KU-ISPL TRECVID 2014 Multimedia Event Detection System

Seongjae Lee, Han Wang, Minseok Keum, Dubok Park, Hyunsik Choi, Zaur Fataliyev, and Hanseok Ko

Intelligent Signal Processing Laboratory, Korea University

Abstract

In this paper, KU-ISPL's system for TRECVID 2014 Multimedia Event Detection (MED) task is described. The system essentially combines various attributes for scene analysis; audio information for Acoustic Scene Analysis (ASA), Automatic Speech Recognition (ASR), visual information for Visual Scene Analysis (VSA), and Optical Character Recognition (OCR). In the fusion process, heterogeneous data from each module is incorporated, and then the rank of event for each video can be generated according to the final score. In order to verify the effectiveness of our system, we have participated with the following tasks of MED; Pre-specified 010EX, Pre-specified 100EX, Ad-hoc 010EX, and Ad-hoc 100EX. The result of our first participation from NIST shows that the Mean Average Precision (MAP) scores of each MED task were 2.3%, 4.6%, 2.1%, and 2.7%, respectively.

Introduction

The importance of video retrieval technique is receiving more attention these days with drastic growth of online video data. As one of the main objectives of an annual workshop, MED in TREC Video Retrieval Evaluation (TRECVID) [1 2] is to process and analyze the video big data provided by NIST. Since the goal of MED is to analyze the multimodal data which is comprised of audio and visual features, it is regarded as a task of developing effective feature extraction scheme of local descriptors. Accordingly, we processed the HAVIC data [3] on our implemented MED system that has 4 local descriptors and a fusion module, and subsequently submitted Pre-specified task (20 events) and Ad-hoc task (10 events) (PS/AH) to the evaluation system. Table 1. Contains the final results we have received from NIST.

Table 1. 2014 MED result of MAP (%)

Pre-Specified (32k videos)		Ad-Hoc (32k videos)	
010Ex	100Ex	010Ex	100Ex
2.3	4.6	2.1	2.7

Methods

2.1 Visual Scene Analysis

In the preprocessing stage, we capture the image signal from a video file at every 10 seconds and extract visual features from the acquired image signals. Robust low-level image features have been

proven to be effective in terms of representing a variety of visual recognition tasks such as object recognition and scene classification. However, since pixels or even local image patches include insufficient semantic meanings for high level visual tasks, such low-level image representations are potentially inappropriate for the tasks. Therefore, in our system, the object bank method [4] is adopted to assure robust VSA with basic visual feature analysis. It is an image representation constructed from the responses of many object detectors, which can be considered as the response of a "generalized object convolution".

To be more specific, a 177 object detectors are applied to an input image at multiple scales. For each object at each scale, a three-level spatial pyramid representation of the resulting object filter map is applied. Then, the maximum response for each object in each grid is computed. By means of the calculated maximum response for each object it is possible to establish a feature vector for each grid which contains the values for all objects. Finally the VSA module receives the concatenated features (44,406 dim) of all grids of an input still frame.

2.2 Acoustic Scene Analysis

2.2.1 System overview

Unlike the previously discussed VSA module, acoustic and speech information for acoustic scene analysis cannot be utilized in all video analysis scenes. For example, input videos may have silent sections due to the diverse recording environment. Moreover, a certain kind of things may not contain meaningful information for each event or may be corrupted by background music. While the general instructional video includes only speech information, and thus the acoustic event detector cannot be used in this case while speech recognition is applicable.

The developed ASA module consists of three components; music detector which determines existence of music, which is a subsequent part of the ASA module, Voice Activity Detection (VAD) for capturing speech sections, and the acoustic event detector.

2.2.2 Self-taught Learning based acoustic feature extraction

The universal dictionary is generated by using spherical K-means clustering. The conventional Non-negative Matrix Factorization (NMF) is widely utilized for acoustic dictionary learning [5] and its non-negativity constraint leads to renowned part-based representation. On the other hand, it also produces fragmented basis. We therefore utilize spherical K-means based dictionary learning [6] to resolve this basis fragmentation problem.

