5-2014

Semi-automatic audio semantic concept discovery for multimedia retrieval

Yipei Wang
Carnegie Mellon University

Shourabh Rawat
Carnegie Mellon University

Florian Metze
Carnegie Mellon University, fmetze@andrew.cmu.edu

Published In

Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1375-1379.

SEMI-AUTOMATIC AUDIO SEMANTIC CONCEPT DISCOVERY FOR MULTIMEDIA 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

Huge amount of videos on the Internet have rare textual information, which makes video retrieval challenging given a text query. Previous work explored semantic concepts for content analysis to assist retrieval. However, the human-defined concepts might fail to cover the data and there is a potential gap between these concepts and the semantics expected from user's query. Also, building a corpus is expensive and time-consuming. To address these issues, we propose a semi-automatic framework to discover the semantic concepts. We limit ourselves in audio modality here. In the paper, we also discuss how to select meaningful vocabulary from the discovered hierarchical sub-categories and provide an approach to detect all the concepts without further annotation. We evaluate the method on NIST 2011 multimedia event detection (MED) dataset.

Index Terms— audio semantic concept discovery, semi-automatic, multimedia retrieval

1. INTRODUCTION

There is a continuing growth of video collections available for searching over the Internet. Given a query, search engine retrieve relevant videos by analyzing their captions or textual descriptions. This initial method faces big problem. A large proportion of the videos are lack of detailed textual information. Even for those with rich textual descriptions, the gap between the content in the video and the given textual information is inevitable. To address the problem, advanced technologies in multimedia content analysis have become very popular recently.

Previous work has explored detecting semantic concepts in multiple modalities to capture the embedded semantics in the multimedia stream [1] [2] [3]. In this paper, we limit ourselves to discuss the audio semantic concepts. The audio semantic concepts are often defined by human based on their understanding of the specific application and their observations of limited data. After annotating a training corpus of these defined concepts, people apply multiple supervised methods for the semantic concept detection. Previous studies have adopted methods in speech recognition and speaker identification [2] [3]. These approaches have been shown to be effective on certain dataset for certain application.

However, there are potential problems of these approaches. Firstly, defining a proper vocabulary of the semantic concepts is time-consuming. Human experts have to observe a large amount of data and summarize the patterns into a number of semantic concepts based on their understanding. In this situation, it is very likely that these semantic concepts fail to cover the data and the vocabulary might be ineffective to retrieve the information needed for the application. These problems become more serious when new videos are added to the collection continuously in practical application. Secondly, the generalization of the human defined vocabulary is another problem. The domain-specific semantic concepts become useless in new domains. And other applications might yield refined concepts to distinguish the subtle difference in semantics. Finally, the lack of generalization introduces further problem. Previous attempts to detect the semantic concepts largely adopted supervised method, which requires large training data to achieve good performance. If we have to build an annotation each time for new data or a new application, it would be extremely expensive.

Previous studies also exploited unsupervised method to learn acoustic units automatically [4] and deployed methods to map them to high-level semantics [5] [6]. The main idea is to learn a sound dictionary, which is similar to learning a phone dictionary in speech recognition. Then, the audio recording can be decoded into a sequence of the basic units in the dictionary. Higher level semantics can be identified through the occurrence pattern of the basic units. The unsupervised method is easy to be generalized to new data and it achieves good performance in the retrieval task. But there is no clear semantics associated with the identified audio segments. It is still impossible to retrieve content-related audios given a text query.

Another problem of the pure unsupervised method is the gap between semantic similarity and acoustic similarity. The acoustically similar sounds might be semantically different while several acoustically distinguishable sounds might be grouped into one semantic concept. Therefore, a well-defined semantic concept should leverage both the acoustic characteristics and human preference.

Following this intuition, we propose a novel framework to discover semantic concepts semi-automatically, which leverages the advantage of data-driven method while adding the constraints of human knowledge. We use manual annotations of broad semantic concepts as the seed and adopt unsupervised approach to expand the semantic concept vocabulary. The rest of the paper is organized as follows. In section 2, we describe the proposed framework and discuss how to select the vocabulary from the discovered results. In section 3, we describe the experiment setup and evaluate the result. We discuss the result in section 4 and conclude in section 5.

