Semantic Concept Annotation of Consumer Videos at Frame-Level Using Audio

Junwei Liang, Qin Jin*, Xixi He, Gang Yang, Jieping Xu, and Xirong Li

Multimedia Computing Lab, School of Information, Renmin University of China
{leongchunwai,qjin,xxlanmi,yanggang,xjieping,xirong}@ruc.edu.cn

Abstract. With the increasing use of audio sensors in user generated content (UGC) collection, semantic concept annotation using audio streams has become an important research problem. Huawei initiates a grand challenge in the International Conference on Multimedia & Expo (ICME) 2014: Huawei Accurate and Fast Mobile Video Annotation Challenge. In this paper, we present our semantic concept annotation system using audio stream only for the Huawei challenge. The system extracts audio stream from the video data and low-level acoustic features from the audio stream. Bag-of-feature representation is generated based on the low-level features and is used as input feature to train the support vector machine (SVM) concept classifier. The experimental results show that our audio-only concept annotation system can detect semantic concepts significantly better than random guess. It can also provide important complementary information to the visual-based concept annotation system for performance boost.

Keywords: Semantic Concept Annotation, Video Content Analysis, Soundtrack Analysis.

1 Introduction

With the explosion of user generated content (UGC) on current social network sites, there has attracted tremendous research interest in developing automatic technologies for organizing and indexing multimedia content [1]. Semantic concept analysis, which consists of annotating and searching a multimedia collection for user-defined concepts, is one of the fundamental analysis tasks in multimedia content understanding. The outcomes of such analysis processes are "high-level" concepts for describing, indexing and searching consumer media [2, 3].

Traditional semantic concept analysis techniques have focused largely on the visual domain. Owing to the fact that vision is the highest bandwidth sensor for humans, it makes sense that machines would also be able to extract significant semantic information from image and video data. However, audio also conveys significant information that can semantically interpret the video content. Audio is extremely useful in certain situations when other sensors such as visual sensor fail to provide reliable

* Corresponding author.

W.T. Ooi et al. (Eds.): PCM 2014, LNCS 8879, pp. 113–122, 2014.

information. For example, when the object is occluded or is in bad illumination, the audio sensors are the key sensors in detecting the presence of objects assuming the objects make sound. Therefore, in the context of multimedia semantic concept analysis applications, audio stream (the soundtrack of a video) can provide important complementary information to visual stream [5, 6]. In this paper, we focus on video concept annotation based only on audio stream.

Most of the state-of-the-art semantic concept frameworks were conducted toward the videos with loose structures such as sports videos [7], surveillance videos [6], or medical videos etc. [8]. In recent years, the consumer generated videos are getting more research attention such as in TRECVID evaluations [3]. Consumer generated videos are unstructured compared to professional contents like films. It brings a lot of technical challenges to analyze them. Huawei organized a grand challenge in the International Conference on Multimedia & Expo (ICME) 2014: Huawei Accurate and Fast Mobile Video Annotation Challenge [9]. The goal of this task is to analyze UGC videos and annotate their contents automatically. The labels to be annotated are 10 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 game" and "party"). The semantic concept annotation within the Huawei challenge is required to be at the frame-level. That means for each frame, we need to make a binary decision about the presence of a specific concept in the frame. Comparing to the semantic concept annotation task at the video level or supra segmental level in previous research, this task requires annotation with finer resolution and is a more challenging task.

The remainder of this paper is organized as follows. Section 2 presents the related work. Section 3 introduces the Huawei grand challenge dataset. Section 4 describes audio semantic concept annotation system. Experimental results are presented in Section 5. Some discussions follow in Section 6 and conclusions are made in Section 7.

2 Related Work

We summarize the prior work in soundtrack analysis from three different dimensions based on the type of data that researchers have focused on:

- First dimension: the quality of the audio data. Early work on audio event classification was largely done on sound databases [10] and clean broadcast or television program audio data [11]. Typical high quality database or broadcast data can be extremely clean, and "foreground" sounds are generally easy to distinguish from "background" sounds due to studio recording conditions. The growing popularity of video sharing services such as YouTube, Dailymotion, Youku in China and so on enables the vast increasing of user generated videos. Analyzing such consumer videos is more challenging.
- Second dimension: number of sound classes. Much early works focused on detecting or distinguishing between a small number of sound classes such as speech, music, silence, noise, or applause. This was solved using various traditional machine learning and signal processing approaches [11-13].

