EXPLORING AUDIO SEMANTIC CONCEPTS FOR EVENT-BASED VIDEO RETRIEVAL

Yipei Wang, Shourabh Rawat, Florian Metze

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, U.S.A
{yipeiw, srawat, fmetze}@cs.cmu.edu

ABSTRACT

The audio semantic concepts (sound events) play important roles in audio-based content analysis. How to capture the semantic information effectively from the complex occurrence pattern of sound events in YouTube quality videos is a challenging problem. This paper presents a novel framework to handle the complex situation for semantic information extraction in real-world videos and evaluate through the NIST multimedia event detection task (MED). We calculate the occurrence confidence matrix of sound events and explore multiple strategies to generate clip-level semantic features from the matrix. We evaluate the performance using TRECVID2011 MED dataset. The proposed method outperforms previous HMM-based system. The late fusion experiment with the low-level features and text feature (ASR) shows that audio semantic concepts capture complementary information in the soundtrack.

Index Terms— multimedia retrieval, audio processing, semantic concept

1. INTRODUCTION

The amount of multimedia data on the Internet has been growing exponentially in the past decades and is expected to continue growing in the future. This yields advanced technologies in multimedia indexing and retrieval.

Considerable research effort has been invested in studying audio-based content analysis. The proposed methods can be generally divided into two major directions. One direction is relying on a direct analysis of the low-level features [1]. The other direction is to detect sound events to bridge the gap between the low-level features and the high-level clip-unit semantics [2] [3] [4] [5] [6]. Recent studies have shown that the second direction outperforms the pure feature-based approach. Foote [3] proposed a music and sound effects retrieval system where mel-frequency cepstral coefficients (MFCCs) plus energy were used as feature vectors. A tree-based vector quantizer (VQ) was applied on the feature vector space to partition it into regions. Sounds were classified by calculating the Euclidean or cosine distances between the histograms of VQ codeword usage within each sound. Lie et al.[4] used SVM classifiers with perceptual and cepstral features on the Muscle Fish data.

A lot of work in sound event detection mainly adopt on methods in speech recognition and speaker identification. For example, hidden markov models (HMM) were employed in [5] to detect 42 semantic concepts defined by a human expert. Another work [6] used support vector machine (SVM) and Gaussian mixture models (GMM) as feature summarization to detect clip-level concepts including 25 overlapping classes (the definition of class exist overlaps).

However, these methods achieve poor performance in the YouTube quality videos, which often include a variety of composite sound events with overlapping or irregular occurrence pattern. The sound events often show complex distributions in feature space.

There are several problems in previous methods mentioned above. The GMM-SVM based method requires calculating the statistical feature representation on the sound event segments. This is not hard to be satisfied in real-life videos since the boundary of sound events is often unknown. The HMM-based method is able to deal with the unknown boundary problem. But it is hard to detect mixed sound events since we often have no training data for the composite sound events. As a generative model, the performance is highly affected when there is rare labeled data. Besides, both of the methods are based on speech-related frequency features. However, due to the differences in the physical natures of sound events, it is expected to make use of other types of feature, e.g. energy and pitch.

Beyond audio semantic concepts detection, extracting effective representation from the detected occurrence of sound events to distinguish video clips requires further research. Recent work mainly generates the bag-of-word feature over the clip [7] for the frequency of concurrency. This method fails to retrieve temporal information, which often provides evidence for clip-level semantics.

The real-world YouTube quality videos yield more robust method to detect the semantic information and effective representation to reflect the information from the complicated occurrence pattern. Thus, we propose a novel framework to solve the problem. We make use of multiple types of features and build classifiers to calculate the occurrence confidence matrix, which embeds richer information than pure prediction result. Then, we investigate heuristic strategies to generate features for event-based retrieval task from the confidence matrix, which attempts to retrieve meaningful information for the video clip. We extend the bag-of-word method to include more temporal information to assist the performance.

The rest of the paper is organized as follows. In section 2, we give an overview of the system. In section 3, we introduce the annotation of audio semantic concepts and analyze the occurrence pattern. The analysis reflects the complex nature of the problem. In section 4, we evaluate the performance on TRECVID2011 multimedia event detection dataset. We further analyze the feature importance for sound event detection in section 4.1. We describe the comparison and fusion experiment with text feature and the low-level feature in section 4.2. Section 5 concludes the paper.