The training dataset for universal dictionary consists of TIMIT speech corpus, music samples, and general environmental sounds. We empirically set the number of basis of speech, music, and general sound as 100, 200, and 1000, respectively by considering the variations that can occur in each category. This procedure finally generates a 1,300-dimensional overcomplete dictionary. We extract the features of input data using NMF, and then transmit them to the predetermined universal dictionary for music detection, VAD, and event detection. These features account for the consisting ratio of basis for each frame of a given input signal.

2.2.3 Music detection

The self-taught learning method [7] extracts both music and non-music features. We applied the pre-determined universal dictionary to music and non-music data sets with NMF algorithm, which generates a 1,300 dimensional acoustic feature. The music dataset consists of 151 samples which are included in HAVIC DB and some songs from a personal music collection, and a 250 non-music dataset is arbitrary chosen from HAVIC DB. The feature vector of each frame is summed to generate representative single feature vector for the sample, and is divided by L2-norm in order to eliminate the dependency of energy quantity. After that, the Support Vector Machine (SVM) with linear kernel is used for the music/non-music classification. From the self-experiment, it was shown accuracy of the method was 93%.

Unlike our initial expectation, only intentionally overlapping music contents were detected as music class, while the background music in the scene was classified as a normal situation. It is assumed that this result was obtained due to the difference of channel characteristics.

2.2.4 Voice Activity Detection

In order to segment speech sections which are to be transmitted to the ASR module, we apply self-taught learning based VAD. In the first step, by applying the universal dictionary to speech data with NMF, the basis associated with speech is chosen. As in the music detector, energy normalization scheme is also applied in order to remove the energy dependency. After summing NMF gain values across the frame, this histogram is sorted and then we set the knee point as a threshold for speech relevant basis. By means of the process, 200 basis with high gain are selected. In the second step, as a speech detection procedure, we calculate the mean of gain from 200 speech basis and 1,100 non-speech basis. When considering the risk of false alarm and miss rate, we assumed that the risk of miss is much higher than the false alarm. The possibility of incorrectly detected speech for the ASR results the number of keywords to be small. Thus, it is likely that there will be critical information loss by missing the speech. Therefore, we had a policy to loosely set the detection threshold, and set the speech-to-nonspeech energy ratio to 0.8 empirically. If the given frame exceeds this threshold, the frame is determined as a speech section. As a post-processing procedure, we employed a smoothing technique which incorporates temporal context by removing short durations of false alarm and miss.

2.3 Automatic Speech Recognition

Among the entire set of videos, data including speech information occupy only a minor portion. Moreover, clear read-style speech is rare compared to the noise corrupted speech signals. It is mainly exists in events of instructional purpose (e.g., Cleaning an appliance, Renovating a home). Accordingly, the ASR is applied to events characterized by the speech contents with certain confidence. When the ASR is applied to event with rare speech presence, the false alarm of keyword may deteriorate performance. ASR is only conducted to the acoustic signal that is detected as a speech segment by the previous VAD module. We adopted Pocketsphinx released by CMU [8] for speech recognition, which limited its target language to English. The Wall Street Journal (WSJ)

Table 2. The number of pre-defined keywords for each event **(Top: PS event set Bottom: Ad-hoc event set)**

E22	Cleaning an appliance	47	E24	Giving a directions	36
E26	Renovating a home	155	E30	Working on a metal crafts project	81
E31	Beekeeping	70	E33	Non-motorized vehicle repair	64
E34	Fixing musical instrument	106	E40	Tuning musical instrument	51
E42	Building a fire	127	E48	Doing a magic trick	86
E49	Putting on additional Apparel	95			

acoustic model and 64k vocabulary 3-gram of Gigaword text corpus publicly released by Keith Vertanen [9] are used for constructing the ASR system.