- Third dimension: the granularity of the audio processing. We can roughly categorize the soundtrack analysis work into two categories: sub-soundtrack classification or entire soundtrack classification. Distinguishing between a small numbers of sound classes can be considered as a sub-soundtrack classification problem. It produces annotations of input data according to a fixed number of classes for which one has trained models. Such sound classes can be aforementioned speech/music etc. There also have been efforts to classify short audio clips with respect to the environment in which they were recorded [14]. The multimedia event detection (MED) using soundtrack is the entire soundtrack classification problem [4]. Modeling the event based on sub-soundtrack classification results has been one type of approaches in such tasks [15, 16]. Though the semantic indexing (SIN) task in TRECVID [4] has a subtask of localizing concepts on frame-level since 2013, few work have used auditory method to help achieve the goal. Similar to the SIN subtask, the Huawei grand challenge can be categorized as a sub-soundtrack classification problem and sound classes are aforementioned 10 concept classes.

3 Data Description

Our experiments are conducted on the development data of the Huawei Accurate and Fast Mobile Video Annotation Challenge (MoVAC 2014). All the videos are collected from the Internet (e.g. Youku and Youtube) and converted to mpeg4 format. The selected 10 semantic concepts cover objects (e.g. “car”, “dog”, “flower”, “food” and “kids”), scenes (e.g. “beach”, “city view” and “Chinese antique building”) and events (“football game” and “party”). There are also some extra videos as background videos which contain none of the predefined 10 concepts.

Table 1. Number of positive and negative frames for each concept in the dataset

Concept	#pos	#neg	%pos
beach	664793	8391067	6.6%
car	1161116	8843824	11.6%
chinese_antique_building	772805	772805	7.7%
city_view	801583	9203357	8.1%
dog	524560	9480380	5.2%
flower	1082986	8921954	10.8%
food	378046	9626894	3.7%
football_game	161873	8391067	16.1%
kids	1525771	8479169	15.2%
party	780240	9224700	8.0%
MEAN	**1030577**	**8974362**	**10.3%**

The development dataset contains 2,666 videos. The video resolution ranges between 640x480 and 1280x720. The videos are normally taken by mobile devices. The recording frame per second (fps) varies among all the videos. Some videos have been post-edited, such as a video on flowers with only music audio. Some videos (16 out of 2666) have no soundtrack. We divide the dataset into a training set (which contains 1764 videos) and a test set (which contains 886 videos). The ground truth label files with manual annotations are in the format of three columns: <concept>\t<start frame index>\t<end frame index>, such as in the following example:

```
Car   1      568
Car   93     1165
Car   1386   1423
Kids  1      1423
```

The detailed information about the amount of positive and negative examples in each concept class is listed in Table 1. As we can see from the table, the amount of negative examples is overwhelmingly larger than the amount of positive examples.

4 System Description

Our semantic concept annotation system using audio only contains the following key components (Figure 1): audio data pre-processing, audio features extraction, concept annotation models and post-processing.

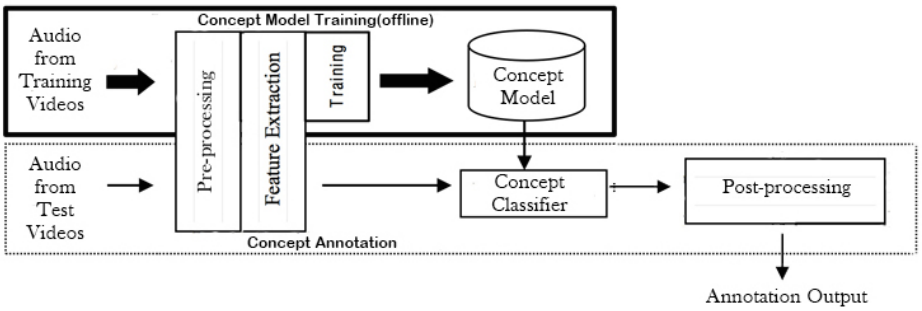

Fig. 1. Audio-only Concept Annotation System Components

Pre-processing. In order to detect concept on frame-level, we chunk the audio stream into small segments with overlap, extract audio features and apply concept detection on those segments. The length of the segment/chunk is a tunable parameter.