2. SYSTEM OVERVIEW

The system works in three steps as is shown in Fig.1. First, we extract the statistics of acoustic feature over fixed length of sliding windows. Then, we apply classifiers trained on annotations of audio semantic concepts to calculate the occurrence confidence matrix, where each column is an occurrence confidence vector for the corresponding sample. Each dimension in the confidence vector represents the prediction confidence for a semantic concept. Finally, we

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

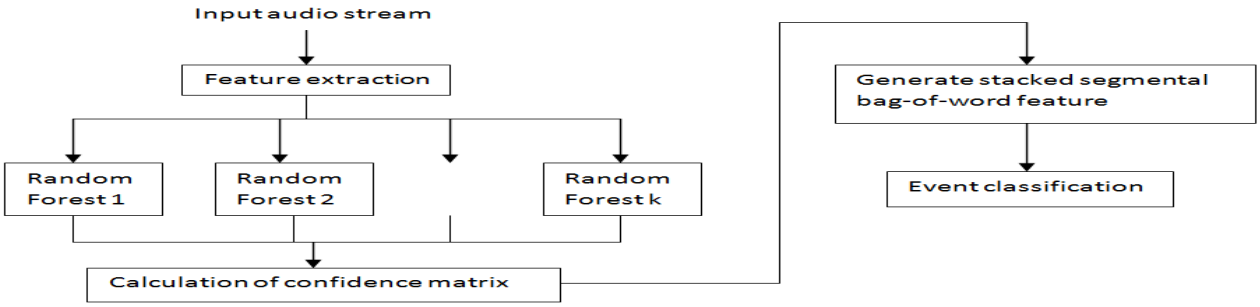

Fig. 1: System Overview

generate clip-level feature from the matrix and evaluate its effectiveness in the multimedia event detection task. The event classification pipeline [8] [9] uses one-against-all rigid kernel regression classifier using χ^2 kernel for each event over bag-of-word feature.

2.1. Feature extraction

We use the toolkit OpenSmile [8] (speech and music interpretation by large space extraction) to extract standard acoustic features, including MFC (Mel-frequency cepstrum), MFCC (Mel-frequency cepstral coefficients), Log-energy and F0 (fundamental frequency). Then, we apply a set of statistical functions over these acoustic features within the fixed length sliding window. The feature set is listed in Table 1.

The left column lists the name of the acoustic features. The right column lists the functions applied on the acoustic features. The total feature dimension is 983.

MFCC(13+13 delta+13 delta_delta)	arithmetic mean, standard deviation, linear regression parameter, quartiles, range
MFC (26+26 delta)	The same set with log-energy + kurtosis
Log-energy	quadratic mean (non-zero value), standard derivation, max Pos, min Pos, linear regression parameter, quartiles, range
F0	quadratic, standard deviation, maximum mean, minimum mean, linear regression, qiartiles, range

Table 1: Feature set list

Besides mean and variance, we also apply other statistics to capture the characteristics of the envelope of acoustic features within the window. For example, the linear regression slope and approximation error rate perform to describe the general changing trend within the window. The quartiles function and extreme point positions are used to describe the shape of the envelope more accurately. However, attempts of using complex statistics to describe the envelope explicitly should be avoided. This would introduce additional computational cost and noise in feature space.

2.2. Classifier

We train classifiers using random forest on the annotations to detect 40 pre-defined semantic concepts. We use the implementation of random forest provided by the toolkit scikit-learn [10]. The random forest uses bootstrapped samples and randomly selected subset of features to grow each decision tree. The prediction is performed by the voting of all the decision trees. The implicit feature selection in building random forest brings the advantage to make use of discriminative feature subspace for classification. In addition, the efficiency of handling high dimensional feature makes it possible to process large-scale data. Compared to SVM [11], random forest is more robust in dealing with high dimensional noisy features with much lower computational cost in training. Here we exploit two strategies to retrieve the occurrence confidence vector of semantic concepts using random forest.

2.2.1. Multi-class classifier

We build one multi-label classifier using random forest on the balanced resampled training data. The predicted label is one of the multiple classes. But we use the prediction probability for each class to get the occurrence confidence vector. This method assumes that the estimated prediction probabilities provide evidence of overlapping sound events.

