Semantic Concept Annotation for User Generated Videos Using Soundtracks

Qin Jin[1,2], Junwei Liang[1], Xixi He[1], Gang Yang[1], Jieping Xu[1], Xirong Li[1,2]
[1]Multimedia Computing Lab, School of Information, Renmin University of China, 100872, China
[2]Key Lab of Data Engineering and Knowledge Engineering, Renmin University of China, 100872, China
{qjin,chunwei,xixilanmi,yanggang,xjieping,xirong}@ruc.edu.cn

ABSTRACT

With the increasing use of audio sensors in user generated content (UGC) collections, semantic concept annotation from video soundtracks has become an important research problem. In this paper, we investigate reducing the semantic gap of the traditional data-driven bag-of-audio-words based audio annotation approach by utilizing the large-amount of wild audio data and their rich user tags, from which we propose a new feature representation based on semantic class model distance. We conduct experiments on the data collection from HUAWEI Accurate and Fast Mobile Video Annotation Grand Challenge 2014. We also fuse the audio-only annotation system with a visual-only system. The experimental results show that our audio-only concept annotation system can detect semantic concepts significantly better than does random guessing. The new feature representation achieves comparable annotation performance with the bag-of-audio-words feature. In addition, it can provide more semantic interpretation in the output. The experimental results also prove that the audio-only system can provide significant complementary information to the visual-only concept annotation system for performance boost and for better interpretation of semantic concepts both visually and acoustically.

Categories and Subject Descriptors

H.5.5 [**Information Interfaces and Presentation**]: Sound and Music Computing –*Signal analysis, synthesis, and processing.*

General Terms

Algorithms; Experimentation; Measurement.

Keywords

Semantic Concept Annotation; Audio Analysis; Consumer Videos.

1. INTRODUCTION

With the rapid introduction of mobile devices in our everyday life, there appears explosively growing amount of user generated content (UGC) videos on social networking services, such as YouTube, Dailymotion, Youku and Tudou in China. Such rich data resources have attracted tremendous research interest in developing automatic technologies for organizing and indexing multimedia content [1]. Automatic semantic concept annotation for UGC videos has been one of the hot research topics. Although the visual information in a video is clearly very important for semantic concept annotation, we believe that the soundtrack may offer a useful and complementary source of information, especially when the visual part fails (occlusion or absence). In this paper we address this challenge by exploiting acoustic information – the soundtrack of a video – to see what useful descriptors can be reliably extracted from this modality.

We summarize the previous related research work in soundtrack analysis and audio concept or event detection from the following three areas of focus: Number of sound classes; Quality of the audio data; Granularity of the audio processing. Much early works focused on a small number of sound classes such as speech, music, silence, noise, or applause. Various acoustic features (such as Mel-Frequency Cepstral Coefficients (MFCCs), zero-crossing rate, short-time energy, pitch, and spectral peak tracks [3-5] etc.) and their derivative features such as posterior probability features [4] plus various classifiers including Gaussian Mixture Models [3], Hidden Markov Models [6], Support Vector Machines [7] and so on have been explored. Early works focused on relatively carefully produced audios, such as broadcast audio or movie soundtracks [3-7]. There is also a lot of research on classifying less constrained environmental sounds [8-11], such as in surveillance applications or context-awareness in mobile devices. In recent years, the growing popularity of video sharing services, such as YouTube, Dailymotion, and Youku, enables the use of a quickly increase amount of user-generated videos. The soundtracks in UGC videos are typically of poor quality and often contain only sparse instances of class-relevant sounds. Thus, working on this data becomes a more challenging task. The soundtrack analysis of consumer videos [12-15] is less researched, particularly when considering the analysis at a finer video resolution, at the frame level. Even though the semantic indexing (SIN) task in TRECVID [16] has been a subtask of localizing concepts at the frame-level since 2013, few studies have utilized a purely auditory method to help to achieve the goal. The bag-of-audio-word [13,15] along with other approaches like pLSA-of-GMM [12] suffer from the drawback that the results are unable to provide semantic interpretation. Rather than specifying and labeling the sub-concepts, we add the semantic interpretation by using "data in the wild" according to particular rules.