2. FRAMEWORK

2.1. Overview

Our proposed method uses an annotation of broad semantic concepts by human as the seed and explores unsupervised method to discover the hidden hierarchical structure of each broad concept. Each audio segment is mapped to a distribution of the acoustic descriptors learned over the data. Hierarchical clustering is applied based on the dissimilarity over the distribution. Each cluster is a candidate sub-category under the broad concept. We describe how to learn the acoustic descriptors using topic models in part 2.2 and how to apply the clustering algorithm to discover sub-categories in part 2.3. Concept vocabulary selection is discussed in part 2.4.

2.2. Learning Acoustic Descriptors

We assume that the sound recording characterizes of the distribution of multiple acoustic descriptors. Therefore, the descriptors are expected to have two properties: (1) they are representative patterns that can cover the data; (2) they are able to capture the characteristics of the sound recording, which help to identify the content but disregard the noise.

The MFCC (mel frequency cepstral coefficients) feature captures the short-term spectral information considering human auditory characteristics and is widely used in speech recognition and audio processing. Here we investigate to learn acoustic descriptors from MFCCs.

Topic modeling methods have been successfully used in text information retrieval. One of the methods, latent dirichlet allocation (LDA), effectively captures the hidden semantic structure in documents. We extend it to learn the acoustic descriptors. We use vector quantization to map the MFCCs into audio word in the codebook learned by k-means algorithm. Then, each audio recording can be treated as a document of words over the codebook. We apply the standard LDA method to learn the topics, which perform as the acoustic descriptors.

We give a brief overview of the latent dirichlet allocation below. Let K be a specified number of latent topics, V be the size of the vocabulary, z_i denotes a latent topic and w_i denotes a word. The generative process is as follows [7].

1.For $k = 1, ...K$, $\phi^k \sim Dir(\beta)$;
2.For each document d in corpus:
i. choose $\theta^d \sim Dir(\alpha)$;
ii.For each word $w_i \in d$:
-Choose a topic $z_i \sim multinominal(\theta)$
-Choose a word $w_i \sim multinominal(\phi^{z_i})$

where ϕ^k is a discrete distribution over a fixed vocabulary that represents the k^{th} topic distribution, θ^d is a document-specific distribution over the available topics, and α, β are hyper parameters for the symmetric Dirichlet distributions that the discrete distributions are drawn from. The joint probability is intractable in general. We use Gibbs sampling to estimate the parameter in the model.

Given a trained model, we can perform the inference on unseen document d_{new}:

$$P(w|d_{new}) = \sum_{k=1}^{K} P(w|z_k) \frac{\gamma_k}{\sum_{j=1}^{K} \gamma_j}$$

where γ_k is the Dirichlet parameter obtained during the inference on d_{new}. We use it as the posterior of topic proportions over a document.

2.3. Hierarchical Sub-categories Discovery

Using method in part 2.2, each audio segment can be mapped to a distribution over the acoustic descriptors. The dissimilarity of audio segments can be measured by the distance of the distributions.

Semantic Concepts should be defined by different levels of similarity based on the application. Therefore, we hope to discover a hierarchy of the concepts so that we can flexibly choose concepts distinguished by certain dissimilarity criterion which is proper for a certain application. Another benefit is that the hierarchical tree would help human annotator understand the data better. Here we adopt agglomerative hierarchical clustering algorithm among all the segments within certain broad semantic concept in the seed annotations. Fig.1 gives examples of the hierarchical cluster trees we learned from the broad semantic concepts using this algorithm.

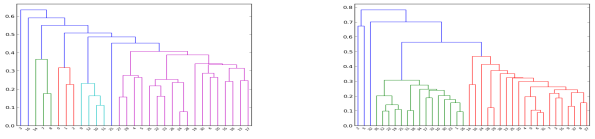

Fig. 1. 'singing'(left) and 'whistle'(right)

2.4. Semantic Concept Vocabulary Selection

We can get a set of sub-categories of semantic concepts by cutting the hierarchical cluster tree at certain level and assigning all the objects below each cut to one sub-category. So the question is where to cut and how to evaluate whether the generated clusters are reasonable or meaningful for the application.

We can judge the quality of the clusters by listening to the audio segments under each sub-category. For selecting meaningful clusters, we can evaluate indirectly through the performance of the application using a set of clusters. Here we discuss how to evaluate the enlarged semantic concept vocabulary for multimedia event detection (MED)[8]. For each video clip, we use the occurrence of semantic concepts as feature and the effectiveness of features is evaluated by the event classification performance. The semantic concept vocabulary only consists of selected sub-categories for the original broad categories. If there is no sub-category, the original broad category is kept in the vocabulary.

