# KU-ISPL TRECVID 2015 Multimedia Event Detection System

Seongjae Lee, Minseok Keum, Cheoljong Yang, Jeongmin Bae, Han Wang, Taeyup Song, Dubok Park, Daehun Kim, Jaeyong Ju, and Hanseok Ko[1]

*Intelligent Signal Processing Laboratory, Korea University*

## Abstract

This paper describes the KU-ISPL Multimedia Event Detection (MED) system employed for competing in the TRECVID 2015 hosted by NIST. The proposed system utilizes diverse audio/visual information source components which consist of a combination of low-level and semantic-level features and adopts an optimized information fusion technique for accurate and robust video event detection. In particular, the local descriptors of the system are composed of the heterogeneous features extracted from the audio/visual information sources of video contents. The fusion process combines the information source at either low-level or semantic-level so that the individual detection score for video events can be judicially appraised in terms of meaningful representation and significance. The results from self-test and the official evaluation have indicated that the proposed system outperforms the previous version.

## Introduction

In 2015, ISPL of Korea University participated in the following MED [1 2] tasks; 10Ex and 100Ex for MED15EvalSub dataset, which is a subset of the SML hardware group. In order to improve the detection performance over that of previous year's further advanced local descriptors and fusion technique were implemented and applied. In terms of the visual descriptors, the following components were employed; Dense Trajectories and TRAJ sub features. The Dense Trajectories (DT) includes HOG, HOF, MBHx, and MBHy while TRAJ sub features play the role of motion information descriptor. The Deep Convolutional Neural Network (DC) is adopted to analyze the object attribute information in the video frame. In particular, it is used to extract both low-level and semantic-level features. The Scale-Invariant Feature Transform (SI) provides the features for entire scene and object information of each video frame.

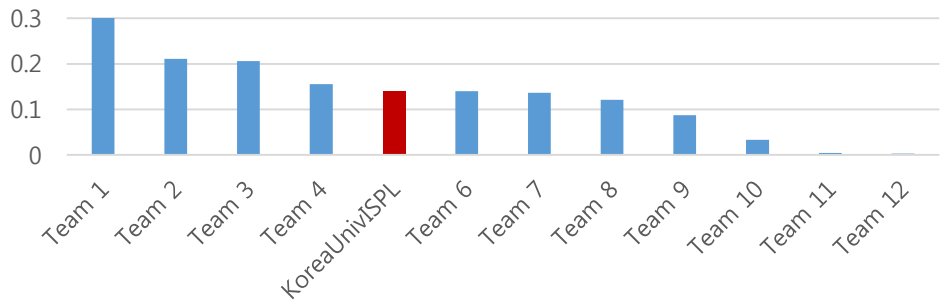


**Figure 1. Result of the InfAP metric of MED 2015 (MED15EvalSub, 10Ex, and SML group)**

[1] *Director of Intelligent Signal Processing Laboratory*

**Table 1. AP score comparison with our 2014 system (MEDEvalSub)**

| 2014 | | 2015 | |
|---|---|---|---|
| 10Ex | 100 Ex | 10 Ex | 100 Ex |
| 2 % | 4 % | 10 % | 9 % |

In order to obtain acoustic scene information, Spectrum Bag-Of-Words (SB) and Temporal Modulation (TM) features analyze the acoustic features in frequency domain by taking various approaches so that a diverse feature set of acoustic information can be yielded. The semantic information which is formed as text word for acoustic and visual scene is also implemented in the employed system. The Optical Character Recognition (OCR) and Automatic Speech Recognition (ASR) module extract the intermittent subtitle and spoken language, respectively. The particular acoustic/visual objects such as vehicle or skid sound in each video frame can be perceived by the Acoustic Scene Analysis (ASA) and Visual Scene Analysis (VSA) module. In addition, we devised an effective fusion method which combines the whole acoustic/visual information for each video event. In addition, we performed a self-test using both the dataset from previous year (e.g. 2014) and the classification accuracy metric for performance validation. As a result, the tendency of the score for each video event was observed following closely to the official evaluation results of NIST. The official results of our second challenge have shown that the Inferred Average Precision (InfAP) score which was the primary metric of assessment recorded 0.1398 among the same competition group as described in Figure 1. Moreover, it is also confirmed that the Average Precision (AP) scores were dramatically improved from that of the previous year [3] as described in Table 1.

