UEC at TRECVID 2011 SIN and MED task

Kazuya Hizume and Keiji Yanai
Department of Computer Science, The University of Electro-Communications, JAPAN
{hizume-k,yanai}@mm.cs.uec.ac.jp

Abstract

In this paper, we describe our approach and results for the semantics indexing (SIN) task and Multimedia event detection (MED) task at TRECVID2011. In our runs of SIN task, we used six features, spatio-temporal (ST) features, SURF, color, face, sound features and word histogram. This year, we use multiple frames selected by calculating the color difference between frames, not all frame. All runs used Multiple Kernel Learning as a fusion method to combine all these features in the same way as last year. Our submitted runs are as follows:

- *UEC1_1: SURF, color, ST, face, sound features*
- *UEC2_2: Run1 & word histogram*
- *UEC3_3: Run2 & sort using a category and video name*
- *UEC4_4: Run2 with TRECVID 2010 training data*

As a result of the full-category SIN task, Run4 yielded the best performance (infAP=0.0452) among four runs.

In MED task, we divide videos to shots which is 150 frames at most and extract SURF, color, ST features from shots. We get the average of the top three shot scores as the original video score.

1. Introduction

Since TRECVID [10] provides not only a large video date set but also a systematic protocol for evaluating video concept detection performance, it is appreciated by the researchers in the field of video/image recognition. Using this valuable date set, we have been testing our system in these years.

For the HLF task in TRECVID2006, we extracted some single types of visual features such as color histograms and edge histograms and classified test frames by the support vector machine (SVM). From the results, we realized that a certain feature cannot satisfy all the concepts. For TRECVID2007, we attempted to adopt a kind of fusion to combine some features to get a result that is effective for any kind of concept. What we did is to apply SVM to the extracted features respectively, and then to fuse these SVM classifiers by linear combination with weights selected by cross validation. This method is more effective, however it is intractable to implement when more than 3 kinds of features are extracted. For the TRECVID2008 HLF task, we still used the thought of developing a framework to fuse a number of features to get more effective performance. At that time we added some new features. In addition, inspired by some papers [2, 13], we implemented a simple version of Adaboost [9] algorithm as a method for late fusion. This method can estimate optimal weights automatically no matter how many kinds of features there are. For the TRECVID2009 HLF task, we explore the feature fusion strategy furthermore. In that year, we used the AP-weighted fusion [14] and Multiple Kernel Learning (MKL) [3, 11] both of which achieved the best performance in our preliminary experiments. For the TRECVID2010 Semantic Indexing Task, we used a novel spatio-temporal (ST) feature [8] which is useful for feature-fusion-based action recognition with Multiple Kernel Learning (MKL). For the TRECVID2011 Semantic Indexing task we use six features including ST feature, word histogram and category name detection and use MKL-SVM in all runs. This year, we don't use Gabor, motion features used last year. We participate in Multimedia event detection task first time this year. We use three features and MKL-SVM for MED task.

2. Overview

Semantic Indexing

This year, we use six features, SURF, color, spatio-temporal (ST) feature[8], face, sound and word histogram. SURF and color features are extracted from multiple selected key frames. Key frames are determined by calculating difference between frame colors. We quantize these features by Bag-of-Features representation, and apply MKL-SVM to model all features.

Multimedia event detection

We use three features, SURF, color and ST feature. Extraction methods of these features are the same as SIN task. We divide each video into shots which consists of 150 frames at most, then extract features from each shot and calculate relevant scores of each shot regarding the given events. The final score is the average among the top three shot scores for each video.

3. Semantic Indexing

3.1. Feature extraction

3.1.1. ST feature

We use a spatio-temporal (ST) feature [8] which is based on the SURF (Speeded-Up Robust Feature) features [1] and optical flows detected by the Lucas-Kanade method [6].

