BUPT-MCPRL at TRECVID 2012*

Zhicheng Zhao, Yanyun Zhao, Yan Hua, Wen Wang,
Decheng Wan, Guoli Jia, Zhixuan Li, Fei Su, Anni Cai
Multimedia Communication and Pattern Recognition Labs,
Beijing University of Posts and Telecommunications, Beijing 100876, China
{zhaozc, zyy, sufei, annicai}@bupt.edu.cn

Abstract

In this paper, we describe BUPT-MCPRL systems for TRECVID 2012. Our team participated in three tasks: known-item search, instance search and surveillance event detection. A brief introduction is shown as follows:

A. Known-item search

This year we submitted 4 automatic runs based on two different approaches, one of which is text-based and the other is visual feature-based. Results of all 4 runs are described in Table 1.

Table 1. KIS results and descriptions for each run

Run ID	Mean Inverted Rank	Description
F_A_YES_MCPRBUPT1_4	0.350	Text search with spell check and synonym expansion
F_A_YES_MCPRBUPT1_1	0.183	Text search with Lemmatization tool
F_A_YES_MCPRBUPT1_2	0.192	Text search with Stemming tool
F_A_NO_MCPRBUPT3_3	0.011	Visual search with visual attention model and concept/object detection

B. Instance search

This year, we mainly focused on the following parts: selection of distance metric, multimodal fusion and results re-ranking, and finally, we submitted 2 runs and achieved a high infAP.

Table 2. INS results and descriptions for each run

Run ID	infAP	Description
F_X_NO_BUPT.MCPRL_3	0.268	Multi-features fusion with SIFT and focus on parts of an image in each topic
F_X_NO_BUPT.MCPRL_2	0.245	Multi-features fusion with CSIFT and concerning global image in each topic

C. Surveillance event detection

This year, we mainly evaluated the events of PeopleMeet, PeopleSplitUp, Embrace, ObjectPut, PersonRuns and Pointing detection. Our system adopted different algorithms in detecting these events accordingly.

1 Known-item Search

Two different methods were proposed. One is traditional text-based and another is improved bio-inspired method used at TRECVID 2011.

1.1 The bio-inspired method

*This work was supported by National Natural Science Foundation of China under Projects 61101212 and 90920001, and by Fundamental Research Funds for the Central Universities, and Network System and Network Culture Foundation of Beijing.

This method based on the framework we used last year, which is an improved approach inspired by human attention, recognition and binding mechanisms. In this approach, a query topic is first parsed by a text analyzer to produce several search cues, and then the cue-based bottom-up saliency map and the top-down cue-guided concept/object detection are fused and refined by the aid of context cues. With this new bio-inspired method we achieved better results than those obtained by concept-based detection method last year.

A) The proposed framework

The proposed KIS framework is shown in Fig.1.1. It mainly includes five parts: a bottom-up attention model for determining salient regions which popped-out by visual stimuli (color, luminance, texture etc), a knowledge base containing various pre-trained object/concept (such as person, car) detectors, a SOM (Self-Organizing Maps) network to map known-item keywords into seven image-related classes, a SVM-based scene classifier for data filtering, and a fusion module to perform content-based retrieval, results fusion and ranking.

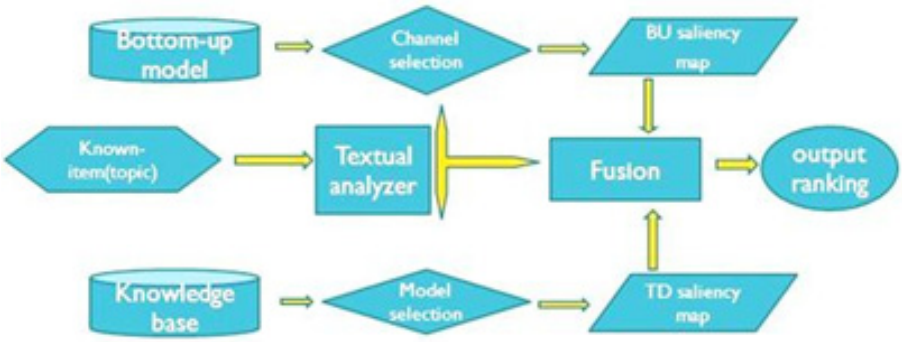

Figure 1.1. The KIS framework of bio-inspired approach.