# Methods

## 1. System overview

The proposed system is based on the officially required MED processes which comprise of Metadata Generation (MG), Event Query Generation (EQG), and Event Search (ES) module. Figure 2 illustrates the overall scheme of the proposed system. In terms of the MG process, the front-end module extracts acoustic and visual data from the input video clip. Independently extracted components are then injected into the 9 local descriptors. Numerically yielded features are regarded as low-level while text-based extracted features as semantic-level. The dimensionality reduction using Fisher Vector (FV) was performed at the EQG step in order to cope with the massive computational complexity for the visual component of low-level feature. 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 DT, and the internal layers in DC. For the sake of modeling the semantic features, the text based dictionary was generated for each video event. The late fusion module calculates likelihood score for each video event as the very last step of ES. At this step, the scores from low-level and semantic feature are combined by means of the dynamic late fusion method so that the optimized result can be obtained.

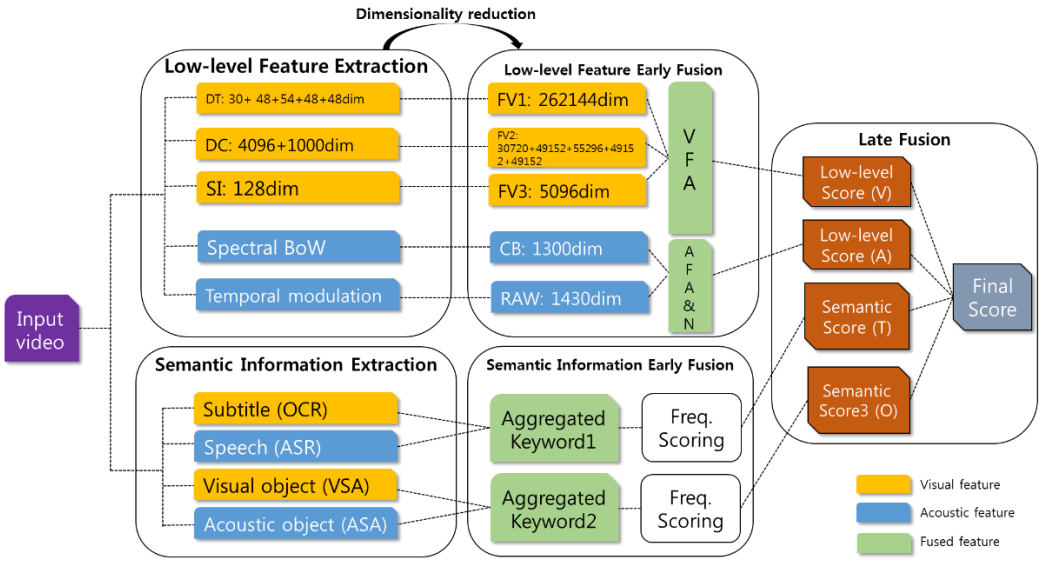


**Figure 2. Overall structure of KU-ISPL 2015 MED system**

# 2. Metadata Generation (MG)

## 2.1 Low-level features

### 2.1.1 Dense Trajectories (DT)

The interesting visual points in video frame can be extracted via DT [10] feature. Furthermore, it is a feasible approach to capture the essential motion information with stacked appearance feature. Because of dense sampling, a tremendous amount of interesting point generation arouses huge complexity for extracting Bag-of-Word of DT. Hence, we re-scaled the input videos to be 320 pixels with a certain aspect ratio and captured only even frames as pre-processing step. The captured frame was then randomly sampled twice for more effective computation as follows. Points were oversampled randomly 5~10 times more for every 60 frames. The sampled points were then sampled randomly once more to be 256,000 points. As a result, the computational time for extracting sub-features of the DT which consist of HOG, HOF, MBHx, MBHy and TRAJ could be reduced by half. Subsequently, the dimensionality of 4 sub-features (HOG, HOF, MBHx and MBHy) were reduced by half by means of adopting Principle Component Analysis (PCA). Finally, we encoded such features with the FV for effective EQG process.