2.2.2. Multiple independent binary classifiers

We build binary classifier using random forest for each audio semantic concept on balanced resampled training data. We exclude the overlapped labels in the negative samples if there exist overlapping sound events. The prediction probability from each binary classifier is used as the occurrence confidence for the corresponding semantic concept.

2.3. Generate feature representation for event classification

We explore different strategies to generate clip-level feature for event classification from the occurrence confidence matrix. To take the balance of the computational complexity and remaining rich information, people often apply statistical functions over the confidence vectors. To find effective representation of the sound event occurrence for event classification, there are two main issues we need to solve: 1) which statistical function is powerful in retrieving information for event classification; 2) how to retrieve the information carried in the temporal pattern in the feature vector.

To address the first issue, we investigated several common used statistics, including the mean value, the maximum value and the standard variance. For the second issue, we propose a solution as described below. Each video is split into a certain number of segments with equal length. Then, we apply the proper statistical function on each segment and stack the results into one feature vector.

3. ANNOTATION

3.1. Introduction

An expert labeled environmental noise for the soundtracks of around 400 videos taken from the training set [12]. The annotation solely referred to the audio cue. Some concepts only appear in several clips. We filtered out clips with such concepts from the training set and 40 concepts are remained finally. They are: crowd (L1), laugh (L2), mumble (L3), speech (L4), speech_ne (L5), cheer (L6), music (L7), music_sing (L8), whistle (L9), squeak (L10), animal (L11), anim_bird (L12), anim_cat (L13), anim_dog (L14), scream (L15), child (L16), singing (L17), tone (L18), human_noise (L19), rustle (L20), scratch (L21), micro_blow (L22), white_noise (L23), washboard (L24), applause (L25), wind (L26), engine_quiet (L27), engine_light (L28), power_tool(L29), engine_heavy (L30), radio (L31), water (L32), knock (L33),thud (L34), clap (L35), click (L36), bang (L37), beep (L38), clatter (L39), hammer (L40).

3.2. Analysis

There is a variety of sound events with overlapping or irregular occurrence pattern in multimedia stream. Therefore, we analyzed the annotation to get an idea of the characteristic of the problem.

First, we surveyed the amount of time each concept appeared within all the video clips (Fig.2). The concept of "speech", "music", "music_sing" and "crowd" occurred more often than other concepts. Each of these concepts makes up over 10% in all the video clips. Most of other concepts occurred rarely. But their occurrences are important in deciding which event the video clip belongs to. The highly unbalanced distribution of semantic concepts makes it difficult to detect those concepts with rare occurrence.

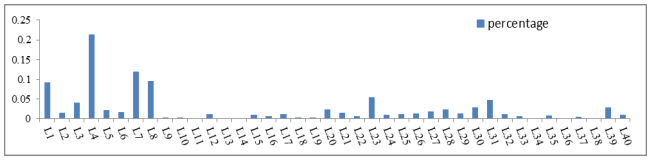

Fig. 2: The frequency of distribution of audio semantic concepts

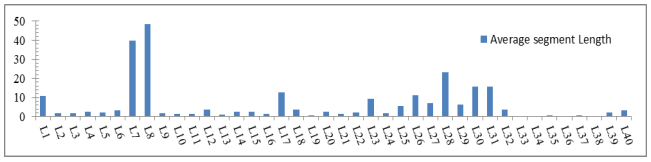

Fig. 3: The average lasting length of audio semantic concepts

Besides, we also investigated the average lasting time of each concept (Fig.3).

Over 30% of the soundtrack is composed of mixed sound events. Therefore, we analyzed the co-occurrence pattern. As is shown in Fig.4, some concepts often occur together, e.g. crowd, cheer and applause. Some concepts often occur with other sound, like animal. The co-occurrence pattern makes the detection more challenging.

4. EXPERIMENTS

The experiments are conducted on the development data from the TRECVID 2011 Multimedia Event Detection (MED) task [13]. The

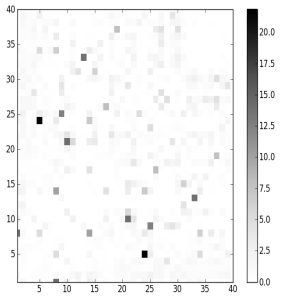

Fig. 4: point-wise mutual information matrix. The axis represents the index of the semantic concept.

dataset includes 3104 video clips for training, and 6642 video clips for testing. In the annotation data, we separate 90% for training and 10% as development data to tune the parameter of the classifier.