B) The SOM Network of Words

The human brain can easily parse text sentences and map the semantic information into visual images. Therefore, if relationships between keywords of search topics and image features/concept can be established, the semantic gap could be narrowed.

In our system, a textual feature vector including noun similarity, noun hierarchy, adjective similarity etc computed by WordNet, is extracted to describe each keyword. A SOM network is then trained to cluster keywords into seven categories: color, texture, shape, person, vehicle, position relations between objects and specific semantic words. The first three categories will give the bottom-up saliency cue, next two give the concept detection (top-down) cue, and the last two respectively give the context and data filtering cues. During KIS search, the query keywords are automatically classified by the trained SOM.

C) The Attention Model

According to the two-channel (Where-what) theory of human visual system, in our framework, a bottom-up combined with top-down attention model is built up to detect informative objects described in the search topic. Firstly, the corresponding channels (color, shape and spectral residual) of the attention model are weighted with the bottom-up cues to locate potential salient regions. And then, proper models are selected from the knowledge base according to the top-down cues to locate objects/concepts interested. Finally, a graph-based inference model is employed to fuse the bottom-up

and top-down saliency maps, and the context cues are used to refine the results.

D) Scene Classification

In order to enhance the search speed and performance, a scene classifier based on Gist and SVM is employed to classify video scenes into two categories: outdoor and indoor. In addition, a black and white video detector is also developed. Both classifiers are used to filter out irrelevant videos.

1.2 Text-based method

This year we submitted 3 runs based on this method, named F_A_YES_MCPRBUPT1_1, F_A_YES_MCPRBUPT1_2 and F_A_YES_MCPRBUPT1_4. We adopted the same automatic text-based search system we proposed last year. The difference of each run is that how each part of the system is carried out. The system is consisted of several main components, including text pre-processing, keywords extracting and processing, text-based retrieval, results fusion and re-ranking. The framework is shown in Fig.1.2.

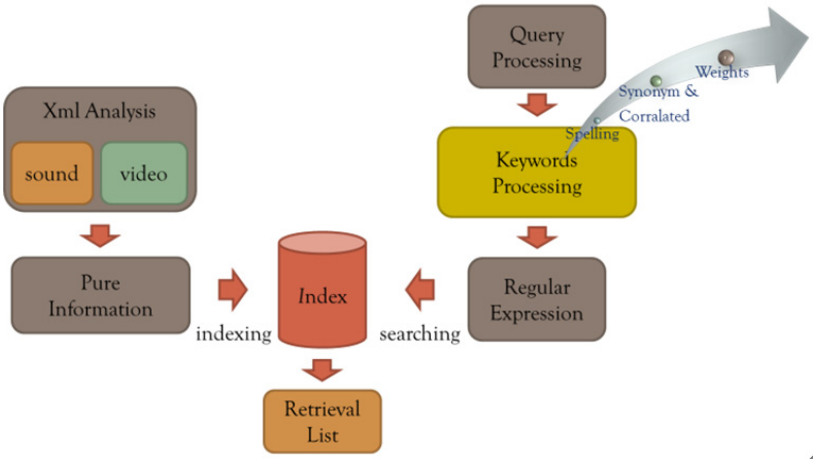

Figure 1.2. The framework of text-based approach.

A) Text Ontology Construction

A text-based ontology is constructed manually for F_A_YES_MCPRBUPT1_4. By using this tree-like lexicon, we can obtain information on whether words are particular colors, language, places, specific terms, sound etc. in search topics in text processing. For F_A_YES_MCPRBUPT1_1 and F_A_YES_MCPRBUPT1_2, we adopted the Lemmatization tool[1,2] and Stemming tool[3] to achieve the same goal, respectively.