The study in this paper is conducted on the data from the grand challenge in the International Conference on Multimedia & Expo (ICME) 2014: HUAWEI Accurate and Fast Mobile Video Annotation Challenge [16]. The goal of this task is to analyze UGC videos and annotate their contents automatically at the frame level: any of the target concepts appear in this frame or not. Comparing to the semantic concept annotation task at the entire-video level or at the supra-segmental level in previous research, this task requires annotation with finer resolution, and it certainly is a more challenging task. In this paper, we focus on six semantic

ICMR'15, June 23–26, 2015, Shanghai, China.

http://dx.doi.org/10.1145/2671188.2749388

concept classes, which cover objects (e.g. "car", "dog", and "kids"), scenes (e.g. "beach"), and events ("football" and "party"). This paper provides a semantically enhanced approach to extract low-level audio features using the traditional bag-of-audio-word method while at the same time maintaining the readability of the results, which we learn from freely available audio data from the Internet, and their descriptions.

2. AUDIO-ONLY ANNOTATION SYSTEM

Our semantic concept annotation system using soundtrack/audio information [19] contains two major components: audio features extraction and concept annotation model training

2.1 Feature Extraction

We first extract the mono soundtrack from a video and re-sampled it to 8kHz. In order to detect concepts at the frame level, we chunk the audio stream into a sequence of 3-sec snippets with 1-sec shift. We then extract audio features and apply concept detection to the snippets. Our frame-based low-level feature is the Mel-frequency Cepstral Coefficients (MFCCs) over short-time window of 25ms with a 10ms shift as our fundamental frame-level feature. Each 25ms frame of a video soundtrack is then represented as a 39-dim MFCC feature vector (13-dim MFCC + delta + delta delta). Each snippet is then represented by a set of MFCC feature vectors. Based on the love-level MFCC features, we derive following other higher-level feature representations.

2.1.1 Bag-of-Audio-Words (BoAW) Feature

The audio words (or acoustic codebook) are generated in a data-driven fashion. To be specific, the codebook is learned by applying the Kmeans clustering algorithm to the 39-dimensional MFCC features from the whole training dataset. Each of the learned K means is considered as one audio word. Each soundtrack snippet is then represented as a distribution over these K codewords by using a soft-assignment (by adding the closest 5 codewords' counts) of MFCC features to these codewords. This is the feature we use in our baseline system.

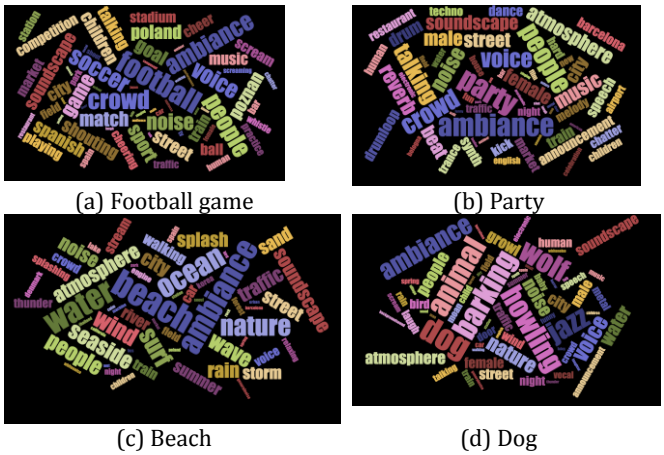
2.1.2 Semantics-Derived (SD) Feature

In order to maintain the good performance of the representation while improving the semantic interpretability of the results, we propose the semantics-derived (SD) Feature. This feature generation consists of four steps: 1) downloading wild data (in this experiment we use data from *freesound*); 2) selecting sound classes via tag selection from the wild data; 3) building SVM classifiers for sound classes; 4) generating feature representations for a snippet by scoring it against the sound class models.