In order to take the representative samples as training data and reduce the computational cost, we apply the following strategy in selecting training samples. If the sound event last longer than the fixed length window, the training samples are selected from the windows falling into the segment with half window length shift. Otherwise, the training sample is selected from the window centered at the sound event segment. Since the boundaries of potential sound events are unknown, the samples for development and test data are extracted from sliding windows with 100ms shift. The number of decision trees in random forest is optimized on the development data.

The evaluation criteria [13] include average Pmiss@TER=12.5, minimum NDC and MAP. The lower Pmiss@TER=12.5 indicates better performance. Also, the lower minimumNDC indicates better performance. The higher MAP means better performance.

	MAP	Pmiss@TER=12.5	minNDC
HMM[5]	—	0.746	0.948
Multi-class(average)	0.0836	0.6846	0.9357
Multiple binary classifiers (average)	0.101	0.677	0.925

Table 2: The performance of HMM-based system and random forest based system

As is shown in Table 2, we first compared the proposed method with previous HMM based method (MAP is not an official evaluation criteria when the work [5] is done). We simply take the average of all the confidence vectors in our proposed framework and use 2s window in feature extraction. As is shown in Table 2, we can observe that our random forests based system outperforms the HMM based system. The system using multiple binary classifiers achieves better performance than using a multi-label classifier.

We explored multiple statistical functions to aggregate the confidence vector within a video to generate the bag-of-word feature for event classification, including taking the maximum, taking the mean and taking the standard variance. The results (Table 3) show that concatenating the mean and standard variance of all the confidence vectors achieves the best performance in general.

To investigate method to retrieve rich information from the temporal occurrence pattern, we apply the splitting strategy mentioned

	MAP	Pmiss@TER=12.5	minNDC
2s maximum	0.0879	0.6686	0.926
2s mean	0.101	0.677	0.925
2s mean + std var	0.11158	0.6746	0.9116

Table 3: The performance of system deployed different statistical functions to generate the feature for event detection.

	MAP	Pmiss@TER=12.5	minNDC
2s segment 1	0.1158	0.6746	0.9116
2s segment 3	0.1266	0.6640	0.9059
2s segment 5	0.1330	0.6588	0.9050
2s segment 10	0.1309	0.6427	0.9004
Fusion 2s (1, 3, 5, 10)	0.1423	0.6477	0.8944
Fusion 0.5s (1, 3, 5, 10)	0.1391	0.6392	0.8895
Fusion 2s + 0.5s (best)	0.1517	0.6308	0.8855

Table 4: The performance of system applied splitting strategies in generating feature for event detection.

in section 2.3 with the experimentally gained statistical function (mean and standard variance). We heuristically split the video into 3, 5 and 10 parts with equal length and apply this method. In Table 4, the results show that we get improvement with certain parameters. This proves that the temporal occurrence characteristics of the sound events is helpful in event classification.

Using shorter window in feature extraction is more effective to capture the sound with short duration. This is because longer window might include segments belonging to other sound events. But the statistics over shorter window might not include sufficient information to distinguish multiple sound events. In order to make use of both of their advantages, we take the late fusion of the system using 2s window and the system using 0.5s window. The fused system achieves the best performance.

4.1. Feature analysis

Random forest performs an implicit feature selection with low computational cost in dealing with high dimensional data. The outcome of the implicit feature selection can be visualized by the Gini importance [9], which provides relative ranking of features. We considered the gini importance calculated in training the random forest classifier for each semantic concept. In Fig.5, each row represents the importance scores of all feature variables for a particular semantic concept. We can see that different kinds of semantic concepts depend on different subset of feature set. Further analysis indicates that most of the concepts depend on MFC feature but vary in different frequency band. Some concepts more depend on energy feature and pitch feature.

4.2. Comparison with low-level feature and ASR

For low-level feature, we take the MFCC k-means system [5]. Table 5 summarizes the result. We can observe that the fused system outperforms the MFCC k-means system.