For designing a new ST feature, we set the premise that we combine it with holistic appearance features and motion features by Multiple Kernel Learning (MKL). Therefore, the important thing is that it has different characteristics from other kinds of holistic features. Following this premise, we extend the method proposed in [7]. In the original method, we detect interest points and extract feature vectors employing the SURF method [1], and then we select moving interest points employing the Lucas-Kanade method [6]. In the original and proposed method, we use only moving interest points where ST features are extracted and discard static interest points, because we expect that it is a local feature which represents how objects in a video are moving. In addition to the original method, we newly introduce Delaunay triangulation to form triples of interest points where both local appearance and motion features are extracted. This extension enables us to extract ST features not from one point but from a triangle surface patch, which makes the feature more robust and informative. The characteristic taken over from the original method [7] is that it is much faster than the other ST features such as cuboid-based features, since it employs SURF [1] and the Lucas-Kanade method [6], both of which are known as very fast detectors. The detail should be referred to [8].

3.1.2. Vector Quantization of Features: Bag-of-Frames

In most of existing works on video shot classification, features are extracted only from key frames. However, the extracted features depend on selected frames, and it is difficult to select the most informative key frame. This year, we select frame by calculating the difference of color between frames. First, we capture a frame along time and reduce it to 80x80. Then, we calculate Euclidean distance of RGB color value for each pixel between the selected frame and the base frame, where the first base frame is the very first frame of video. If the distance is greater than the threshold, the captured frame is determined as a frame to extract the features and become the new base frame. The extracted features is vector-quantized and converted into the bag-of-features (BoF) representation within each shot. While the standard BoF represents the distribution of local features within one image, the BoF employed in this paper represents the distribution of features within one shot which consists of several frame images. We call this BoF regarding one video shot as bag-of-frames (BoFr). SURF and color features are extracted from selected frames. While ST features are obtained from every N frame images, we vector-quantize them like the local features.

Also, we adopt spatial pyramid matching technique[4] to BoF representation. We divide the selected frames to 2×2 regions, and generate BoF vectors within each region. We applied this technique to SURF and color features, because these features are extracted from one frame.

3.1.3. Local pattern

We use SURF [5] as a local pattern feature. The local patches are sampled randomly, and they are vector-quantized to convert them into BoFr vectors. The codebook are built by performing the k-means clustering with features extracted from one key frame of all the shots in the training videos. We set the size of the codebook as 1000. Since we use a spatial pyramid with 1×1 and 2×2 regions, totally we generate a 5000 dimensional feature vector.

3.1.4. Color

We extract RGB color histogram features from all pixels of selected frames of each shot. In the same way as SURF, we generate a 5000 dimensional BoFr vector.

3.1.5. Faces

We perform face detection by using Haar-like features[12]. We detect from all frames of each shot, and treat the largest number of face as a 1 dimensional feature.

3.1.6. Sound

For audio feature, we extract mel-frequency cepstrum coefficients (MFCCs) from shots. We use MFCC, log power, ΔMFCC, ΔΔMFCC, Δlog power and ΔΔlog power, 39 dimensional feature. We translate this feature to a 5000 dimensional BoFr vector.

3.1.7. Word histogram

We generate the word histogram by counting words that occur more than three times in the video metadata, and it is weighted in the logarithm of the reciprocal of frequency. The word histogram is a 9539 dimensional features vector.

3.1.8. Feature Fusion Fusion with Multiple Kernel Learning

Multiple Kernel Learning (MKL) is an extension of a support vector machine (SVM). MKL treats with a combined kernel which is a weighted liner combination of several single kernels, while a normal SVM treats with only a single kernel. MKL can estimates weights for a linear combination of kernels as well as SVM parameters simultaneously in the train step. The training method of a SVM employing MKL is sometimes called as MKL-SVM. MKL-SVM is a relatively new method which was proposed in 2004 in the literature of machine learning [3], and recently MKL is applied to image recognition.

Since by assigning each image feature to one kernel MKL can estimate the weights to combine various kinds of image feature kernels into one combined kernel, we can use MKL as a feature fusion method.

In this paper, we use the multiple kernel learning (MKL) to fuse various kinds of image features. With MKL, we can train a SVM with an adaptively-weighted combined kernel which fuses different kinds of image features. The combined kernel is as follows:

$$K_{comb}(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{K} \beta_j K_j(\mathbf{x}, \mathbf{y})$$

$$\text{with } \beta_j \geq 0, \ \sum_{j=1}^{K} \beta_j = 1. \quad (1)$$

where β_j is weights to combine sub-kernels $K_j(\mathbf{x}, \mathbf{y})$. MKL can estimate optimal weights from training data.