Audio Features. The codebook model is a common technique used in the document classification (bag-of-words) [18] and image classification (bag-of-visual words) [19]. The similar bag-of-audio-words model has also been applied in the sound track analysis work [5, 17]. In our system, we use bag-of-audio-words model to represent each audio segment. It is represented by assigning low level audio features to a discrete set of

codewords in the vocabulary (codebook) thus providing a histogram of codeword's counts. These codewords are learnt via unsupervised clustering. The discriminative power of such a codebook is governed by the size of the codebook and by the assignment of features to codeword's [14]. In this paper we apply this model to the low level MFCC features. The MFCC features are computed every 25ms with 10ms shift and are in 39 dimensions (13 MFCC + 13 delta + 13 ddelta). The vocabulary is learnt by applying kmeans clustering algorithm with K=4096 on the whole training dataset. Each audio segment is then represented as a distribution over these 4096 codewords' by using soft-assignment of MFCC features to these codewords' (as shown in Figure 2).

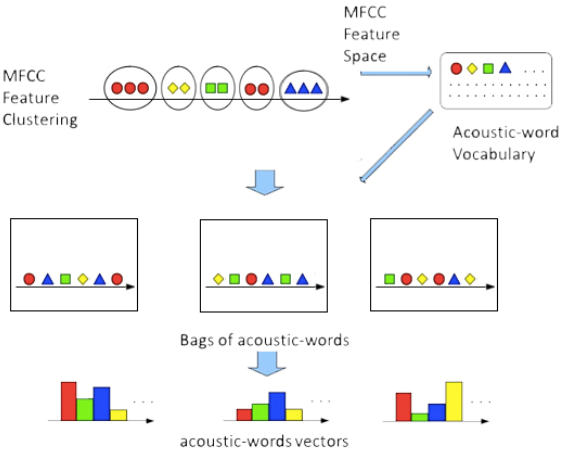

Fig. 2. Data-driven MFCC based Bag-of-Audio-Words model

Concept Annotation Models. After we calculate all the bag-of-audio-words features, we train two-class SVM classifiers for each of the 10 concepts. As shown in table 1 that the training data is overwhelmed by negative examples, we train classifiers by the Negative Bootstrap algorithm [20]. The algorithm takes a fixed number (*N*) of positive examples and iteratively selects negative examples which are most misclassified by current classifiers. The algorithm randomly samples *10*x*N* 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 negative candidate examples. The top *N* most misclassified candidates are selected and used together with the *N* positive examples to train a new classifier. In order to improve the efficiency of the training process, we use Fast intersection kernel SVMs (FikSVM) as reported in [21].

Post-processing. Intuitively, if a concept occurs within a video, it is usually not an instantaneous appearance. It normally lasts for certain duration. Therefore, we conduct boundary padding and cross-segment smoothing over the raw annotation results. We expand the beginning and ending of the detected segments. We also merge two detected segments if they belong to the same concept and the gap between them is below a certain threshold.

5 Experimental Results

We use the average precision to evaluate the concept annotation performance for each concept class:

$$AP = \frac{1}{R}\sum_{j=1}^{n} I_j \times \frac{R_j}{j} \quad (1)$$

where R is total number of relevant segments of that concept, n as the total amount of segments, I_j=1 when the j^{th} segment are relevant otherwise I_j=0 and R_j is the number of relevant segments in the first j segments.