2.1.2.1 Wild Data Downloading

We find that freesound.org [20] is a good repository covering a wide range of user-collected recordings. We first search audios for the six target concepts (car, dog, kids, beach, football, and party) by using the concept name as a textual query in the freesound search engine. This returns audio segments, which contain similar tags or description to the query text. Next we use each of the returned audio segments as a query to search for acoustically similar audio segments in the freesound similarity search engine, which returns audio segments that have similar statistics over audio features, such as the zero-crossing rate. We keep all the text similarity results and the top 20 acoustic similarity results for each search. Finally, we filter out the audios with bad quality, or those with abnormal coding. Eventually, we collect 35681 audio files. The resulting wild audio dataset has a total duration of about 763 hours, and the median of a file's duration is 20 seconds. We split the freesound dataset into a training set of 28544 files and a testing set using the remainder of the files.

2.1.2.2 Sound Class/Tag Selection



(a) Football game (b) Party

(c) Beach (d) Dog

Figure 1. Tag Selection Examples for some Concepts

Each wild audio file contains tags created by the uploader. In order to select semantic classes with sufficient samples to train reliable classifiers, we select the top n most frequent tags that appeared in each concept's search results. We choose 150 most frequent tags from each concept's text similarity search results and 300 most frequent tags from their acoustic similarity search results. Some concepts share tags, for example, "football game" and "party" share tags like 'noise' and 'crowd', which indicates their acoustic similarity. At the same time, "football game" also contains tags that "dog" doesn't, such as 'shout' or 'cheer', which indicates acoustic differences. In this way, we obtain 579 tags. Examples of tags selected for the concepts "football game", "party", "beach", "dog" are shown as a word cloud according to their occurrence frequencies in Figure 1.

2.1.2.3 Sound Class Model Training

We train SVM classifiers for each of the 579 semantic classes. We chunk each audio file into 3-sec snippets with 1-sec overlap. We extract BoAW features for each snippet using a codebook clustered on the whole wild data training set by the Kmeans algorithm. For each semantic class, we use audio files that contain a particular tag as positive samples, and keep the rest as negative samples to train a classifier for each of the 579 semantic classes.

2.1.2.4 Sound Class Scoring

We use the sound-class classifiers trained from the freesound data to derive a new feature representation. A soundtrack snippet is compared to the sound-class classifiers, and the SVM scores from the 579 classifiers are concatenated to form a 579-dimensional new feature vector for each snippet. Considering it to be straight-forward to generate the BoAW features using freesound data as well, we generate the 4096 audio words using the freesound data via Kmeans clustering, and create the BoAW features accordingly. Since here the audio words are learned from freesound data, we name this bag-of-word feature as FS-BoAW. We also combine the 579-dimensional feature with the 4096-dimensional FS-BoAW feature to form a new feature representation (concatenated after scaling the 579-dimensional feature into the same range of the FS-BoAW feature). We call these features 'stand-alone SD features' and 'combined SD features', respectively.

2.2 Concept Annotation Models

After we extract the features, we train a binary SVM classifier for each of the six concepts. Because the training data is overwhelmed by negative examples, we train classifiers using the Negative Bootstrap algorithm [21]. The algorithm takes a fixed number (*N*) of positive examples and iteratively selects those negative examples, which are misclassified the most by the current classifiers. The algorithm randomly samples *10xN* number of negative examples from the remaining negative examples as candidates at each iteration. An ensemble of classifiers trained in the previous iterations is used to classify each of the negative candidate examples. The top *N* most misclassified candidates are selected and used together with the *N* positive examples to train a new classifier. The algorithm takes several bags of positive examples and performs the training independently on each of the positive bags, and the classifiers are compressed at the end. In order to improve the efficiency of the training process, we use the Fast intersection kernel SVM (FikSVM) as reported in [22].