3.2. Experiments

Table 1 shows four runs we submitted. For All runs, MKL-SVM is used for the classification method. We increase the features or the data for each run. As the based approach, we use five features(SURF, color, ST, face, sound) in Run1, and add word histogram in Run2. In Run3, we check the original video name of shots. If the video file name contains the words in each category, we add shots that are divided from the video to the top of ranking. In Run4, we add the training data at TRECVID2010 SIN task. That data has only 130 categories, so the result of the other 216 category is same as the result in Run2.

Figure 1 shows the result of all runs of the evaluated 50 categories among the submitted 346 categories. Figure 4 shows the weight estimated by MKL for Run4. Our team reached rank 56 (among 68 team) for the full-category SIN task as shown in Figure 3 and rank 70 (among 102 team) for the light-category. The good results of our teams are Anchorperson, News_Studio, Reporters and Skating. Looking at Figure 4, weight of face features of these categories is greater than other categories, so MKL is applied

Table 1. 4 runs for the semantics indexing task in TRECVID2011.

Runs	Description	full	light
Run1:UEC1_1	Combine SURF, color, ST, face, sound features and Multiple Kernel Learning (MKL)	0.0271	0.0198
Run2:UEC2_2	Run1 & word feature	0.0182	0.0076
Run3:UEC3_3	Run2 & add shots that contains a category name to the top of ranking	0.0202	0.0151
Run4:UEC4_4	use TRECVID2010 dataset for Run2	0.0452	0.0336