We experiment with different length of chucking in the pre-processing step: 3sec-Chunk+1sec-Shift; 5sec-Chunk+2sec-Shift; 7sec-Chunk+3sec-Shift. The results are very close. Longer chunk only improves less than 1% on mean AP. Longer chunk takes more time to process, we therefore choose the 3sec-Chunk+1sec-Shift setup as our best setup. We choose *N=3000* in the Negative Bootstrap Training. The second column in Table 2 shows the annotation performance of each concept class using audio only with the best setup. We achieve a mean average precision of 30% on all 10 concepts. Some concept classes, which are acoustically easy to distinguish such as "football game", "dog", "kids", clearly achieve much better performance than others, with AP of 72.8%, 47.9%, and 40.5% respectively. From the results, we can see that the concept annotation based on audio stream only achieves significantly better performance than random guess. Since significant semantic information is conveyed in the visual stream, we also develop the concept annotation system using visual stream. Intuitively, the audio and visual streams contain complementary information for interpreting a semantic concept. We then explore to combine these two systems. The linear fusion weights of the two systems are tuned on a held out dataset. As shown in Table 2, although the visual concept annotation system achieves much better performance than the audio system, combining them achieves certain improvement over all of the 10 concept classes.

As we inspect the data closely, we find out that since the manual annotation of videos is produced mainly according to visual evidence, there are a certain amount of audios that are unrelated to the concepts they are labeled. Therefore, we try to clean these acoustically false "positive" examples by listening to them and deciding whether we human can identify the concept(s) based on audio only. We only focus on three concepts, "kids", "football game", and "dog", which intuitively can be easily distinguished using acoustic cues. We conduct the following procedure to clean the data. The videos satisfy either one of the following criteria will be excluded: 1) the videos with music-only audio that may has nothing to do with the visual content and the music only relates to the editor's taste which is random; 2) the videos with loud background noise that we cannot hear any of the labeled concept(s), such as a dog swimming in a coming wave with strong wind blowing at the beach; 3) the videos with no target concept's sound at all, for examples, a sleeping dog or a kid quietly sitting on the chair. In the end, we exclude 10-30% amount of annotations for these three concepts, both in the training set and the test set. We then conduct the annotation experiment on this cleaned data set with the same setup. Table 3 compares the annotation performance of our audio-only system on the original data and on the cleaned one. We can see from the table that, there is obvious improvement for each concept class.

This suggests that in the future work we can look for better ways to address the video annotation problem from the audio sensor point of view, both manually and automatically. Some work in [22] can be related to this task.

Table 2. Perofrmance of audio-only, visual-only and fusion systems

Concept	Audio Only	Visual Only	Fusion	Audio Weight
beach	12.8%	69.7%	70.0%	0.14
car	24.5%	77.7%	77.8%	0.15
chn_anti_bldg	19.9%	75.2%	76.2%	0.21
city_view	22.1%	73.3%	74.9%	0.27
dog	47.9%	60.6%	68.2%	0.39
flower	26.0%	80.8%	81.4%	0.21
food	7.0%	59.7%	59.8%	0.06
football_game	72.8%	98.2%	98.5%	0.22
kids	40.5%	53.9%	60.7%	0.50
party	25.8%	85.0%	86.1%	0.22
MEAN	**29.9%**	**73.4%**	**75.3%**	**-**

Table 3. Performance comparison (original data vs cleaned data)

Concept	Original	Cleaned
dog	47.9%	55.3%
football_game	72.8%	75.6%
kids	40.5%	43.1%
MEAN	**53.7%**	**58.0%**

The evaluation criteria in the Huawei grand challenge is defined as follows (we call it Huawei accuracy):

$$\begin{cases} accuracy = sign(score \geq th) \\ sign(x) = \begin{cases} 1 & if\ x\ is\ true \\ 0 & else \end{cases} \\ score = \frac{\sum_{i=1}^{10} Interval_{ip} \cap Interval_{ig}}{\sum_{i=1}^{10} Interval_{ip} \cup Interval_{ig}} \end{cases} \quad (2)$$

$Interval_{ig}$ means the ground truth annotation interval(s) for concept i, and $Interval_{ip}$ means the prediction interval(s) for the same concept. An extra threshold will be used to get a discrete score of 1 or 0 for each prediction (the threshold is set 0.5 in the evaluation). Finally, all the scores will be summed up to get the Huawei accuracy. We achieve 19.2% Huawei accuracy (on original data) with our audio-only annotation system (with 3sec padding to all segments in the post-processing). Meanwhile, visual-only annotation system achieves 63.1% Huawei accuracy, and the fusion of both systems improves the Huawei accuracy to 65.6%.

