KU-ISPL TRECVID 2016 Multimedia Event Detection System

Seongjae Lee, Daehun Kim, Suwon Shon, Seongkyu Mun, Minkyu Shin, Youngseng Chen, Sejong Hyun, Harris Mohammed, and Hanseok Ko[1]

Intelligent Signal Processing Laboratory, Korea University

Abstract

KU-ISPL system for TRECVID 2016 Multimedia Event Detection (MED) is presented in this paper. The deep learning-based local descriptors extract heterogeneous metadata collected from input in a frame-by-frame manner. It consists of Acoustic Scene Analysis (ASA), Visual Scene Analysis (VSA), Visual Motion Analysis (VMA), and Subtitle Information Analysis (SIA). Such metadata can be modeled through the deep learning or the statistical modeling based Event Query Generation (EQG) process depending on the types of metadata. In addition, since the different characteristic of the multimedia events hinders detection performance, a fusion process which combines those various metadata effectively is regarded as significant enhancement factor for MED. Hence, mitigating the detection problem, the system that performs not only elaborate metadata data extraction but also two-fold-constructed metadata fusion method is proposed. Unlike the conventional fusion approach, it is composed of Adaptive Metadata Weighting (AMW) and Dynamic Feature Selection (DFS). It applies selective metadata components to the corresponding multimedia event adaptively so the property of multimedia events can be reflected to the detection system adequately. The experimental results using HAVIC and YFCC set from TREC Video Retrieval Evaluation (TRECVID) 2016 demonstrate that the proposed system attains improved MED performance compared to the conventional approaches as it recorded remarkable higher score in both self-assessment and official evaluation of TRECVID 2016.

Methods

1. System overview

Figure 1 illustrates the overall schemes of the proposed system for MED 16 [1 2]. The local descriptor extracts acoustic and visual metadata from the video input. The dimensionality reduction using Fisher Vector (FV) [5] performs at the video event generation module in order to reduce the massive computational complexity for the visual component of metadata. Each level of feature can be combined at the early fusion module effectively. The module for low-level features essentially normalizes and concatenates similar features such as HOG, HOF, MBHx, MBHy, and TRAJ which are considered as sub-features of VMA. For the sake of modeling the SIA features, the text-based semantic dictionary is generated for each video event. The late fusion module calculates likelihood score for each video event as the very last step. At this step, the scores from low-level and semantic feature are combined by means of the proposed late fusion method so that the optimized result can be obtained.