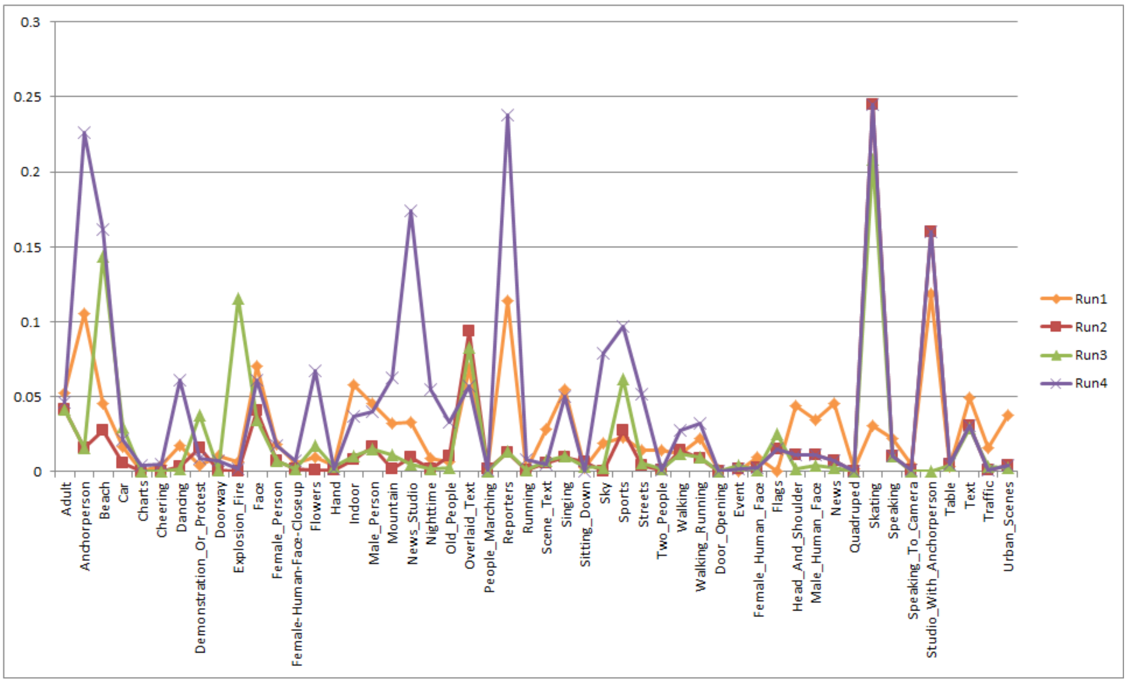

Figure 1. The comparison of results of 4 runs

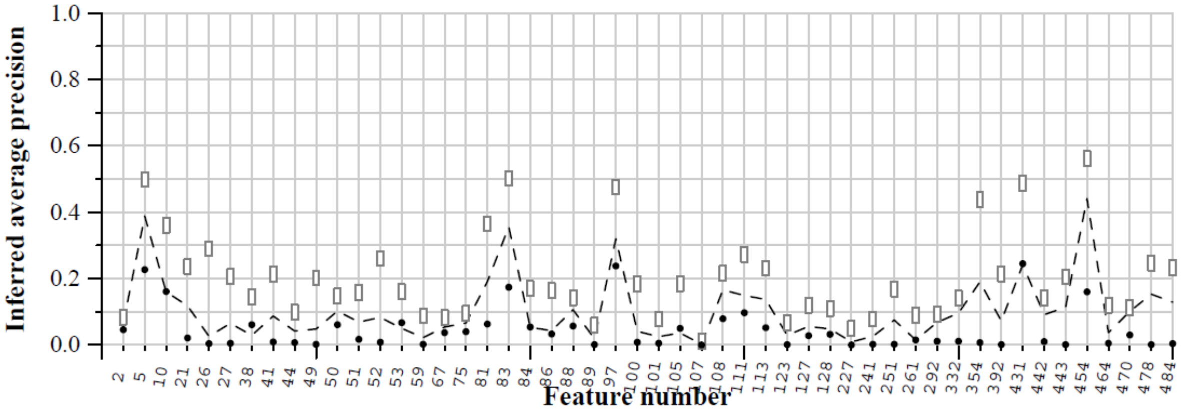

Figure 2. The comparison with median, best and Run4 of full category in TRECVID 2010.

properly. The new categories added in TRECVID2011 are often that the weight of word histogram is larger than the other features. We can't get enough the image features because the training data in TRECVID2011 is smaller than the training data in TRECVID2010. At result, 8 categories out of 16 categories are that the re-

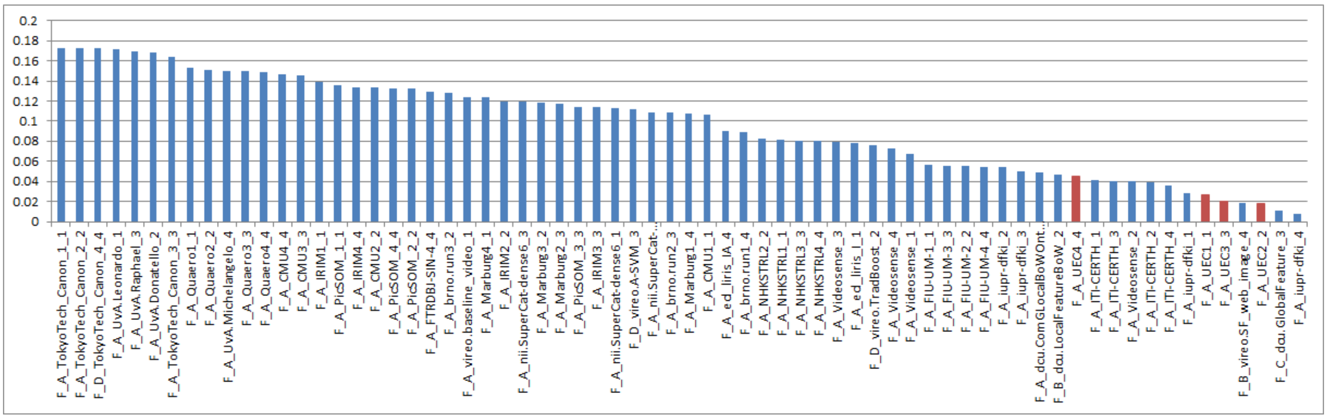

Figure 3. The comparison with results in TRECVID 2011. Red lines show the full-category results of UEC team among 68 runs.