### 2.1.2 Scale-Invariant Feature Transform (SI)

SIFT feature [4] has been widely used in various image recognition tasks. It is specialized to extract the spatial feature around the Difference-Of-Gaussians (DOG) key-points. In our task, the feature with 128 dimensions can be extracted from the corresponding key-point. It then can be projected into 64 dimensions via PCA for dimensionality reduction. In order to encode such information effectively, the FV is then utilized with the vocabulary of size 512 [7]. Hence, the dimensionality of FV would be 65,536 for each SIFT feature. Due to the computational efficiency,

we sampled single frame from 60 frames of video sequences. The cumulated SI information on each video and the video descriptor is then the average of these FVs in sampled frames.

### 2.1.3 Deep Convolutional Neural Network (DC)

We extracted attribute features based on the Convolutional Network (ConvNets) which was comprised of 19 weight layers for obtaining both low-level features and semantic features [5]. The attribute features were trained from the ImageNet 2014 [6] which consisted of 1.4 million images with total of 1,000 Synsets such as “Instrument”, “Canine”, “Vehicle”, and “Bird”. In case of the low-level feature part, we adopted 2 outputs of internal structures instead of using it to not calculate the final detection confidence score. The detailed description of the 2 internal structures are as follows; 4,096 nodes of the last convolution layer which represents the convolutional patch patterns. Outputs of the last full connected layer with 1,000 nodes were directly related with classification class. We sampled such features as single frame from 20 frames of video sequences and averaged the total value which was calculated from the entire video sequence.

### 2.1.4 Spectral BoW (SB)

Spectral feature represents frequency shape of short time frame in temporal sound in video clip. The magnitude spectrum in our system is calculated from a window of 64ms length with 50% overlap by applying Fourier transform. To mitigate the problem wherein spectrum solely captures weak information of an acoustic event from its short time analysis, we utilized the bag-of-words feature which represented the distribution of spectrum codes in the video. Spherical K-means was used for spectrum code generation. As a result, the trained codebook consisted of 100 from speech, 200 from music, and 1,000 from other general scene. The spherical K-means method is preferred because it generates codes which is not dependent on the sound magnitude. Cosine similarity was used for similarity measure between audio input and codebook so that the dependency of sound magnitude could be eliminated while only spectral shape were incorporated effectively.

### 2.1.5 Temporal Modulation (TM)

Since the spectral feature can only represents partial information of the sound without consideration of temporal configuration, we incorporated subband TM features. It was calculated from 20 subbands based on bark scale. The window size was set to 2 seconds while period was set to 1 second. Also, the highest modulation frequency was set at 10Hz. Resultant feature dimension is 1408 (22 by 64). The TM feature, also represented as bag-of-words with the codebook was generated by K-means method and consisted of 2,000 codes.

### 2.1.6 Background music classification

The official HAVIC data is mostly comprised of user uploaded videos and some videos are overlapped with BGM. In this case, acoustic information becomes useless for the MED process. For example, among the 100Ex dataset, about 21% was classified as music. Moreover, the Bike trick

(E021) event contains half of the data as music. Accordingly, we used SB feature based Support Vector Machine (SVM) for BGM classification.

## 2.2 Semantic features

### 2.2.1 Visual Scene Analysis (VSA)

The visual scene information in our system represents the appeared objects in particular video frame. The feature extraction is exactly same as the DC feature mentioned in section 3.1.3. We regarded a numeric vector from 1,000 nodes at the final layer as corresponding object detection score. In order to use meaningful information for EQG and ES, we trained text-based VSA dictionary for event classes via top-*k* attribute features which were calculated from appearance frequency. The value of *k* is determined empirically as described in Figure 3. From the training set, the recognition rate asymptotically converged to 82.42% with *k=39*, and was applied to all video event dictionaries identically.

