BUPT-MCPRL at TRECVID 2013*

Zhixuan Li, Yuhui Huang, Kaiqi Zhang, Zhicheng Zhao, Yanyun Zhao, Fei Su
Multimedia Communication and Pattern Recognition Labs,
Beijing University of Posts and Telecommunications, Beijing 100876, China
{zhaozc, zyy, sufei}@bupt.edu.cn

Abstract

In this paper, we describe BUPT-MCPRL systems for TRECVID [6] 2013. Our team participated in two tasks: automatic instance search and surveillance event detection. A brief introduction is shown as follows:

A. Automatic instance search

In our work, we divide the topics into 2 kinds, i.e. object and person according to the query description, and treat differently for each kind. However, because of the errors in key-frame extraction, we get a low infAP score.

Table 1. INS results and descriptions for each run

Run ID	infAP	Description
F_X_NO_BUPT.MCPRL_2	0.014	BoW scheme
F_X_NO_BUPT.MCPRL_3	0.019	BoW scheme with global feature

B. Surveillance event detection

This year, we focus on the events of ObjectPut, PersonRuns, Pointing, PeopleMeet, PeopleSplitUp, and Embrace. These events are divided into three groups. Our system adopted different algorithms in detecting events accordingly.

1 Automatic instance Search

1.1 Object retrieval

We adopt both local and global features. For local feature, we first choose key-points by Hessian-affine detector and describe them using the SIFT, and then generate a 63k generic codebook by approximate k-means clustering [1] with training images crawled randomly from Flickr [2]. Then, each descriptor is assigned to the closest cluster center in feature space and all the local features are aggregated followed by the BoW scheme. For global feature, we use a 512-dims HSV correlation histogram to describe the global color distribution of the image.

For 4 given query images and their corresponding masks of each topic, visual vocabularies as well as HSV correlation histogram are first extracted from each image. Then for each query, we have two vocabulary sets, i.e. $V_i = \{v_{i1}, \dots, v_{iv}\}$, $M_i = \{m_{i1}, \dots, m_{im}\}$, where V_i is an unordered set of visual words from the whole image and M_i is an unordered set only containing visual words from ROI set by the mask. Note that both V_i and M_i are unordered sets, which means we are only interested in whether the visual word appears in the image, but ignore the times it appears due to the sparseness of the aggregated feature. We then take the union of M_i, i.e. $M = M_1 \cup M_2 \cup \dots \cup M_n$ as the

*This work is supported by Chinese National Natural Science Foundation (90920001, 61101212, 61372169), National High and Key Technology Research and Development Program (2012AA012505, 2012BAH63F00), National S&T Major Project of the Ministry of S&T (2012ZX03005008), the Special Funds of Beijing Municipal Co-construction Project, the Fundamental Research Funds for the Central Universities, and Network System and Network Culture Foundation of Beijing.

vocabulary set we need to pay special attention.

Then, for each reference image, we also get an unordered set of visual words and a HSV correlation histogram as its feature. Also, the idf coefficient for each visual word is calculated from the reference set as $idf_i = lg(N/N_i)$, where N is the total number of images and N_i is the times visual word i appears in an image.

After the features from both query and reference images are extracted, the similarity between each query image and reference is calculated as below:

1) Find the intersection between query i and reference j: $U = V_i \cap R_j$;
2) Set a weight α_k to each $u_k \in U$ in this way: if $u_k \in M$, a high weight is set (in our experiment, we set $\alpha = 3$), otherwise, $\alpha_k = idf_k$;
3) Similarity between query and reference is: $sim\left(V_i, R_j\right) = \sum_k \alpha_k \ /(|V_i| + |R_j|)$.

After step 3, we have a similarity score for each reference image. Then, RANSAC is used to further identify the spatial relationship between each query image and reference. We then combine the number of inliers together with the similarity score computed before to get the final score for each image: $score = log(inlier \cdot \alpha) \cdot exp(sim \cdot \beta)$. In our experiment, we set $\alpha = 1, \beta \ = 10$.