B) Text Processing

- Spelling mistakes in queries as well as in metadata and ASR data were corrected.
- With the aid of NLP tools, we eliminated much redundant information in topics and metadata, and extracted the keywords according to their weights.
- Use a custom Stopword list to extract keywords from metadata and query.
- Thanks to Youdao dictionary, we expanded the extracted keywords by finding their synonyms and correlated categories.
- Recomputed the weights of each processed-keyword based on their importance, followed by forming them into regular expressions to search.

C) Text-based Retrieval

In text-based retrieval, keywords from ASR data and metadata were first converted to a stream of plain-text tokens, as explained in Figure 2, and then were sent to Lucene to build text index respectively. Same process was applied to search topics when doing retrieval. For each query, the index of ASR data and metadata were searched individually, and results were combined and re-ranked.

1.3 Results and analysis

The final KIS results of all automatic search runs are shown in Fig.1.3, and 4 red bars are MIR of our 4 different runs: the higher three is based on text only and the lowest one is content-based which is proposed above. Since we submitted 2 runs of F_A_YES_MCPRBUPT1_1 and F_A_YES_MCPRBUPT1_2 in incorrect format, the evaluated MIR are 0.001. We evaluated the performance according to the ground truth, and the actual MIR should be 0.183 and 0.192 respectively.

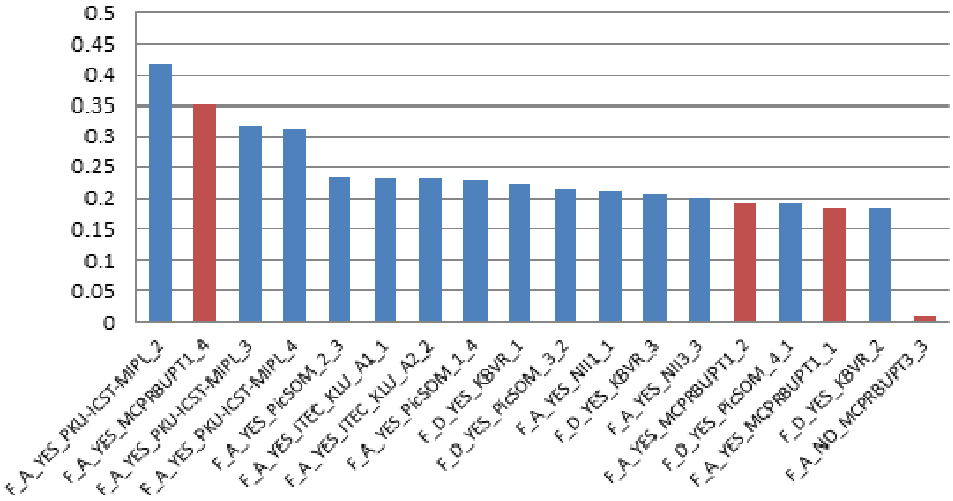

Figure.1.3. Results of all KIS runs.

Fig.1.3 indicates that one of our text-based approaches achieved the best MIR score 0.350 in our 4 runs, which is in second place among all automatic runs, and other two text-based runs achieved 0.192 and 0.183 respectively, and the MIR score of the bio-inspired approach is 0.011. As can be seen from the results, the two runs F_A_YES_MCPRBUPT1_1 and F_A_YES_MCPRBUPT1_2 with Lemmatization and Stemming only reach the average level among all runs, while F_A_YES_MCPRBUPT1_4 achieve a high MIR than others. This trend suggests that spell correction and synonymy expansion play an important role for text-based retrieval.