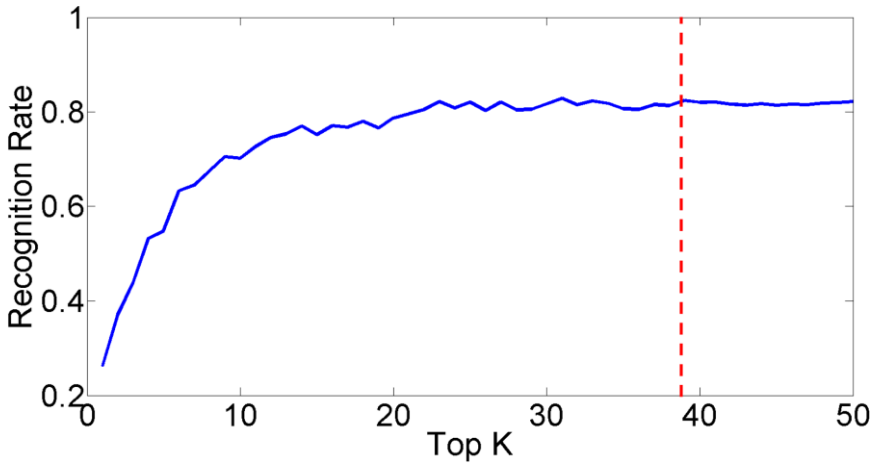


**Figure 3. Convergence of VSA score (*k*=red-dotted line) from the self-test**

### 2.2.2 Acoustic Scene Analysis (ASA)

We incorporated some semantic attribute for acoustic scene analysis. The used acoustic database for training ASA was QMUL freefield1010 consisting of 7,690 audio samples whose duration was 10 seconds for each sample [8]. We constructed an audio codebook with 7,690 codes for each low-level acoustic features (SB and TM). Due to their intrinsic characteristics, the sampling rate of spectral and temporal modulation was of different scale. Spectral feature occurred every 8ms while temporal modulation feature took at every 1 second. For spectral feature, each audio sample was modeled by 10 mixture GMM in order to fully accommodate dense sampling rate, and KNN was utilized for TM because there were not enough samples.

### 2.2.3 Automatic Speech Recognition (ASR)

We have utilized the ASR transcription gratefully provided by LIMSI-CNRS [9].

### 2.2.4 Optical Character Recognition (OCR)

Even though appearance frequency of subtitle in general video clip tends to sparse when it compares to speech information, it plays the role as crucial part for MED task since the subtitle can be a decisive evidence to make a decision of particular video event. Accordingly, we created 37 wordbooks for recognizing the 20 events based on the keyword appearance frequency, and captured still frame for every 2 seconds to extract the subtitle information. The main problem of OCR in our task is that the considerable outliers such as text-like object would hinder correct recognition for non-purified videos. To alleviate such problem, we reduced dimension of image to the pre-specified size, and selected a uniform background image. In addition, we employed statistic information and global variance for detection of the uniform background image with pre-specified threshold which was determined heuristically. Finally, the keywords which was converted from subtitle in video clip can be yielded by means of using neural network based OCR algorithm.

# 3. Event Query Generation (EQG) and Event Search (ES)

## 3.1 EQG

We developed different versions of EQG module for low-level and semantic-level due to the difference of feature type (numeric and text). Similar to our previous system scheme, we employed linear SVM model for low-level EQG. The early-fusion was partially carried out for grouping similar features as pre-processing step. It was applied to DT and DC feature which are composed of the sub-features (HOG, HOF, MBHx, MBHy, and TRAJ) and the sub-layers (Convolution and Output layer), respectively. Since the utilization of the extracted low-level features from all video frame arouses tremendous computational complexity for EQG, the FV method was adopted to reduce the dimensionality. As described in the following Table 2, the elapsed time of EQG for 20 events was reduced at the self-test using COTS machine.