3. EXPERIMENTS

3.1 Data Description

The HUAWEI grand challenge dataset contains of 2,666 UGC videos. Originally, the concepts to be annotated are 10 semantic concept classes, covering objects (e.g. "car", "dog", "flower", "food" and "kids"), scenes (e.g. "beach", "city view" and "Chinese antique building") and events ("football" and "party"). We discovered that the soundtracks of videos samples for "flower", "food", "city view" and "Chinese antique building" contain mostly post-edited pure music. We therefore exclude these four concepts in our study on this dataset, only focusing the remaining six concepts. The data set contains a total duration of about 70 hours of videos excluding those videos with pure music. We divide the dataset into a training set (1014 videos), a development set for tuning model parameters and fusions weights (362 videos), and a test set (656 videos). The amount of negative examples is 10 times larger than the amount of positive examples. The entire dataset has manual labels at the frame level.

As we looked more closely into the provided ground truth files, we came to believe that the provided manual labels are based on visual content. For example, the videos with image of dogs in the foreground and with children talking from behind the camera are only labeled with "dog" but not with "kids". The videos where a dog's barking can be heard in the soundtrack but no dog to be seen in the video do not have a "dog" label. In order to further investigate the semantic concept annotation from both an audio and a visual point of view, we need consistent ground truth with both audio and visual evidences. Therefore, we hand-labeled the entire dataset by only listening to the soundtracks without looking at the videos to generate the new audio-driven ground truth for the six target concepts.

We build an audio/soundtrack annotation system trained using the audio-driven ground truth. We also build a visual annotation system trained using the visual-driven ground truth. In following experiments, we use the audio-driven ground truth to evaluate the soundtrack annotation system and the intersection ground truth for comparing the audio-only, visual-only and fusion systems.

3.2 Baseline Results

We use the average precision to evaluate the concept annotation performance for each concept class. The detailed baseline results are shown in Table 1. The baseline system with HW-BoAW features of 4096 dimensions (which means the 4096 audio words were learnt from the target HUAWEI dataset) yields a mean AP (mAP) of over 0.497 on all six concepts. We randomly generate a ranked result for several times to get the mean random guess results of mAP 0.106, which is close to the percentage of positive examples in the testing set. From these results, we can see that the concept annotation based on the soundtrack only achieves significantly better performance than random guessing does. We can also see that some concept classes, which are acoustically easy to distinguish such as "football game", "kids", clearly yield much better performance than others (with APs of 0.793 and 0.575, respectively). We also generate the 4096 audio words via clustering the freesound data for the BoAW features (FS-BoAW). The FS-BoAW achieves a comparable performance to the HW-BoAW. This implies some potential benefit for the FS-BoAW features: when the labeled target training data is limited or unavailable, we can still generate a similar feature representation using audio words learned from wild data.

Table 1: *Baseline annotation performance comparison*

	beach	car	dog	fb-game	kids	party	**mAP**
HW-BoAW	0.161	0.489	0.476	0.793	0.575	0.490	**0.497**
FS-BoAW	0.162	0.480	0.461	0.797	0.571	0.476	**0.491**
Random guess	*0.055*	*0.165*	*0.018*	*0.153*	*0.106*	*0.139*	***0.106***

(c) Test Example for Kids

Figure 2. Semantic Interpretation from SD Features

3.3 Semantics-Derived Features Results

As described in section 2.1.2, we build 579 semantic concept classifiers from the freesound data and use them to generate the SD feature representation. Concept classifiers for the six target concepts are then trained using the SD features, and annotation of the testing data follows. The stand-alone SD features yield a mAP of 0.443, and the combined SD features yield a mAP of 0.499 over the six concepts. Although the combined SD features only slightly outperform the baseline system, it provides a lot of additional semantic information in the output, as shown in Figure 6. The word clouds are generated according to the value of each dimension in the SD features. In Figure 2, we see for the test video example of "kids", the top ranked SD feature dimensions correspond to "street", "city", "subway". We can infer that the video is recorded near streets in the city rather than on the country side, which we cannot determine by just using the BoAW features or even the visual cues.