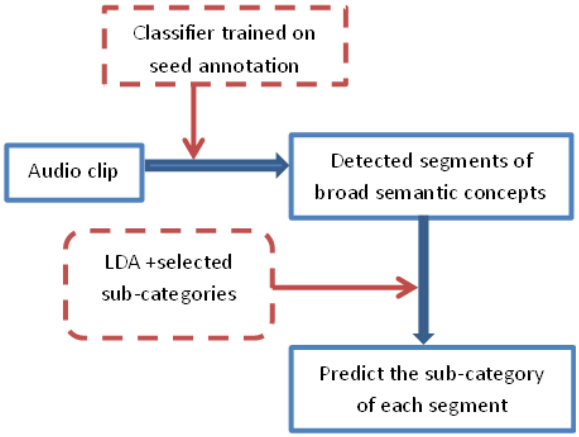

Fig. 2. Diagram for predicting semantic concepts.

To avoid further annotation for sub-category concepts, we design a scheme (Fig.2) to predict the probability for all semantic concepts in the video clip. First, we detect segments of the broad semantic concepts using classifiers trained on the seed annotations. Then, for each detected segment, we infer the distribution of acoustic descriptors using LDA models and predict its corresponding sub-categories (if exists). The probability of a sub-category is calculated as:

$$P(c_i|f) = \begin{cases} P(B|f) & \text{if} c_i = \arg\min_{i=1,,M} \|f - f_{c_i}\|^2 \\ 0 & \text{else.} \end{cases}$$

where c_i denotes the i^{th} the sub-category of broad concept B. f is the feature vector generated from the distribution of acoustic descriptors. f_{c_i} is the centroid of a cluster, which is calculated by taking the average of all samples within the cluster.

The occurrence probability of each semantic concept is estimated by its lasting length weighted by the prediction confidence:

$$d_{c_i} = \frac{P(c_i) * length(c_i)/length(clip)}{\sum_{j=1}^{M} P(c_j) * length(c_j)/length(clip)}$$

3. EXPERIMENT

3.1. The dataset and seed annotations

The experiments are conducted on the development data from the NIST 2011 Multimedia Event Detection task [8]. The dataset includes 3104 video clips for training, and 6642 video clips for testing.

We asked a human expert to manually label the audio semantic concepts for around 380 videos in the training set [9]. The name of the semantic concept refers to the event kit description and is assigned based on human understanding. After filtering the concepts that appear too rarely, we use 40 semantic concepts as broad semantic concepts and use the labeling as our seed annotations. The semantic concepts are: crowd, laugh, mumble, speech, speech_ne, cheer, music, music_sing, whistle, squeak, animal, anim_bird, anim_cat, anim_dog, scream, child, singing, tone, human_noise, rustle, scratch, micro_blow, white_noise, washboard, applause, wind, engine_quiet, engine_light, power_tool, engine_heavy, radio, water, knock, thud, clap, click, bang, beep, clatter, hammer.

3.2. Experiment setup

The parameters are experimentally set. In learning acoustic descriptors, MFCC feature is extracted using 32ms sliding window with 10ms shift. The codebook of MFCCs has 4096 word. The number of topics in LDA is set to be 100 and the model is learned over the whole dataset. For hierarchical clustering, we use the Euclidian distance as metric and use complete linkage in clustering.

In selecting sub-categories, we need to choose a threshold to cut the hierarchical tree generated in the clustering algorithm. Here the cutting threshold is a proportion of the maximum distance among all pairs of samples. We also experimented with other cutting criteria while the flat clusters are basically the same.

The rarely occured semantic concept will not contribute to distinguish the content of videos. Therefore, we heuristically filter out the clusters with low proportion. The rest of the clusters are selected as sub-categories. The original semantic concepts will be replaced by these sub-categories in the expanded vocabulary.

3.3. Result and Analysis

From the observation of the hierarchy built by the algorithm for each broad class, we find out that the number of clusters and the distribution of these candidate sub-categories vary from class to class. There are also some outliers, which fall into none of the natural groupings.

We further listen to the samples in the discovered sub-categories and most of the results sound reasonable. For example, the music in background is separated from the normal one, the heavy wind and breeze fall into different clusters, and etc. Some sound events would co-occur with other sound in some acoustic scene. We find out that some separated clusters are due to different overlapping sound. For example, the segments of 'crowd' with overlapping speech are separated with those without speech.