**Table 2. Elapsed time for low-level dataset (Total playback time: 90.8 hours)**

| | DC | DT | SI | MF | TM |
|---|---|---|---|---|---|
| Time (sec) | 3,660 | 13,420 | 17,080 | 6,100 | 5,100 |
| RTF | 0.01 | 0.04 | 0.05 | 0.02 | 0.02 |

For the sake of semantic-level EQG, the text based dictionary for OCR, ASR, VSA, and ASA was established by means of employing high-frequency and meaningful keyword extraction method using Term Frequency- Inverse Document Frequency (TF-IDF).

## 3.2 ES

In order to yield the final detection score, various features can be extracted and combined by means of applying the same method as carried out in EQG process. So there are 9 scores which corresponds to the extracted features from MG can be calculated. The scores for each video event

from 9 local descriptors were finally integrated at the ES process as illustrated in Figure 2. In this process, the scores were transmitted to the late-fusion module which was designed by our new fusion strategy. Such process is able to optimize the final fusion result dynamically through the multiple fusion stages. As described in the experimental result section, our fusion strategy for late-fusion is outperformed when it compared to the baseline fusion method.

## Experimental results

The goal of the self-test was to verify the developed individual modules prior to the official evaluation process. In addition, the fusion strategy was established from the result. The experiment was conducted using 2,000 video samples given by 100Ex dataset from MED 2014. There were 20 video events which were identical to MED 2015 used to MG, EQG, and ES. The dataset was separated 50%-to-50% for training and test, respectively. Since the all dataset were labeled, we used classification accuracy as primary metric for all trials. Table 3 depicts the self-test results without late fusion. There were all low-level features which included early fused sub-features considered to yield classification accuracy. From the above result, we designed late-fusion module and confirmed the drastic performance improvement made as described in Table 4.

**Table 3. Classification accuracy of each local descriptor**

| Feature type | Sub-feature type | Classification accuracy (%) |
|---|---|---|
| DC | Feature layer (4096 nodes) | 31.2625 |
| | Label layer (1000 nodes) | 16.3327 |
| | Early fusion (4096 + 1000) | 30.8500 |
| DT | HOF | 47.4950 |
| | HOG | 44.3888 |
| | MBHx | 21.5431 |
| | MBHy | 36.1723 |
| | TRAJ | 42.5852 |
| | Early fusion (ALL) | 50.2004 |
| SI | - | 64.6293 |
| MF | - | 14.7000 |
| TM | - | 14.5000 |
| **Average** | | **36.4204** |

**Table 4. Effect of the fusion strategy**

| Fusion method | Classification accuracy (%) |
|---|---|
| Baseline | 60.5167 |
| Stage 1 | 64.5375 |
| **Stage 1 + 2** | **70.0500** |