Although F_A_YES_MCPRBUPT3_3 only achieved 0.011, the MIR score is higher than the result using the same method at TRECVID 2011. Furthermore, it should be pointed out that among all 361 queries, only 37 topics are searched with the proposed method, while the remaining results are randomly generated. If consider the 37 queries only, 9.37% MIR would be obtained, which is shown in Fig.1.4. This trend gives us the confidence that the proposed bio-inspired method is promising if the attention model and knowledge base are further improved.

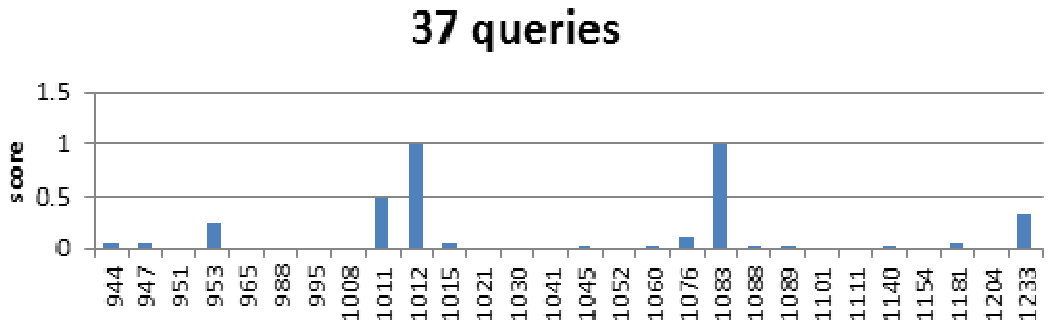

Figure 1.4. Results of 37 search topics.

2 Instance Search

The proposed automatic instance search system is consisted of several main components, including visual query pre-processing, key-frames and features extraction, key-frames retrieval, multimodal fusion and results re-ranking. The framework of our INS system is shown in Fig.2.1.

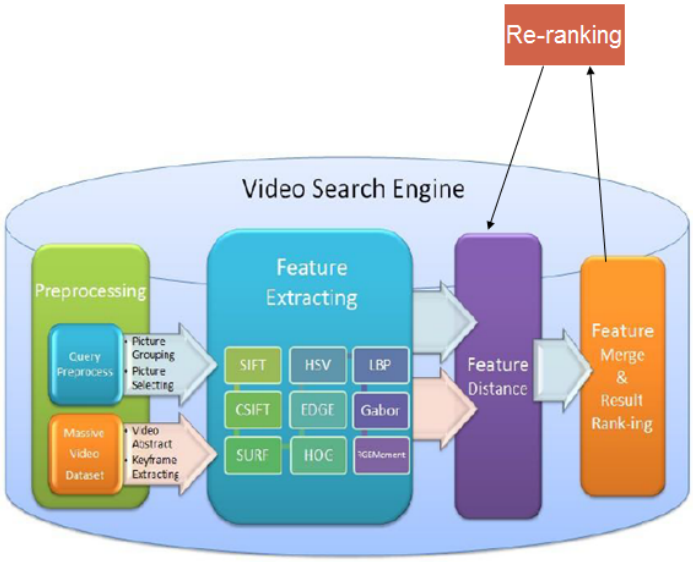

Figure 2.1. The framework of INS system.

2.1 Feature Selection

We extracted several visual features at regional and global levels[4, 5], the details of which are listed in Table 2.1.

Table 2.1. Extracted visual features

Features	Description
HSV Histogram	512 dims HSV color histogram for global partition
RGB_Moment	225 dims RGB color moment feature for global partition
SIFT	SIFT feature and BoW method with 50000 visual words
CSIFT	CSIFT feature and BoW method with 10000 visual words
Gabor Wavelet	3-scale and 6-direction Gabor feature with 3*3 regional partition
EDH	145 dims histogram by concatenating global and regional EDH

LBP	256 dims histogram of each LBP code with global partition
PHOG	3060 dims histogram with 8*8 regional partition and 9 directions
HOG	2520dims histogram with 10*7 regional partition and 9 directions

Compared with our system of TRECVID 2011, some changes are made in feature extraction:

- Except Harris-Laplace key points, we also use dense sampling method to detect key points for CSIFT and SIFT.
- A Pyramid of Histograms of Orientation Gradients (PHOG) descriptor [6] which consists of a histogram of orientation gradients over each image sub-region at each resolution level is added to describe an image by its local shape and spatial lay out of the shape.
- Instead of using k-means, we use approximate k-means (AKM) to cluster visual codebooks with BoW method. [7] shows that AKM gives a very similar performance with less computational costs.