For text feature, we run Janus [14] to get ASR transcription and calculate TF-IDF as clip-level feature [5]. Table 6 summarizes the

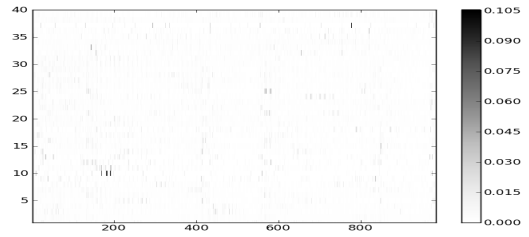

Fig. 5: Horizontal axis represents the index of the feature. 1 to 10 are energy feature, 11 to 127 are MFCC feature, 128 to 413 are MFC feature and 414 to 433 are F0 feature. The dimensions of 434 to 856 are the corresponding delta feature and the remain dimensions are the delta-delta feature of MFCC and F0.

	MAP	Pmiss@TER=12.5	minNDC
MFCC k-means	0.1985	0.5591	0.8268
Semantic best	0.1517	0.6308	0.8855
Fused System	0.2180	0.5453	0.8177

Table 5: The performance of semantics features, low-level feature and the fused system.

result. We can observe that the fused system got significant improvement. We find out that only around 30% of the sound track includes speech. For segments without speech, semantic concepts play an important role in capturing the semantics in the video.

	MAP	Pmiss@TER=12.5	minNDC
Text feature	0.1318	0.6977	0.8868
Semantic best	0.1517	0.6308	0.8855
Fused System	0.1947	0.5867	0.8334

Table 6: The performance of semantics features, text feature and the fused system.

5. CONCLUSIONS

We present our framework to effectively capture the audio semantics in dealing with the real-world YouTube quality videos. The experiment results show that our random forest based system outperforms previous HMM based system in MED task. We also compare the semantic features with text feature and low-level feature for the event-based retrieval task. The combined system through late fusion proves that semantic features provide complementary information.

6. ACKNOWLEDGMENTS

The work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contact number D11PC20068. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

7. REFERENCES

[1] Zhu Liu, Yao Wang, and Tsuhan Chen, "Audio feature extraction and analysis for scene segmentation and classification," *VLSI Signal Processing*, vol. 20, no. 1-2, pp. 61–79, 1998.

[2] Erling Wold, Thom Blum, Douglas Keislar, and James Wheaton, "Content-based classification, search, and retrieval of audio," *IEEE MultiMedia*, vol. 3, pp. 27–36, 1996.

[3] Jonathan Foote, "Content-based retrieval of music and audio," *SPIE*, vol. 3229, no. 6, pp. 138–147, 1997.

[4] Lie Lu, HongJiang Zhang, and Stan Z. Li, "Content-based audio classification and segmentation by using support vector machines," *Multimedia Syst.*, vol. 8, no. 6, pp. 482–492, 2003.

[5] Qin Jin, Peter Franz Schulam, Shourabh Rawat, Susanne Burger, Duo Ding, and Florian Metze, "Event-based video retrieval using audio," in *INTERSPEECH*, 2012.

[6] Keansub Lee and Daniel P. W. Ellis, "Audio-based semantic concept classification for consumer video," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 18, no. 6, pp. 1406–1416, 2010.

[7] Gerard Salton and Michael McGill, "Intorduction to modern information retrieval," *McGraw-Hill Book Company*, 1984.

[8] Lei Bao, Longfei Zhang, Shoou-I Yu, Zhen zhong Lan, Lu Jiang, Arnold Overwijk, Qin Jin, Shohei Takahashi, Brian Langner, Yuanpeng Li, Michael Garbus, Susanne Burger, Florian Metze, and Alexander G. Hauptmann, "Informedia @ trecvid2011," in *TRECVID*, 2011.

[9] Lu Jiang, Alexander G Hauptmann, and Guang Xiang, "Leveraging high-level and low-level features for multimedia event detection," in *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012, pp. 449–458.

[10] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay, "Scikit-learn: Machine learning in python," *CoRR*, vol. abs/1201.0490, 2012.

[11] Corinna Cortes and Vladimir Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[12] Susanne Burger, Qin Jin, Peter F. Schulam, and Florian Metze, "Noisemes: Manual annotation of environmental noise in audio streams," *Technical report CMU-LTI-12-07*, 2012.

[13] TRECVID, "http://www-nlpir.nist.gov/projects/tv2013/tv2013.html," .

[14] Hagen Soltau, Florian Metze, Christian Fügen, and Alex Waibel, "A one-pass decoder based on polymorphic linguistic context assignment," *In Proc. Automatic Speech Recognition and Understanding (ASRU), Madonna di Campiglio, Italy*, 2001.