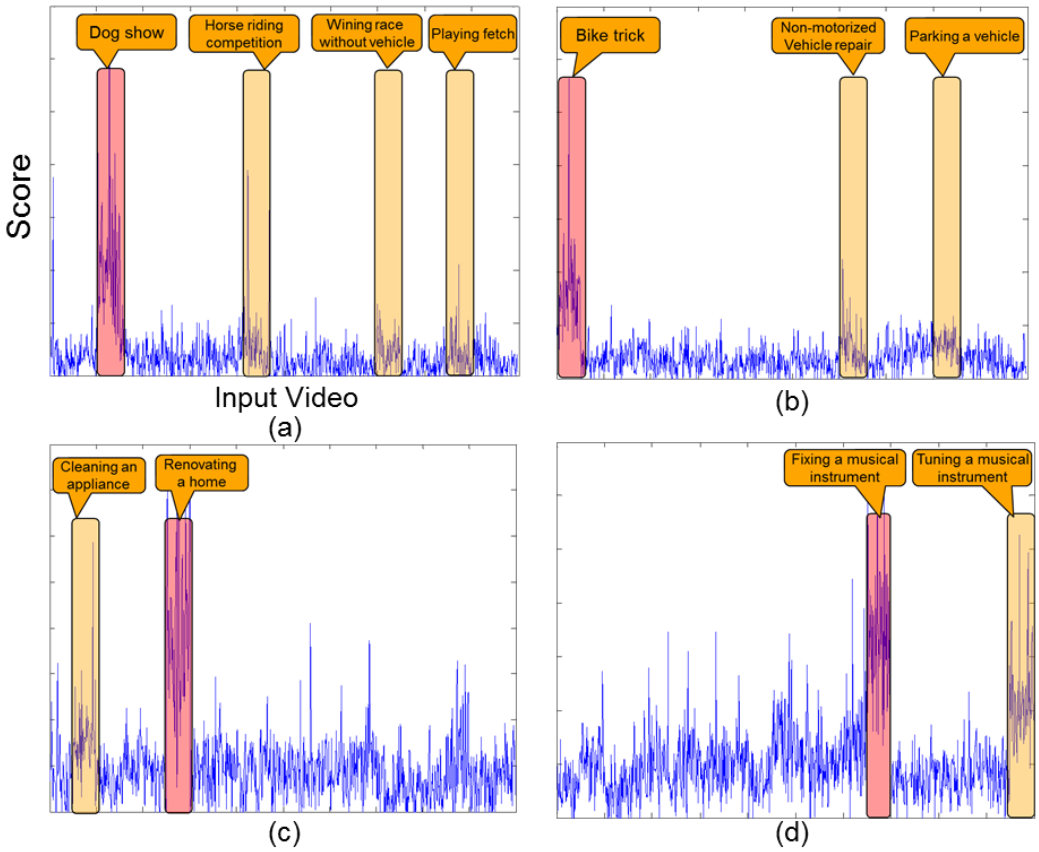


**Figure 4. Similarity of score between TRECVID 2015 events (Red area: True event)**

Figure 4 illustrates the similarity between the 20 video events (E021 to E040). We additionally verified the system's improved performance as the detection scores are recorded high value among similar events. For example, if the true event label for input video was "E021-bike trick", the score for events which motorcycle would appeared frequently such as "E033-Non-motorized vehicle repair" and "E037-Parking a vehicle" is simultaneously higher than other unrelated events (Figure 4. (b)).

# Conclusion

The result of our second challenge was shown dramatically improved over the last year by increased feature types and advanced fusion strategies. The self-test was also shown helpful for our system since it not only verified the local descriptors but also supported toward establishing the fusion strategy. We plan to continue to improve our current system for the next round of opportunity in 2016.

# 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., Wessel, K., Smeaton, A. F., and Quéenot, G. (2015) TRECVID 2015 -- An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In proceedings of TRECVID 2015, Gaithersburg, MD, USA
3. Lee, S., Wang, H., Keum, M., Park, D., Choi, H., Fataliyev, Z., and Ko, H., (2014) KU-ISPL TRECVID 2014 Multimedia Event Detection System. In Proceedings of TRECVID 2014, Gaithersburg, MD, USA
4. Lowe, D. G. (2004) Distinctive image features from scale-invariant keypoints. International journal of computer vision. 60(2): 91-110
5. Simonyan, K., & Zisserman, A. (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
6. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., and Fei-Fei, L. (2014) Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 1-42
7. Sánchez, J., Perronnin, F., Mensink, T., and Verbeek, J. (2013) Image classification with the fisher vector: Theory and practice. International journal of computer vision, 105(3):222-245
8. Stowell, D, and Mark D. Plumbley. (2013) An open dataset for research on audio field recording archives: freefield1010. arXiv preprint arXiv:1309.5275
9. Gauvain, J, Lori L, and Gilles A. (2002) The LIMSI broadcast news transcription system. Speech communication 37(1):89-108
10. Wang, H., Klaser, A., Schmid, C., and Cheng-Lin, L. (2011) Action recognition by dense trajectories. Computer Vision and Pattern Recognition, IEEE Conference on, Colorado Springs, CO, USA