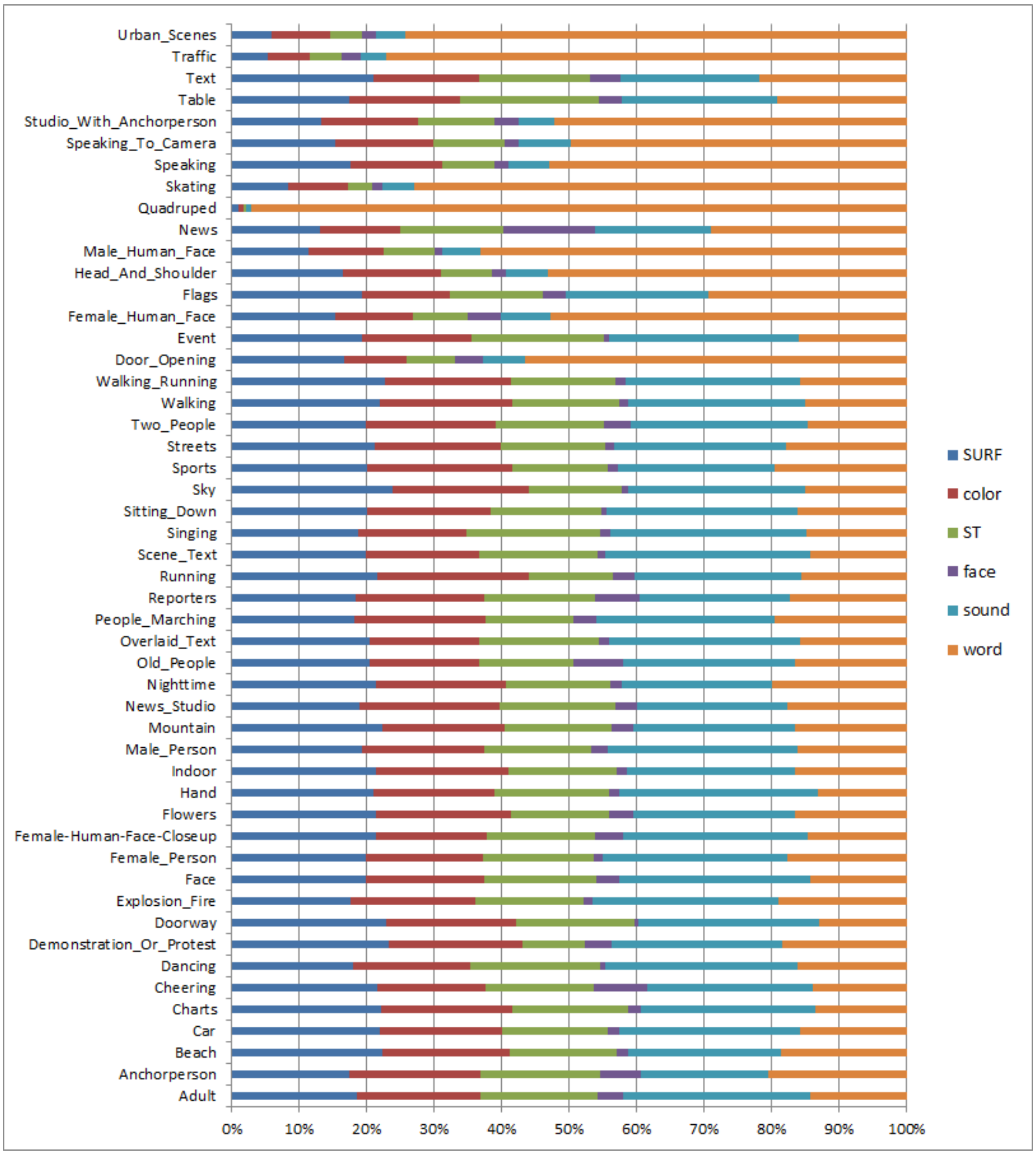

Figure 4. Estimated weights by MKL with full category in TRECVID 2011 Run4.

sult of Run1 which don't include word histogram is better than other runs. In the old categories, the result of Run1 is also worse than Run2 at most categories, so word histogram should be fused in different way, not MKL.

4. Multimedia event detection

4.1. Dividing videos into shots

We apply the system used in the SIN task to the MED task. Our system is intended to recognize per shot, not to recognize per video, so we need to divide the videos of the dataset in the MED task into shots. Each video is divided into shots which consists 150 frames at most, then the features are extracted from each shot.