---

[1] *Director of Intelligent Signal Processing Laboratory*

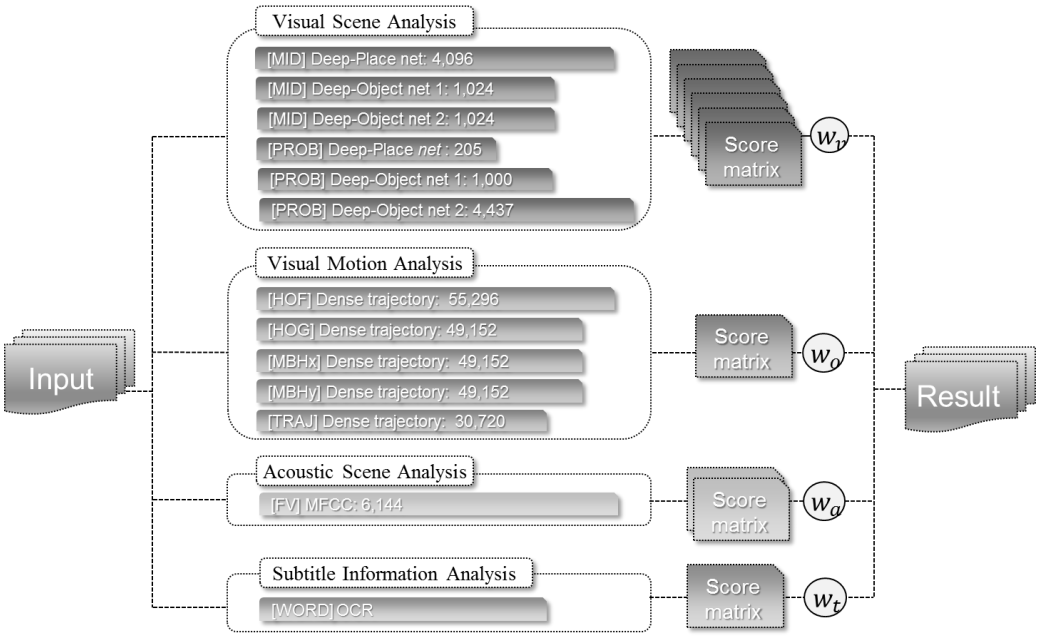

Figure 1. Overall structure of KU-ISPL TRECVID 2016 MED system

2. Metadata Generation (MG)

2.1 Visual Scene Analysis

CNN models recently proved its robust performance on the diverse visual recognition tasks such as object detection and scene classification. The proposed VSA module adopts 3 CNN models which are popularly used in many practical tasks. For the object detection, 2 CNN models developed from GoogleNet (GN) [6] and MediaMill (MM) [7] are adopted. GN model consists of 24 layers, including 23 convolutional layers, 1 fully connected layer with 1,000 nodes. In addition, the size of all the convolutional filters is 3→3 and the stride is to only 1 pixel except the first convolutional filter is that the size and stride of it are 7*7 and 2, respectively. There are 4 max-pooling layers and 1 Avg-pooling layers in the network model. Such structure allows to explore global and more detailed features from feature map with deep layers. It includes the number of 1×1 filters in the projection layer after the built-in max-pooling. Every reduction/projection layers use rectified linear activation and the mid-level features are obtained from layer "Pool5/7x7" with 1,024 nodes. During the MG process, the proposed local descriptor extracts mid-level features and output probabilities, and computes the average values respectively. Since the structure of MM model is based on GN, its convolutional layers and pooling layers are almost same. On the contrary, the output layer is extended to 4,437 nodes for robust classification. The mid-level features are obtained from layer "pool5/7x7" with 1024 nodes. Given an input video clip, the proposed descriptor extracts mid-level features and output probabilities, and then calculates the average values. The average processing time for single video input using such 2 models is 11.58 seconds and 10.94 seconds, respectively. In terms of scene classification, the proposed system employs Places205-AlexNet (AN) model [8]. The model contains 5 convolutional layers and 3 fully-connected layers. The convolutional filters have various sizes from 3x3 to 11x11. During training process, random sampling is applied to both data augmentation and dropout layer. The model is trained by using AN model with 2.5 million

images classified as 205 scenes. The proposed descriptor extracts the mid-level feature (fc7) and the final score for each categories of places for each video frame respectively. The average processing time per video using AN model is 9.99 seconds.

2.2 Visual Motion Analysis

The proposed method adopts the Dense Trajectory (DT) [9] for robust VMA. The visual descriptors of DT are the trajectory by optical flow, HOG, HOF, MBHx and MBHy. The dimensionality of each descriptors is 30, 96, 108, 96, and 96, respectively. In this paper, such dimensionalities except trajectory descriptor can be reduced to half by performing Principle Component Analysis (PCA), and id is encoded as Fisher Vector-based words. The input videos are re-scaled to 320 pixels wide while maintaining an aspect ratio. In order to reduce computational loads, DT uses the even input frame and generates the interesting points for obtaining Bag of Words (BoW) are generated by random sampling method which selects 256,000 samples for each frame. The proposed module adopts two stages for random sampling. The first stage samples random points which assigns 5~10 times for each 60 frames. These points are then sampled randomly again as 256,000 points. As a result, it is able to reduce the processing time for extracting dense trajectory about half. The encoded BoW finally can be used to establish video event model in the EQG process.