4. FUSION OF AUDIO AND VISUAL

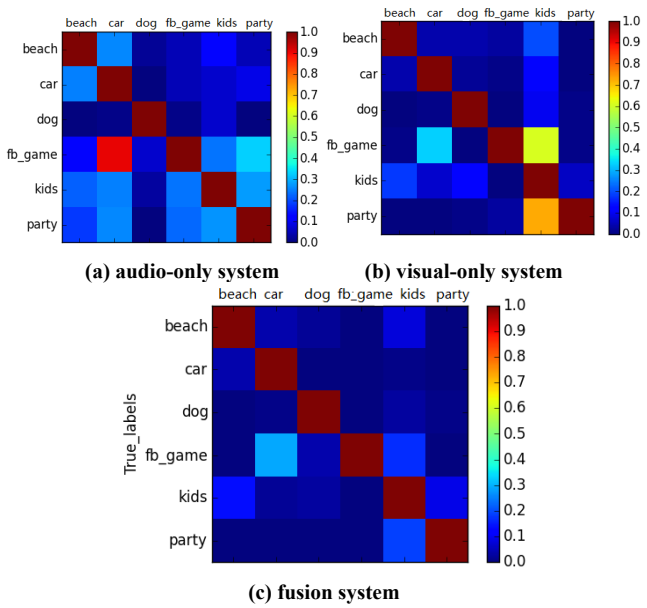
In the following experiments, we evaluate the fusion performance on the contents that are both visually and acoustically identifiable. Please note that the intersection of the audio-driven and visual-driven ground truth is used to evaluate the fusion system. The fusion weights are learned following the approach in [23].

The experimental results in Table 2 show that the audio and visual streams contain complementary information for interpreting a

semantic concept. A relative improvement of 30% is achieved when combining both modalities. Figure 3 shows the confusion matrices for the three systems. Each matrix in Figure 10 is normalized to the same scale so that they can be compared to one another. In Figure 3(a) and 3(b) we see that although the audio-only system makes more mistakes overall, it makes fewer mistakes distinguishing "kids" from "football game" and "kids" from "party". After fusion, the visual system's weakness has been reduced, as we show in Figure 3(c).

Table 2: *Annotation performance comparison among the audio-only, visual-only and fusion systems*

Concept	Audio-only	Visual-only	Fusion	Audio Weights
Beach	0.133	0.447	0.542	0.42
Car	0.212	0.374	0.448	0.38
Dog	0.476	0.133	0.497	0.95
Fb-game	0.798	0.945	0.969	0.41
Kids	0.529	0.265	0.578	0.82
Party	0.401	0.808	0.839	0.30
mAP	**0.425**	**0.495**	**0.645**	**-**



(a) audio-only system **(b) visual-only system**

(c) fusion system

Figure 3. Confusion matrices for different systems.

5. CONCLUSIONS

In this study, we illustrate the viability of classifying concepts of consumer videos based only on soundtrack data. The audio only system achieves a significantly better performance than just random guessing. We show how to make use of freely available wild data, and how to bridge the semantic gap using the bag-of-audio-word method. We explore explicitly adding semantic detail to the feature representation. We propose a semantics-derived feature representation by using a large-scale dataset which we downloaded from freesound, and which we combine with a bag-of-word feature representation. The combined semantics-derived feature not only provides semantic information in the output but also outperforms the baseline features. Finally, we fuse the audio-only system with the visual-only system, and such achieves a relative improvement of 30%. Our simple strategy to enhance the semantic interpretation using wild data shows promising results. In future work, we plan to explore latent and embedding approaches for the semantic enhancement.