4.2. Feature extraction

For the MED task, we use three features, SURF, color and ST feature. The feature extraction methods are the same as the SIN task.

4.3. Score and threshold decision

The score is calculated per shot by MKL-SVM. The average of the top three scores of video shots is used for the original video score. Threshold is calculated from scores of learning videos classified by MKL-SVM. We use 100 shots selected from each of the other categories for negative shot, and calculate the average of score for positive or negative shot. We use the mean of the each average for threshold and normalize it with difference between maximum and minimum scores for each category. Each scores of evaluation videos is also normalized with this difference.

4.4. Experiments

Figure 5 and Table 2 shows the result score. events that there is a big move across the screen (e.g. Flash mob gathering, Parade, Parkour) is relatively good results, but events that do the work at hand (e.g. Working on a sewing project, Making a sandwich, Grooming an animal) is bad. This means that our system can not respond to events such as a series of small movements occur. Moreover, our team a number of FA in our team is very large compared to other teams while a small number of miss. Since we use the average of the top three scores of video shots, if there is at least one high-scoring shot our system recognize that the video belongs to the category. This approach cannot detect events that occur continuously, so it is difficult to say that target events can be detected properly. This is a big task.

5. Conclusion

In the Semantics indexing task (SIN) of TRECVID2011, we got multiple key frames by calculating the difference between frames. We used SURF, color, sound features and word histogram in addition to ST features and the number of faces as features, and used Multiple Kernel Learning to combine them. In the best runs among our submission, we have achieved 0.0452 average precision(AP). The results differed from our expectation that it was the good method to combine word features and MKL.

In the Multimedia event detection (MED), we divided videos to shots which are 150 frames at most, then extracted SURF, color, ST features from shots. The original video score was the average of the top three shot scores.

As future work, we plan to explore the key frame selection method and how to handle features. We verify the propriety of this selection method by comparing to use all frames. We should try to use different fusion techniques or features rather than fusion all features by MKL simply. For MED task, we do not process the video for the time directional, so need to improve our system using techniques such as scenario-based.

References

[1] B. Herbert, E. Andreas, T. Tinne, and G. Luc. Surf: Speeded up robust features. *Computer Vision and Image Understanding*, pages 346–359, 2008.

[2] W. Jiang, S. Chang, and A. Loui. Kernel sharing with joint boosting for multi-class concept detection. In *Proc. of CVPR Workshop on Semantic Learning Applications in Multimedia*, 2007.

[3] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[4] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. of IEEE Computer Vision and Pattern Recognition*, pages 2169–2178, 2006.

Table 2. 4 runs for the semantics indexing task in TRECVID2011.

—Title	#Targ	#Sys	#CorDet	#FA	#Miss	PFA	PMiss	NDC	Min PFA	Min PMiss	Min Dec. Thresh
E006	186	31821	175	14745	11	0.4661	0.0591	5.8795	0.0036	0.9462	0.8098
E007	111	31821	98	18696	13	0.5896	0.1171	7.4797	0.0064	0.8018	0.8017
E008	132	31821	131	19261	1	0.6078	0.0076	7.5976	0.0171	0.4167	0.7700
E009	95	31821	90	14675	5	0.4626	0.0526	5.8288	0.0086	0.7684	0.8232
E010	87	31821	85	20995	2	0.6616	0.0230	8.2846	0.0004	0.9885	0.9061
E011	140	31821	138	22317	2	0.7044	0.0143	8.8108	0.0010	0.9714	0.9215
E012	231	31821	226	18474	5	0.5848	0.0216	7.3244	0.0089	0.7749	0.7976
E013	104	31821	100	19835	4	0.6254	0.0385	7.8478	0.0149	0.6346	0.7817
E014	78	31821	68	10850	10	0.3418	0.1282	4.3965	0.0101	0.6795	0.7165
E015	81	31821	63	11301	18	0.3560	0.2222	4.6684	0.0000	1.0000	0.9454