2.3 Acoustic Scene Analysis

The i-vector paradigm is adopted for extracting robust ASA feature [10]. The training step is conducted on 39-dimeionsal feature space. At first, 13-dimentiosnal MFCC features are extracted using 25ms Hamming window which is captured by 10ms frame shift manner. The feature warping [11] is then applied to 13-dimensional feature vector for 3 seconds-length sliding window. Delta and acceleration coefficient are additionally appended to produce 39-dimensional feature vectors. Using 39-dimensional MFCC features, Gaussian Mixture Model-Universal Background Model (GMM-UBM) is trained to generate 16 and 256 components for 10ex and 100ex dataset, respectively by considering the amount of each dataset. Using the obtained GMM-UBM parameters, total variability matrix is trained as 400-dimensional total factor using the same dataset which were used for GMM-UBM training. Then, 400-dimensional i-vectors are extracted using total variability matrix Finally, the i-vectors are linearly project on to more discriminative subspace [12] and are normalized by whitening and scaling method so that their lengths are able to fit on the unit length. The model i-vectors of each acoustic scenes are obtained by averaging all i-vectors for each ASA model. In the ES step, the scores between the established ASA model and the search data can be calculated by cosine similarity of two i-vectors.

2.4 Subtitle Information Analysis

The SIA (Subtitle Information Analysis) is to produce a text file with the output of each video separately. It creates 37 wordbooks for recognizing the events in the keyword and capturing frame each 2 seconds to recognize the subtitle in that particular frame. The Artificial Neural Network (ANN)-based SIA algorithm [13] is purposely modified to extract the subtitle information only from

the uniform background in the video. The image variance threshold and the image-resolution normalization is fixed to a determined values.

3. Event Query Generation and Event Search

The diverse dimensionality of each local descriptor is one of the significant factors for robust EQG process. For instance, using too high dimensionality such as FV for deep learning approach usually causes training problem. So the proposed EQG method adopts both statistical and deep-learning approach for VMA and VSA, respectively. The statistical approach is based on Support Vector Machine (SVM) which employs linear kernel. The deep-learning architecture for EQG is composed of 5 layers and variable nodes. The Event Search (ES) module is composed of Adaptive Metadata Weighting (AMW) and Dynamic Feature Selection (DFS). The basic principle of the proposed fusion method is based on the diverse features of video event. For example, video event "Bike trick" usually lacks of speech and acoustic signal since not only the acoustic component of the most event is quite noisy but also background music is usually inserted to the video by user. On the other hands, the visual information such as OFA and VSA are even more robust than acoustic metadata in this event. The AMW module learns the weight values for each video event so the generated weight values then can be applied to the input video. As the next step, the DFS considers every combination of metadata so 2^n (n= number of types of metadata) combination pool can be generated. For example, the different combination (e.g., "Bike trick": VMA+VSA+OCR, "Giving directions": VMA+ASA) can be generated via training step of DFS. Finally, the unique combinations would apply for each corresponding video event.

Results and Discussion

The goal of the self-evaluation was to verify the developed local descriptors and fusion method. The experiment was conducted using 2,000 videos given by HAVIC database set [14]. There were 20 video events which are used as detection class. The dataset was separated as training and testing for each video event. Figure 2 illustrates the average processing time of EQG for each video event. The VMA which includes all DT features recorded the highest complexity due to its dimensionality. Figure 3 presents the performance comparison result between the conventional systems [5 6] and the current version for TRECVID official evaluation. From both self and official evaluation, the proposed system was shown to attain the drastic performance improvement.