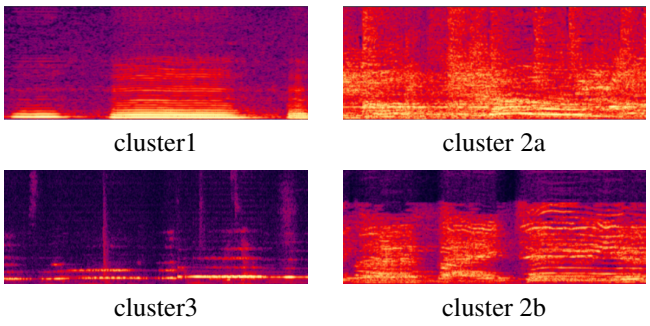

Fig. 3. Spectrum of randomly selected audio recording within clusters.

To further examine the result, we also observe the frequency spectrum. Here we discuss the 'singing' example. Excluding the outliers, we discover 3 sub-categories by certain threshold in flattening clusters. Cluster1 sounds like opera singing; cluster3 includes humming; another cluster includes two types of singing, multiple people singing together (cluster2a) and singing with music accompaniment (cluster2b). As is shown in Fig.3, the pattern in the spectrum reflects that our method is able to distinguish acoustic characteristics. The two types in cluster2 look more similar in spectrum compared to the other clusters. It would require cutting the tree at a lower level to distinguish the subtle difference.

3.4. Vocabulary Selection for MED

We build multiple binary classifiers based on random forest algorithm [10] using the seed annotations. The acoustic features are extracted from 2s sliding window with 100ms shift. So the prediction granularity is 100ms. The event classification pipeline [11] uses one-against-all rigid kernel regression classifier using χ^2 kernel for each event over the feature. The performance of the event classification task is evaluated by the normalized MAP (the higher number indicates better performance) since each event has different numbers of positive samples and negative samples.

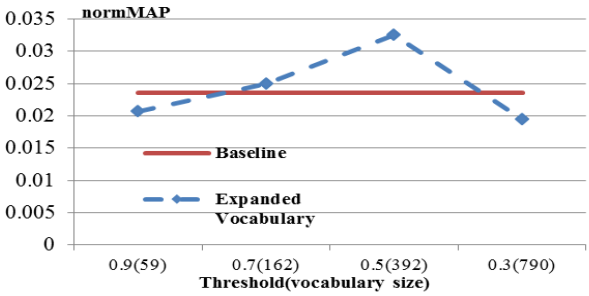

Fig. 4. The normalized MAP for event detection task using different vocabulary of semantic conepts.

As is shown in Fig.4, we experiment with cutting threshold of 0.5, 0.7 and 0.9 of the maximum distance. The results show that we can get significant improvement using a proper expanded vocabulary compared with the baseline using original broad concepts. we can also observe that the performance degrades after splitting broad semantic concepts into too many sub-categories. This is understandable because the introduced noise overwhelm the informativeness brought by the sub-categories.

4. CONCLUSION

In this paper, we present a novel framework to discover audio semantic concepts semi-automatically. The framework leverages LDA models to discover hidden patterns in acoustic feature space and use seed annotations to distill human understanding. The proposed method allows us to build an effective vocabulary of semantic concepts on data quickly with an initial annotation of broad semantic concepts. The hierarchy would also help human understand the structure within each semantic concept.

5. ACKNOWLEDGMENT

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.

6. REFERENCES

[1] Shih-Fu Chang, Dan Ellis, Wei Jiang, Keansub Lee, Akira Yanagawa, Alexander C. Loui, and Jiebo Luo, "Large-scale multimodal semantic concept detection for consumer video," in *Multimedia Information Retrieval*, 2007, pp. 255–264.

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

[3] 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.

[4] Sourish Chaudhuri, Mark Harvilla, and Bhiksha Raj, "Unsupervised learning of acoustic unit descriptors for audio content representation and classification," in *INTERSPEECH*, 2011, pp. 2265–2268.

[5] Sourish Chaudhuri and Bhiksha Raj, "Unsupervised structure discovery for semantic analysis of audio," in *NIPS*, 2012, pp. 1187–1195.

[6] Sourish Chaudhuri and Bhiksha Raj, "Unsupervised hierarchical structure induction for deeper semantic analysis of audio," in *ICASSP*, 2013, pp. 833–837.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

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

[9] 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.

[10] Leo Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[11] 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.