2.2 Similarity Computation

Various normalizations are used for visual features and different distances are used to compare feature's similarity. For features using BoW method (SIFT and CSIFT), we used Gaussian_ normalization, for other features representing shape, color and appearance (EHD, HOG, PHOG, RGB_Moment, HSV_Correlogram, LBP, Gabor) we used sum_normalization. Furthermore, we used Bhattacharyya distance instead of Euclidean distance to compute HSV_Correlogram similarity.

2.3 Result fusion and re-ranking strategy

Firstly, search results based on different visual features are linear fused by preset weights experiential obtained at self-test phase, and then a query expansion method is used to update and re-rank results. The main steps are as follows:

- Choose top 10 results from initial result list as new search instances to re-search.
- Results weighted fusion: all 11 search lists by re-ranked according to (1)

$$w(x) = 1.25^{-x} \tag{1}$$

  where $x \in 0,...,9$ is the ranking of search result.
- Normalization: after re-ranking, the score is normalized to be the final result.

2.4 Results and Analysis

Fig.2.2 shows final search results with 2 runs of different feature merging strategies. One run achieves 0.268 MAP which indicates that our method is feasible.

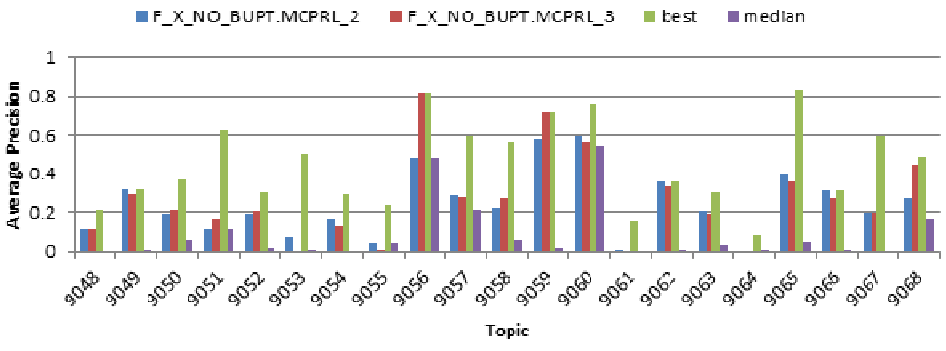

Figure 2.2. Our results vs median and the best.

3 Surveillance Event Detection

This year, we mainly focused on the events of Embrace, ObjectPut, PeopleMeet, PeopleSplitUp, PersonRuns and Pointing. Our system adopted different algorithms in detecting these events accordingly.

3.1 Automatic System

3.1.1 PeopleMeet and PeopleSplitUp Detection

Our system for PeopleMeet and PeopleSplitUp detection is typically the same as, except several modifications, that of the last year. PeopleMeet and PeopleSplitUp detection is based on the low-level features. As these two events are multi-persons activity, they are considered in the whole frame area. We applied MoSIFT[8] features as our low-level features which presents the large video data in an elegant way. When extracting the features, some prior knowledge is introduced. For example, in cam3 there's no chance for a PeopleMeet or a PeopleSplitUp event to occur at the top of scene, so we just discard the features drawn from this area. After extracting the features, bag of the words method is used to find the meaningful features' centers to avoid divergence. Then we build a cascade classifier, each stage of which is a random forests[9] classifier, to train and classify the samples generated by the sliding window method.