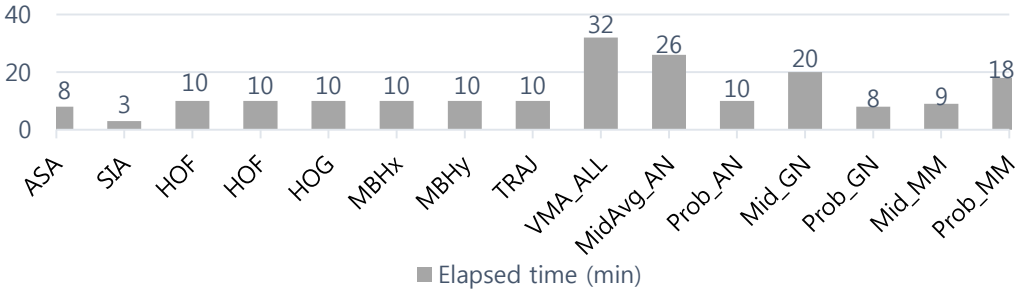

Figure 2. EQG time for each video event

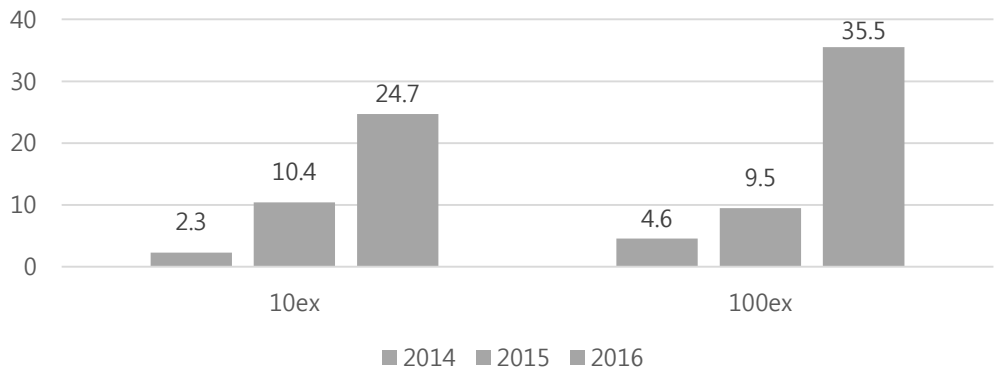

Figure 3. Official Average Precision (AP) score comparison of the KU-ISPL systems

References

1. A. Smeaton, F., P. Over, and W. Kraaij, "Evaluation campaigns and TRECVID", In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, ACM Press, New York, Feb 2016
2. G. Awad, J. Fiscus, M. Michel, D. Joy, W. Kraaij, A. F. Smeaton, G. Quéenot, M. Eskevich, R. Aly, and R. Ordelman, "TRECVID 2016: Evaluating video search, video event detection, localization, and hyperlinking", In Proceedings of TRECVID 2016, Gaithersburg, MD, USA, 2016
3. S. Lee, H. Wang, M. Keum, D. Park, H. Choi, Z. Fataliyev, and H. Ko, "KU-ISPL TRECVID 2014 Multimedia Event Detection System", In Proceedings of TRECVID 2014, Gaithersburg, MD, USA, 2014
4. S. Lee, M. Keum, C. Yang, J. Bae, H. Wang, T. Song, D. Park, D. Kim, J. Ju, and H. Ko, "KU-ISPL TRECVID 2015 Multimedia Event Detection System", In Proceedings of TRECVID 2015, Gaithersburg, MD, USA, 2015
5. J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice", International Journal of Computer Vision, 105(3), 222-245, 2013
6. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich. "Going deeper with convolutions", In Proceedings of CVPR, Columbus, Ohio, 2015
7. P. Mettes, D. Koelma, and C. Snoek, "The ImageNet shuffle: Reorganized pre-training for video event detection", In Proceedings of ICMR, New York, Feb 2016
8. B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database", Advances in Neural Information Processing Systems, Montreal, Canada, 2014
9. H. Wang, A. Klaser, C. Schmid, and L. Lin, "Action recognition by dense trajectories", In Proceedings of CVPR, Colorado, USA, Jun 2011
10. N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification", IEEE Trans. Audio, Speech, Lang. Process, vol.19, No.4, 788–798, May 2011

11. J. Pelecanos, and S. Sridharan, “Feature warping for robust speaker verification”, In Proceedings of The Speaker Recognition Workshop Odyssey, Crete, Greece, Jun 2001
12. D. Romero, and C. Wilson, “Analysis of i-vector Length Normalization in Speaker Recognition Systems”, In Proceedings of Interspeech, Florence, Italy, 2011
13. N. Rao, A. Sastry, A. Chakravarthy, and Kalyanchakravarthip, “Optical character recognition technique algorithms”, Journal of Theoretical and Applied Information Technology, vol.83. No.2, 275-282, Jan 2016
14. S. Strassel, A. Morris, J. Fiscus, C. Caruso, H. Lee, P. Over, J Fiumara, B. Shaw, B. Antonishek, and M. Michel, “Creating HAVIC: Heterogeneous audio visual internet collection”, In Proceedings of The Eight International Conference on Language Resources and Evaluation, Istanbul, Turkey, May 2012
15. B. Thomee, D.A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, L. Li, "YFCC100M: The new data in multimedia research", Communications of the ACM, 59(2), pp. 64-73, 2016