6 Discussions

From our system annotation results, we discover the salient co-occurrence of "car" and "city view", "football game" and "car". We then check the ground truth (the manual annotation data) and notice that the "kids" concept do often co-occur with other concept such as "football game" and "beach", the "car" concept often co-occurs with "city view" and "football game" （Figure 3）. In Figure 3 the number (in thousands) in the table cell indicates the number of co-occurrence frames between the concept on X axis and the one on Y axis. Such co-occurrence information can potentially help with automatic annotation. In future work, we will consider using the co-occurrence information (such as co-training) to improve the system.

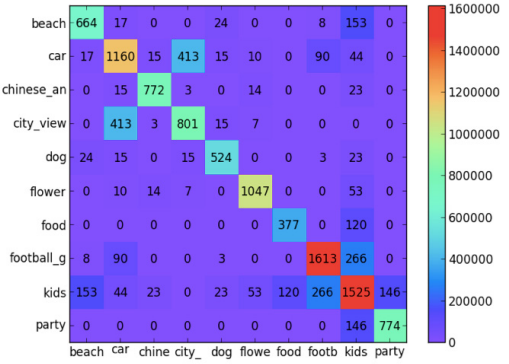

Fig. 3. Concepts' frame-level co-occurrence in the manual annotation

As shown in table 2, we can see that the performance of audio-only system varies among different concepts. This is not surprising, because although some concepts are visually consistent, they hardly have acoustic consistency. For example, the audios from "food" videos vary from incredibly noisy to completely silent. Some concepts cannot be an audio-concept at all, such as "flower". Most of the audios from flower videos are post-edited music. Even if we get a reasonable result on this dataset, we are not modeling the flower concept, rather the editor's music taste for flower. If the music taste changes, the model will fail. Therefore, more attention should be paid to the consistency of the audio data.

We can also see from results in table 2 that combining audio-only and visual-only systems benefits in general. It is even more beneficial on certain concepts, such as "kids" and "dog". In extreme cases when visual evidence is not available, such as the object is hidden but its sound is heard, audio-only system is the only solution for concept annotation in this case. For example, for a snap shot in the video tr0213.mp4 with kids talking behind the camera at a circus performance (Figure 4 (a)), audio-only system successfully detects the "kids" concept while visually is impossible. For another snap shot in the video tr0209.mp4 with kids talking behind the camera about

(a) tr0213.mp4 (b) tr0209.mp4

Fig. 4. Snap shot with kids' voice in the background

the chicken (Figure 4 (b)), audio-only system also successfully detects "kids", while visual-only system may only be able to detect "chicken" or "animal". These examples indicate that for acoustically salient concepts when visual system fails to detect, audio annotation systems will be the best solution.

7 Conclusions

This paper presents our semantic concept annotation system using audio stream only for the Huawei grand challenge. The system uses bag-of-audio-words representation based on the low-level features and negative bootstrap SVM concept classifier. The experimental results on the challenging Huawei UGC video data show that our audio-only concept annotation system can detect semantic concepts significantly better than random guess. When combining with visual-only concept annotation system, it helps in general and more significantly on certain concepts. In the future work, we will explore different low-level features to build the acoustic vocabulary, since MFCC may not be the best feature to distinguish non-vocal sound. We will study the impact of audio data consistency on annotation performance and explore the potential of utilizing the concept co-occurrence property. We will also study building audio annotation systems on more acoustically salient concepts that visual method might fail.

Acknowledgements. 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), and SRFDP (No. 20130004120006).