3.1.2 Embrace, ObjectPut, PersonRuns and Pointing Detection

For Embrace, ObjectPut, PersonRuns and Pointing detection, the main frame is similar with that of PeopleMeet and PeopleSplitUp detection. Since these four events include only two persons at most, we divide a frame into 12 blocks so that each block covers much fewer people than the whole frame does. Such operation can block out mass noises. For each block we build a codebook and detect events respectively. To detect these events We applied two kinds of features, one of which is MoSIFT and the other is a new introduced one.

We introduce a novel interest point detector based on optical flow vorticity and divergence to detect these events. First, a dense optical flow field is calculated. Then, we compute the vorticity and divergence of the field. At last the points with larger vorticity or divergence value is selected as interest points. The descriptor of the points is its trajectory and HOG/HOF[10].

3.2 Interactive System

The interactive system is an extension of the automatic system. The framework of the interactive system is shown in figure 1. For each time the automatic system returns the results, a manual intervention is applied to select the correct detections and dislodge the false positives. Then the samples including both correct and incorrect detections are relabeled and sent to the training set to retrain the classifier, which is worked as a feedback system. After several runs of this process, the performance of the classifier will be improved in the future tasks. In our experiment, the automatic system is trained using data in 08dev. Then data in 08eval is used to update the classifiers, where the manual selection is replaced by comparing the output labels and annotation labels. We considered the process mentioned above as training stage. Then we test our system on 09eval data and accomplished manual selection. The result of interactive system is shown in Fig. 3.1.

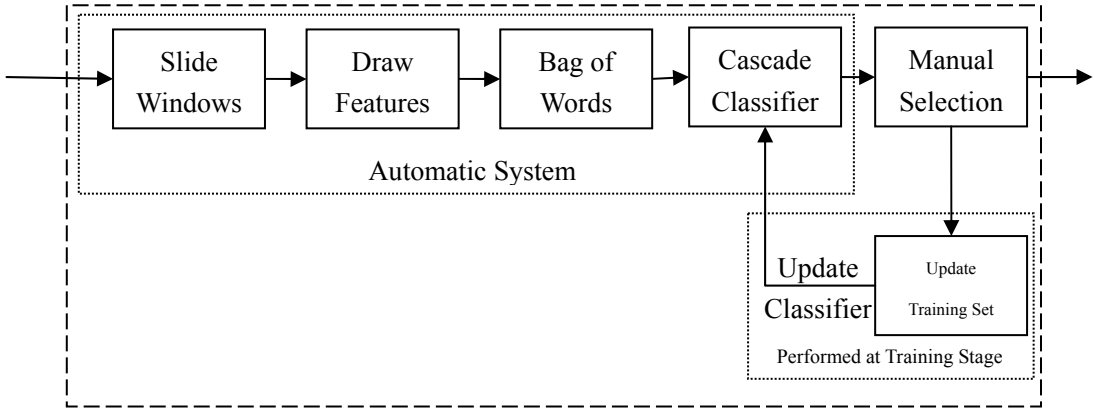

Figure 3.1. The framework of interactive system

3.3 Result and Conclusions

There are some problems in our SED algorithm needed to be solved. For individual behaviors, person detection and tracking are the two most important parts in improving system performance. The second one is to research more discriminative features to describe human motion. The third is that the rate of false alarms should be further suppressed.

