used in the literature as classification method.

Table 1 summarizes features employed for decision tree SVM where rows refer to extracted features and columns are the classes at all three classification levels. For example, at the first level of classification tree, to discriminate between speech and non-speech clips, features such as ZCR and MFCC can be applied. Table 2 reports the accuracy and error recognition rate (ERR) resulted from the tests for 11399 clips. Length of a clip is 1 second. In this initial investigation, only nine classes of audio are tested. Table 3 compares the accuracy of SVM-DT over SVM. For all classes, the recognition rates using SVM-DT is outperformed than using ordinary SVM. Overall classification accuracy is exceeded over 80%. Note that there is strong agreement of manual and automatic classification.

On going work of this research is to find the best decision parameter sets for classifying the other 2 audio classes (i.e., silence vs non-silence in frame level.

Table 1. Encoded Table for Classes and Selected Features

Features / Classes	ZCR	STE	SF	NFR	SR	MFCC
Speech	1	0	0	0	0	1
Non-speech	1	0	0	0	0	1
Pure Speech	0	0	0	1	0	1
Commentator's Speech	0	0	0	1	0	1
Background Crowd	0	0	0	0	0	1
Environmental Sound	0	0	0	0	0	1
Silence	0	0	1	0	1	0
Non-silence	0	0	1	0	1	0
Commentator's Excited Speech	1	1	0	0	0	0
Speech Over Crowd	1	1	0	0	0	0
Cheering Over Crowd	0	0	0	0	0	1
Crowd Only	0	0	0	0	0	1
Whistle	0	0	0	1	0	1
Audience Clapping	0	0	0	1	0	1

Table 2. Accuracy and Error Recognition Rate

Classes	Accuracy (%)	ERR (%)
Speech	97.05	2.95
Commentator's Speech	97.52	2.48
Background Crowd	92.54	7.46
Environmental Sound	97.20	2.80
Speech Over Crowd	89.23	10.77
Commentator's Excited Speech	88.37	11.63
Cheering Over Crowd	87.40	12.60
Crowd Only	80.04	19.96
Audience Clapping	80.02	19.98

Table 3. Classification Accuracy of SVM and SVM-DT

Classes	SVM (%)	SVM-DT (%)
Speech	90.84	97.05
Commentator's Speech	89.93	97.52
Background Crowd	64.76	92.54
Environmental Sound	90.14	97.20
Speech Over Crowd	86.93	89.23
Commentator's Excited Speech	80.88	88.37
Cheering Over Crowd	63.06	87.40
Crowd Only	79.99	80.04
Audience Clapping	60.34	80.02

6. Conclusion

This proposed system focus on developing an effective scheme to apply audio content analysis for improving video structure parsing and indexing process. In this research, effective multi-label audio classification is presented combining SVM and decision tree. The six feature sets; ZCR, STE, SF, SR, NFR and MFCC are used. The SVM is used as a preprocessing of decision tree to select strong instances to generate rules. The SVM is also inserted as a decision of binary tree to select strong instances to generate rules is used. The performance of the algorithm is found to be impressive in terms of accuracy and error rates. Future direction of this work will include to obtain the strong rules for classes (whistle and clapping) incorporation of other

features if it is necessary and to conduct a series of experiments.

References

1. N. Dimitrova et al., "*Entry into the Content Forest: The Role of Multimedia Portals*," IEEE MultiMedia, vol. 7, no. 3, July–Sept. 2000, pp. 14-20.
2. R.S. Heller and C.D. Martin, "*A Media Taxonomy*," IEEE MultiMedia, vol. 2, no. 4, Winter 1995, pp. 36-45.
3. S. Eickeler and S. Muller, "*Content-Based Video Indexing of TV Broadcast News Using Hidden Markov Models*," Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP), IEEE Press, Piscataway, N.J., 1999, pp. 2997-3000.
4. H. Miyamori and S.-I. Iisaku, "*Video Annotation for Content-Based Retrieval Using Human Behavior Analysis and Domain Knowledge*," Proc. Int'l Conf. Automatic Face and Gesture Recognition 2000, IEEE CS Press, Los Alamitos, Calif., 2000.
5. 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.
6. H.-G. Kim, J. Jeong, J.-H. Kim, and J.Y. Kim, "*Real-time highlight detection in baseball video for TVs with time-shift function*," IEEE Trans. on Consumer Elec., vol. 54, no. 2, 2008.
7. M. Xu, C. Xu, L. Duan, J. Jin, and S. Luo, "*Audio keywords generation for sports video analysis*," ACM TOMCCAP, vol. 4, no. 2, 2008.
8. Z. Liu, Y. Wang, and T. Chen, "*Audio Feature Extraction and Analysis for Scene Segmentation and Classification*", Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology, 1998, pp. 61-80, 1998.
9. Vesa Peltonen, Juha Tuomi, Anssi Klapuri, Jyri Huopaniemi, Timo Sorsa "*Computational Auditoy Scene Recognition*", ICASSP 2002, Peltonen, V.; Tuomi, J.; Klapuri, A.; Huopaniemi, J.; Sorsa, T.; Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on , Volume: 2 , 2002Pages:1941 – 1944.
10. 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.
11. 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).

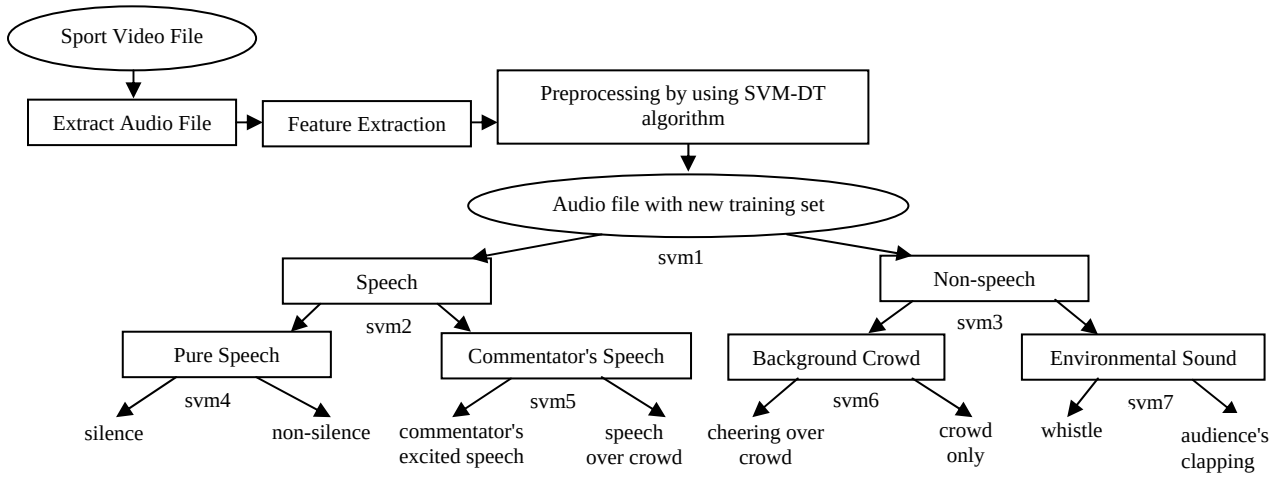

Figure 2. Proposed System Design