References

1. Snoek, C., Worring, M.: Concept-based Video Retrieval. Foundations and Trends in Information Retrieval (2009)
2. Chang, S.F., Ellis, D., Jiang, W., Lee, K., Yanagawa, A., Loui, A.C., Luo, J.: Large-Scale Multimodal Semantic Concept Detection for Consumer Video. In: International Workshop on Multimedia Information Retrieval (MIR) (2007)

3. Naphade, M.R., Smith, J.R., Tesic, J., Chang, S.F., Hsu, W., Kennedy, L., Hauptmann, A., Curtis, J.: Large-Scale Concept Ontology for Multimedia. IEEE Journal MultiMedia 13(3) (2006)
4. Over, P., Awad, G., Michel, M., Fiscus, J., Sanders, G., Kraaij, W., Smeaton, A.F., Quéenot, G.: TRECVID 2013 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In: Proceedings of TRECVID. NIST, USA (2013), http://www-nlpir.nist.gov/projects/tvpubs/tv13.papers/tv13overview.pdf
5. Lee, K., Ellis, D.P.W.: Audio-Based Semantic Concept Classificationfor Consumer Video. IEEE Transactions on Audio, Speech, and Language Processing 18(6) (2010)
6. Atrey, P.K., Kankanhalli, M.S., Jain, R.: Information Assimilation Framework for Event Detection in Multimedia Surveillance Systems. In: Multimedia Systems, pp. 239–253 (2006)
7. Kolekar, M.H., Sengupta, S.: Semantic concept extraction from sports video for highlight generation. In: International Conference on Mobile Multimedia Communications (MobiMedia) (2006)
8. Luo, H., Fan, J.: Building Concept Ontology for Medical Video Annotation. In: ACM Multimedia (2006)
9. ICEM 2014 Huawei Accurate and Fast Mobile Video Annotation Challenge, http://www.icme2014.org/huawei-accurate-and-fast-mobile-video-annotation-challenge
10. Wold, E., Blum, T., Keislar, D., Wheaten, J.: Content-based Classification, Search, and Retrieval of Audio. IEEE Multimedia 3(3) (1996)
11. Saunders, J.: Real-time Discrimination of Broadcast Speech/Music. In: ICASSP (1996)
12. Scheirer, E., Slaney, M.: Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator. In: ICASSP (1997)
13. Williams, G., Ellis, D.P.W.: Speech/Music Discrimination Based on Posterior Probability Features. In: Eurospeech (1999)
14. Ma, L., Milner, B., Smith, D.: Acoustic Environment Classification. ACM Transactions on Speech and Language Processing 3(2) (2006)
15. Eronen, A., Peltonen, V., Tuomi, J., Klapuri, A., Fagerlund, S., Sorsa, T., Lorho, G., Huopaniemi, J.: Audio-based Context Recognition. IEEE Trans. on Audio, Speech, and Language Processing 14(1) (2006)
16. Brown, L., et al.: IBM Research and Columbia University TRECVID-2013 Multimedia Event Detection (MED), Multimedia Event Recounting (MER), Surveillance Event Detection (SED), and Semantic Indexing (SIN) Systems. In: TRECVID Workshop (2013)
17. Jin, Q., Schulam, F., Rawat, S., Burger, S., Ding, D., Metze, F.: Categorizing Consumer Videos Using Audio. In: Interspeech (2012)
18. Xue, X.B., Zhou, Z.H.: Distributional Features for Text Categorization. IEEE Transactions on Knowledge and Data Engineering 21(3) (2008)
19. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: CVPR 2007 (2007)
20. Li, X., Snoek, C., Worring, M., Koelma, D., Smeulders, A.: Bootstrapping Visual Categorization With Relevant Negatives. IEEE Transactions on Multimedia 15(4) (2013)
21. Maji, S., Berg, A., Malik, J.: Classification using international kernel support vector machines is efficient. In: CVPR 2008 (2008)
22. Zha, Z.-J., Wang, M., Zheng, Y.-T., Yang, Y., Hong, R., Chua, T.-S.: Interactive Video Indexing with Statistical Active Learning. IEEE Transactions on Multimedia 14(1), 17–27 (2012)