In addition, the similarity of the HSV correlation histogram between each query and reference image is calculated in terms of the Mahalanobis distance. We fuse the score of local and global features with their linear combination.

1.2 Person retrieval

In the off-line phase, a face detector is adopted to find the face regions in each reference image, followed by a 3200-dims LBP feature extraction to describe each detected face. In the on-line phase, a similar process is done for each query image. However, for query image, instead of the whole image, the face detector is only used in the ROI confined by the given mask. Then, the Euclidean distance is used to measure the distance between LBP features from each query and reference face region. Since each topic contains four query images, for each reference image, its shortest distance to all the query face regions is chosen as the distance to the topic. In this way, we obtain the rank list for each topic.

1.3 Result

We submitted two runs this year. Because of the errors in key-frame extraction, some of the indices of our key-frame are not in correspondence to the ground truth data. Also, we only extract one key frame for each shot which is far from enough to describe video content. Thus, we get inferior MAP score, shown in Table.1.

For the first run, we only use local features in object retrieval. For the second run, we adopt a combination of both local and global features. We see a slight improvement in the second run which indicates the global feature works.

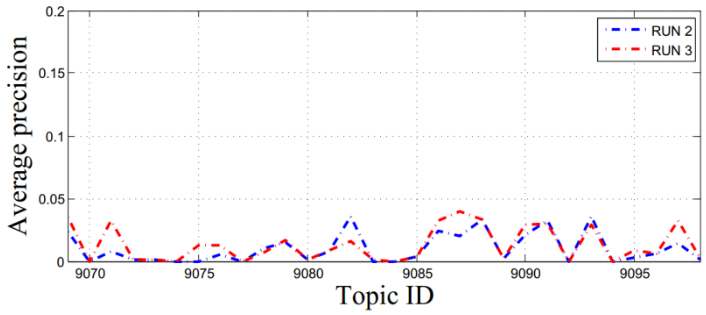

Figure 1. Results of two submitted runs.

2 Surveillance Event Detection

This year, we focused on the events of ObjectPut, PersonRuns, Pointing, PeopleMeet, PeopleSplitUp, and Embrace. These events are divided into three groups. Our system adopted different algorithms in detecting events accordingly.

2.1 Automatic System

2.1.1 PeopleMeet, PeopleSplitUp and Embrace Detection

We calculated the probability of occurrence and built a hot map for these events respectively. When extracting the features, we just applied algorithms on the predefined hot region of a frame. We applied dense features [3] as our low-level features. After that, bag of words method was used to find the meaningful features' centers to avoid divergence. Then we built a cascade classifier, each stage of which is a SVM classifier, to train and classify the samples generated by the sliding window method.

2.1.2 ObjectPut, PersonRuns Detection

Since ObjecetPut and PersonRuns events are single person activity, we applied pedestrian detection and tracking algorithms [4] to get window for the target. Then we extracted dense features on the target window as samples and sent them into cascade SVM classifier for training, which is similar with the multi-persons activity.

2.1.3 Pointing Detection

For the pointing event, we also got the target window by pedestrian detection and tracking algorithms. Then we located the upper body of the target in the window. After that we extracted a five-dimensional feature, which includes body length and angles between body and limbs [5]. These features describe human body as a "stickman". The classifier is the same as other events.

2.2 Interactive System

The interactive system is an extension of the automatic system. The framework of the interactive system is shown in figure 2. For each time the automatic system returns the results, a manual intervention is applied to select the correct detections and dislodge the false positives. For the elapsed time limit, we just replayed the events whose scores exceeded the score threshold chosen, so that the interactive time was less than 25 minutes. During interactive procedure, we corrected wrong event labels. This process reduced the false alarm significantly, but contributed little to the missing activities.