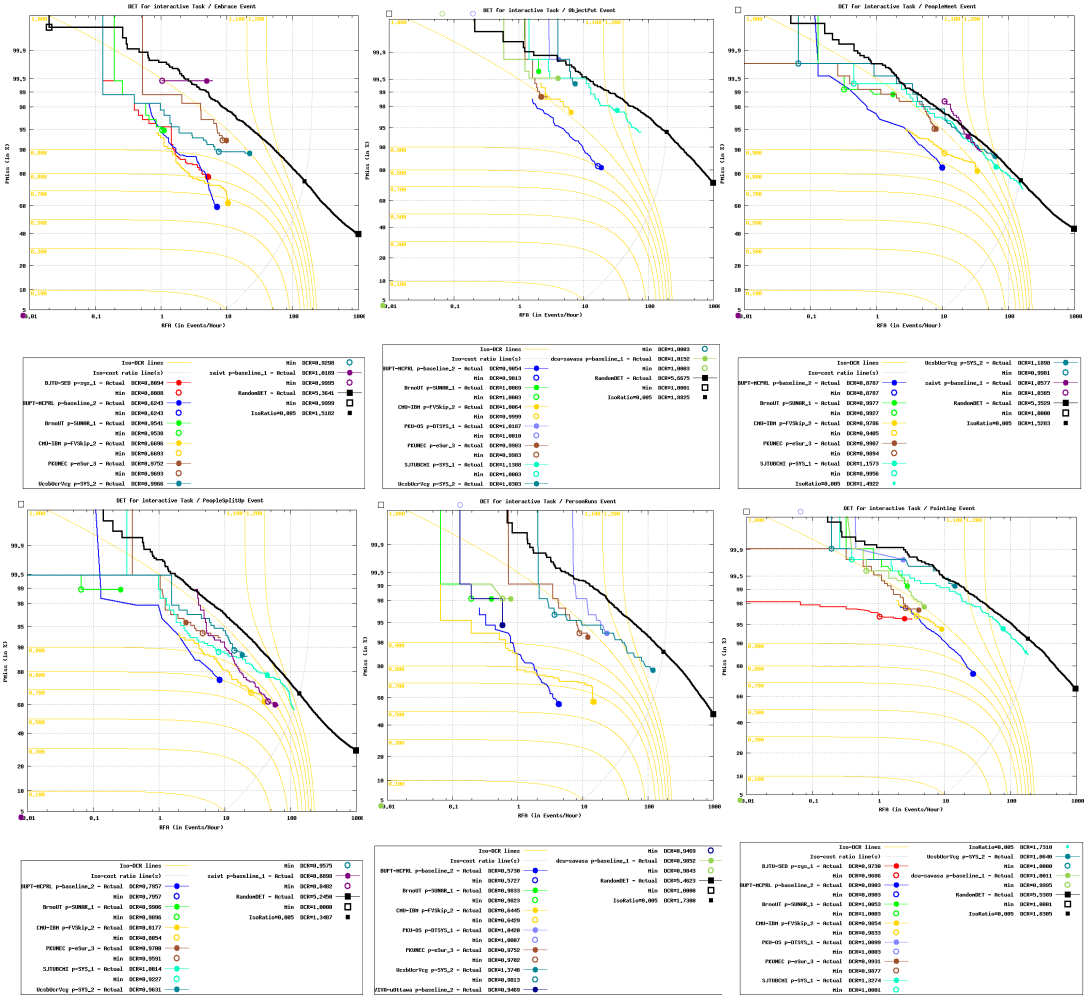

Figure 3.2. Our results of interactive system

References

[1] Kristina Toutanova and Christopher D. Manning. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70.

[2] Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259.

[3] Online available at: http://tartarus.org/martin/PorterStemmer.

[4] Xiaoming Nan, Zhicheng Zhao, Anni Cai et al, “A Novel Framework for Semantic-based Video Retrieval”, ICIS 2009.

[5] Zhicheng Zhao, Yanyun Zhao, Zan Gao, Xiao ming Nan et al, “BUPT-MCPRL at TRECVID 2009”, In: Proceedings of TRECVID 2009 Workshop.

[6] A. Bosch, A. Zisserman et al “Representing shape with a spatial pyramid kernel”, CIVR 2007

[7] J. Philbin ,O. Chum et al, “Object retrieval with large vocabularies and fast spatial matching”, CVPR 2007.

[8] M. Chen and A. Hauptmann. Mosift: Recognizing human actions in surveillance videos. Computer Science Department, 2009.

[9] Breiman, Leo. Random Forests. Machine Learning 45 (1): 5-32.

[10] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In ECCV, 2006.