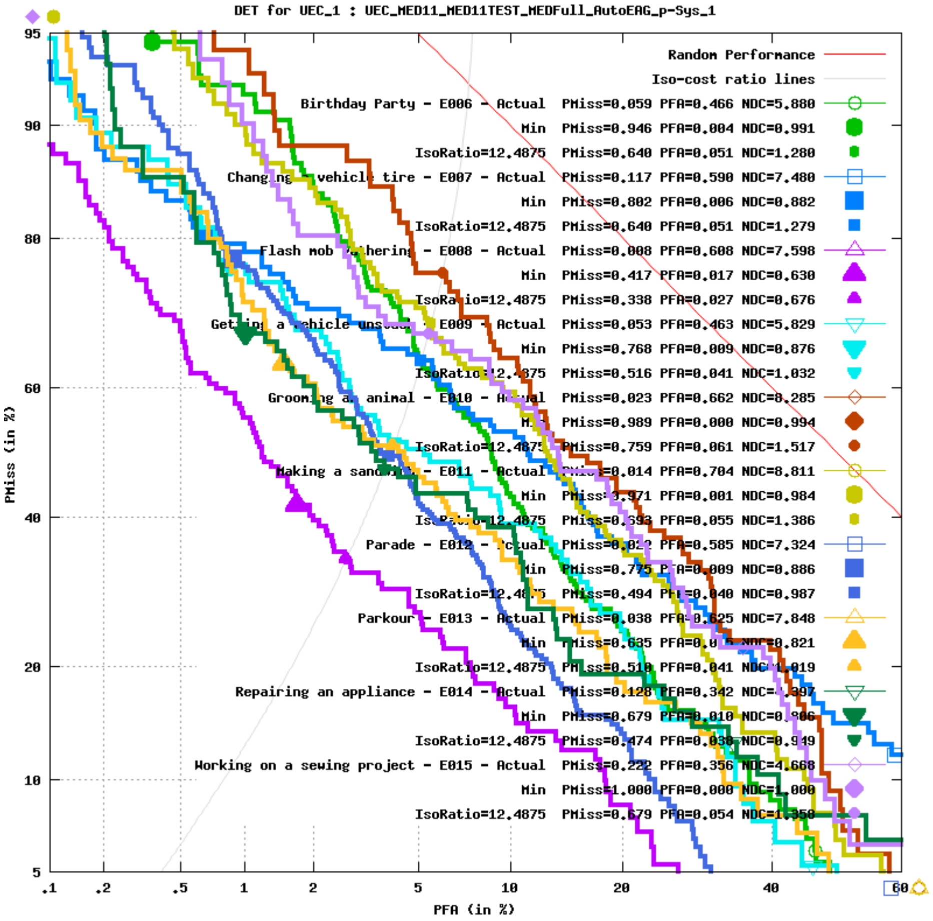

Figure 5. result MED task in TRECVID 2011.

[5] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[6] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. of International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

[7] A. Noguchi and K. Yanai. Extracting spatio-temporal local

features considering consecutuveness of motions. In *Proc. of Asian Conference on Computer Vision(ACCV)*, 2009.

[8] A. Noguchi and K. Yanai. A surf-based spatio-temporal feature for feature-fusion-based action recognition. In *Proc. of ECCV WS on Human Motion: Understanding, Modeling, Capture and Animation*, 2010.

[9] R. Schapire, Y. Freund, and R. Schapire. Experiments with a New Boosting Algorithm. In *Proc. of International Conference on Machine Learning*, pages 148–156, 1996.

[10] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *Proc. of ACMMM WS on Multimedia Information Retrieval*, pages 321–330, 2006.

[11] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *Proc. of IEEE International Conference on Computer Vision*, pages 1150–1157, 2007.

[12] P. Viola and M. Jones. Rapid Object Detection Using a Boosted Cascade of Simple Features. In *Proc. of IEEE Computer Vision and Pattern Recognition*, volume 1, 2001.

[13] D. Wang, X. Liu, L. Luo, J. Li, and B. Zhang. Video diver: generic video indexing with diverse features. In *Proc. of ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 61–70, 2007.

[14] M. Wang and X. S. Hua. Study on the combination of video concept detectors. In *Proc. of the 16th ACM international conference on Multimedia*, pages 647–650, 2008.