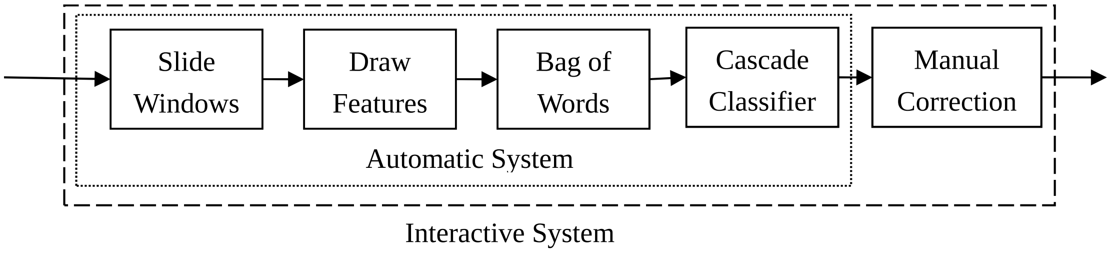

Figure.2. The framework of interactive system

2.3 Result and Conclusions

Our results of SED 2013 are shown in figure 3. During the SED test, we were confronted with some problems. First, we missed a large amount of positive samples while we were exacting the features of single person activities. An improvement on the pedestrian detection is requisite. Second, compared with single event, the hot region is still too large, in which the features may interfere with each other. The features and region restriction need to be researched further.

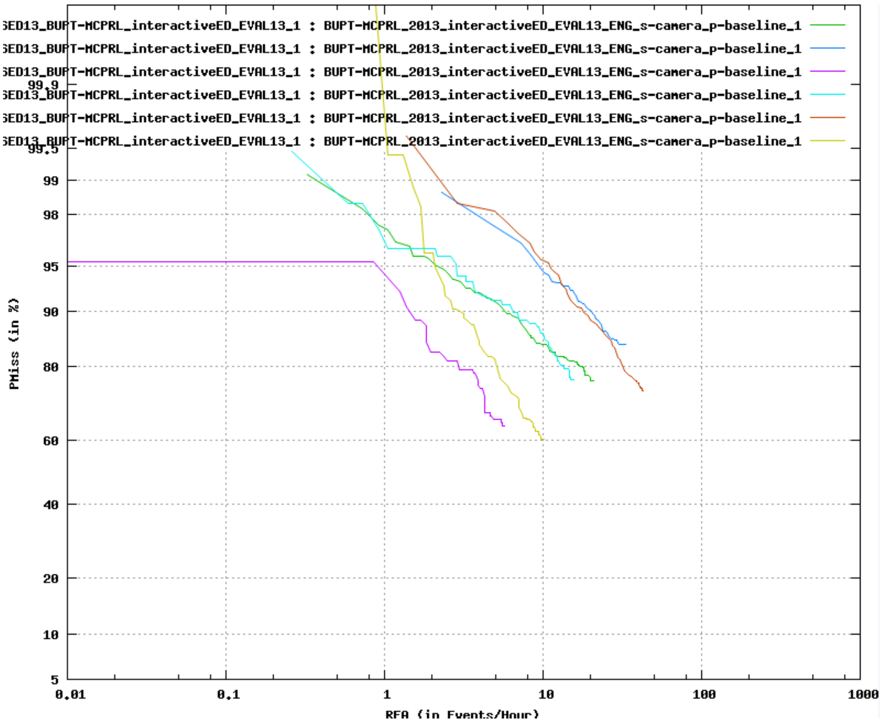

Figure.3. Results of interactive system

References

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

[2] www.flickr.com

[3] Heng Wang, Alexander Klaser et al. Action Recognition by Dense Trajectories. CVPR, 2011.

[4] Amir RoshanZamir, AfshinDehgan et al. GMCP-Tracker: Global Multi-object Tracking Using Generalized Minimum Clique Graphs. ECCV, 2012.

[5] M. Eichner, M. Marin-Jimenez et al. 2D Articulated Human Pose Estimation and Retrieval in (Almost)Unconstrained Still Images. Computer vision(2012) 99:190-214.

[6] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders,W. Kraaij, A. F. Smeaton, and G. Queenot. Trecvid 2013 { an overview of the goals, tasks, data, evaluation mechanisms and metrics. In Proceedings of TRECVID 2013. NIST, USA, 2013.