6. ACKNOELDGEMENT

This work is supported by the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (No. 14XNLQ01), the Beijing Natural Science Foundation (No. 4142029), NSFC (No. 61303184), SRFDP (No. 20130004120006).

7. REFERENCES

[1] C. Snoek, and M. Worring: Concept-based Video Retrieval. Foundations and Trends in Information Retrieval, 2009.
[2] P. Over, et al.: TRECVID 2013 -- An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. TRECVID 2013, USA.
[3] J. Saunders: Real-time discrimination of broadcast speech/music. ICASSP, 1996.
[4] G. Williams and D. P. W. Ellis: Speech/music discrimination based on posterior probability features. Eurospeech, Budapest, 1999.
[5] T. Zhang and C.-C. J. Kuo: Audio content analysis for online audiovisual data segmentation and classification. IEEE Tr. Speech and Audio Proc., vol. 9, no. 4, pp. 441–457, 2001.
[6] J. Ajmera, I. McCowan, H. Bourlard: Speech/music segmentation using entropy and dynamism features in a hmm classification framework. Speech Communication, (40), 351–363, 2003.
[7] K. Lee and D. P. W. Ellis: Detecting music in ambient audio by long window autocorrelation. ICASSP, 2008.
[8] D. P. W. Ellis and K. Lee: Minimal-impact audio-based personal archives. ACM Workshop on Continuous Archival and Retrieval of Personal Experiences, New York, NY, October 2004.
[9] S. Chu, S. Narayanan, and C.-C. J. Kuo: Content analysis for acoustic environment classification in mobile robots. AAAI Fall Symposium, Aurally Informed Performance: Integrating Machine Listening and Auditory Presentation in Robotic Systems, 2006.
[10] A. Eronen, et al.: Audio-based context recognition. IEEE TASLP, (14), no. 1, pp. 321–329, Jan. 2006.
[11] E. Wold, et al.: Content-based Classification, Search, and Retrieval of Audio," IEEE Multimedia, 3(3), 1996.
[12] K. Lee and D.P.W. Ellis: Audio-Based Semantic Concept Classification for Consumer Video, IEEE TASLP, 18(6), 2010.
[13] Q. Jin, P. Schulam, S. Rawat, S. Burger, D. Ding, F. Metze, "Categorizing Consumer Videos Using Audio," Interspeech, 2012.
[14] L. Ma, B. Milner, and D. Smith: Acoustic Environment Classification. ACM Trans. on Speech and Language Processing, 3(2), 2006.
[15] L. Brown, et al.: IBM Research and Columbia University TRECVID-2013. In: TRECVID Workshop, 2013.
[16] ICME 2014 Huawei Accurate and Fast Mobile Video Annotation Challenge http://www.icme2014.org/huawei-accurate-and-fast-mobile-video-annotation-challenge.
[17] X.B. Xue, Z.H. Zhou: Distributional Features for Text Categorization. IEEE Transactions on Knowledge and Data Engineering, 21(3), 2008.
[18] J. Philbin, et al.: Object retrieval with large vocabularies and fast spatial matching. CVPR 2007.
[19] J. Liang, et al.: Semantic Concept Annotation of Consumer Videos at Frame-level Using Audio. Pacific-Rim Conference on Multimedia (PCM), 2014.
[20] Freesound data repository: http://www.freesound.org
[21] X. Li, C. Snoek, M. Worring, D. Koelma, A. Smeulders: Bootstrapping Visual Categorization With Relevant Negatives. IEEE Transactions on Multimedia, 15(4), 2013.
[22] S. Maji, A. Berg, J. Malik: Classification using international kernel support vector machines is efficient. In: CVPR 2008.
[23] X. Li, C. Snoek, M. Worring, A. Smeulders, "Fusing concept detection and geo context for visual search", ICMR 2012.