2.4 Optical Character Recognition

The occurrence of subtitle information of video is relatively sparse compared with speech and acoustic information. Accordingly, we created 150 wordbooks to recognize the event based on keyword appearance frequency, and sampled the still frame every 2 seconds to extract subtitle information. However, there are many outliers disturbing the correct recognition, when the whole image region was considered. To avoid such problem, we reduced the dimension of image to a pre-specified size, and detected the text region as a pre-processing stage. The conventional methods based on Harris corner [10] and MSER [11] have shown poor computational complexity, and are vulnerable to artifacts generated from compression. We therefore applied the Canny edge detector to overcome the limitations of such methods. Then, the dilation process is conducted in order to find local maxima in edge images. We used the disk-shaped structuring element for dilation. The final text region is detected by applying a pre-specified threshold from the dilated image as shown in Fig. 1. Then, we extract the text using Neural-network based OCR, and the keywords are finally recognized.

(a) Input image (b) Edge image (c) Dilation image (d) Detected region

Fig. 1. The procedure of text region detection.

2.5 Score fusion

We propose two types of score, “numeric score” and “text score” to fuse the heterogeneous data from local descriptors (ASA, ASR, VSA, and OCR). The numeric score, $Score_{numeric}$ and text score, $Score_{text}$ represent the average score of ASA and VSA descriptor, and the frequency

number of keywords from ASR and OCR module, respectively. The numeric score may solely rely on the result of VSA module due to the absence of acoustic information in particular video inputs. Using the following equation, the fusion module carries out the final scoring process.

$$Score_{event_i} = (1 + Score_{text}) * Score_{numeric}$$

where $Score_{event_i}$ stands for the matching level of each video event, and used to rank the candidate videos. The text score, $Score_{text}$ was regarded as additional information, since the occurrence of speech and subtitle is relatively lower than the acoustic event and visual information. It can be generated by the matching frequency between the extracted word set from ASR or OCR module and the pre-defined keywords of each event (word dictionary).

Discussion and conclusion

Based on the obtained results from participated tasks, we have not only confirmed that meaningful information from audio/visual module affects retrieval results but also observed that presence of keyword information in the input improves system performance. As a result, we have learned the great importance of semantic information tagging process. For next year's task, we plan to mitigate the weakness of our current system and improve the performance.

Acknowledgements

This research was supported by Seoul R&BD Program (WR080951)

References

1. Smeaton, A. F., Over, P., and Kraaij, W. (2006) Evaluation campaigns and TRECVid. In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval. ACM Press, New York, NY, 321-330
2. Over, P., Awad, G., Michel, M., Fiscus, J., Sanders, Greg., Wessel, K., Smeaton, A. F., and Quéenot, G. (2014) TRECVID 2014 - An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In proceedings of TRECVID 2014
3. Strassel, S., Morris, A., Fiscus, J., Caruso, C., Lee, H., Over, P., Fiumara, J., Shaw, B., Antonishek, B., and Michel, M. (2012) Creating HAVIC: Heterogeneous Audio Visual Internet Collection. In Proceedings of 8th LREC., 2573-2577
4. L.-J. Li, Su, H., Xing, E. P., and L. Fei-Fei. (2011) Object bank: A high-level image representation for scene classification and semantic feature sparsification. Neural Information Processing Systems
5. Lee, D. D., and Seung, H. S. (2001) Algorithms for non-negative matrix factorization. In Advances in neural information processing systems, 556-562
6. Buchta, C., Kober, M., Feinerer, I., and Hornik, K. (2012) Spherical k-means clustering. Journal of Statistical Software, 50:10, 1-22

7. Raina, R., Battle, A., Lee, H., Packer, B., and Ng, A. Y. (2007) Self-taught learning: transfer learning from unlabeled data. In Proceedings of the 24th international conference on Machine learning, 759-766
8. Huggins-Daines, D., Kumar, M., Chan, A., Black, A. W., Ravishankar, M., and Rudnicky, A. I. (2006) Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. IEEE International Conference on Acoustics, Speech and Signal Processing, 1520-6149
9. http://www.keithv.com/software/
10. Zhao, X., Lin, K., Fu, Y., and Hu, Y. (2011) Text from corners: A novel approach to detect text and caption in videos. IEEE Transactions on Image Processing, 20:3, 790-799
11. Chen, H., Tsai, S. S., Schroth, G., and Chen, D. M. (2011) Robust text detection in natural images with edge-enhanced Maximally Stable Extremal Regions. IEEE International Conference on Image Processing, 2609-2612