PKU-IDM @ TRECVid 2010: Pair-Wise Event Detection in Surveillance Video

Kaihua Jiang [b1], Zhipeng Hu [a*], Zhongwei Chen [c*], Guochen Jia [a], Teng Xu [a], Qiong Hu [c], Guangcheng Zhang [b]
Yaowei Wang [a], Lei Qin [c], Yonghong Tian [a+], Xihong Wu [b], Wen Gao [a]

[a] National Engineering Laboratory for Video Technology, Peking University
[b] Speech and Hearing Research Center, Peking University
[c] Key Lab of Intel. Inf. Proc., Institute of Computing Technology, Chinese Academy of Sciences
[+] Corresponding author: Phn: +86-10-62758116, E-mail: yhtian@pku.edu.cn

Abstract

In this paper, we describe our system for the surveillance events detection task in TRECVid 2010. We focused on pair-wise events (e.g., PeopleMeet, PeopleSplitUp, Embrace) that need to explore the relationship between two active persons. For our team had participated in the TRECVid SED task in 2009, we developed the system based on the old one. The improvements are three-fold. First, we refined the background subtraction method of last year. Some better background frames are automatically selected to train and update the background model and the background reconstruction is performed at pixel level instead of frame level. Second, we employed a MPL (Multi-Pose Learning) based method for head-shoulder detection, which can effectively improve the detection recall. Third, a structural SVM (SVM-HMM) classifier is employed for pair-wise events detection. According to the comparative results in the TRECVid SED formal evaluation, our experimental results are promising.

1. Introduction

This year we chose four events and focused on pair-wise events (e.g., PeopleMeet, PeopleSplitUp, Embrace) that need to explore the relationship between two active persons. As our team had participated in the TRECVid SED task in 2009, we developed this year's system based on the old one (eSur). The improvements are three-fold.

First, we refined the background subtraction method of last year. Some better background frames with fewer foreground objects are automatically selected as training samples to train and update the background model by comparing video frame with a Gaussian background model, and the background reconstruction is performed at the pixel level instead of the frame level. Experimental results show that the method can detect the foreground objects sensitively with much lower false alarms than the classic background modeling methods.

Second, within the extracted foreground region, we used the cascaded HoG(Histograms of Oriented Gradients) [8] for head-shoulder image reorientation, and apply Multiple Pose Learning-based RealBoost for classifier learning. The online Boosting method is then used for tracking each detection part. Intermediate experimental results show that our human detection and tracking technique, together with background modeling, obtains better performance than last year.

Third, a structural SVM classifier is employed for pair-wise events detection. As the events videos are inherently sequential data, we introduced the Hidden Markov Support Vector Machine (SVM-HMM) to model and classify the interactive events with consideration of the statistical dependencies over adjacent frames. Features like distance between two persons are extracted from every frame. Instead of simply concatenating the features into a vector, we treat them as sequential data to exploit not only the discrete information from individual frames, but also the sequence and correlation information among frames. The final detections are parsed from raw sequential results generated by SVM-HMM.

The remainder of this paper is organized as follows. In section 2, we present our system framework briefly. Background subtraction is described in section 3. In section 4, we describe our head-shoulder detection and tracking approach. In section

[1] These persons are equally important in the contest. This work is partially supported by grants from the Chinese National Natural Science Foundation under contract No. 61035001, No. 60973055, No. 61072095 and No. 61003165, National Basic Research Program of China under contract No. 2009CB320906, and Fok Ying Dong Education Foundation under contract No. 122008.

5, we present our approach for detecting different events in given surveillance video sequences. Experimental results and analysis are given out in section 6. Finally, we conclude this technical report in section 7.

2. The eSur System Framework

The diagram of our eSur system is shown in Fig.1. The whole system is similar to the one we developed last year and the main difference lies in the event detection module. Last year we used classic linear SVM classifiers and automata to classify and identify different events in this module. However, this module is completed replaced by SVM-HMM and outlier classifier this year. Besides, some significant improvements are achieved in the background subtraction module and the human detection and tracking module.

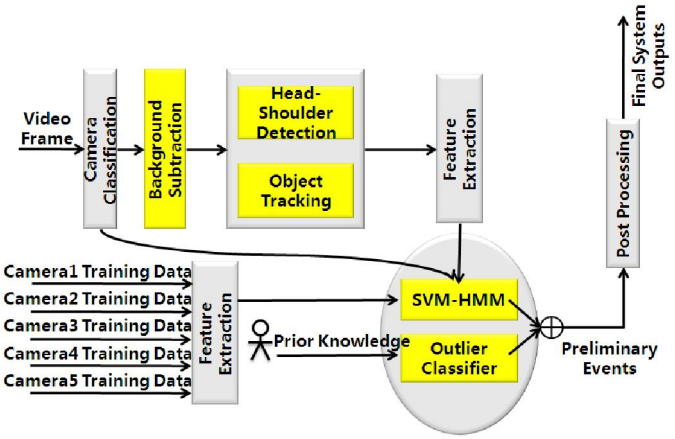

Fig.1 Diagram of our system, eSur

3. Background Subtraction

In our framework, background subtraction is used to extract foreground regions to accelerate the head-shoulder detection and the tracking process. At the same time, the detection and tracking false alarms are decreased effectively.

We proposed a selective eigenbackground method, which is a reformation of the method we used last year. In the training stage, the dimensionality of the training samples is reduced to build a Gaussian model G_m. Then those training samples containing fewer foregrounds are selected to compute the initial eigenbackgrounds according to their similarities to the Gaussian model.

In the subtraction stage, the dimensionality of the input frame vector is reduced to update the Gaussian model G_m in a running average style as in GMM[1]. If the similarity of the frame to the Gaussian model is sufficiently high, incremental PCA is performed to update the eigenbackgrounds. Then the most descriptive eigenbackground is selected for each pixel to reconstruct the background, according to the minimum absolute value of the eigenbackground element. This process is formulated in Equ. (1) – (3), where B(i) is the reconstructed background value of the ith pixel, ψ_{ki} is the reconstructed background frame, u_{ki} is the selected eigenbackground for the ith pixel to reconstruct the background, x is the input frame vector and $u_j(i)$ is the ith element of the jth eigenbackground.

$$B(i) = \psi_{ki}(i) \quad (1)$$

$$\psi_{ki} = u_{ki}u_{ki}^{T}x \quad (2)$$

$$u_{ki} = \min_j\{|u_j(i)|\} \quad (3)$$

At last, adaptively thresholding is applied to the absolute difference image between the input frame and the reconstructed background image to get the foreground mask image.

4. Detection and Tracking

4.1 Head-Shoulder detection

Pedestrian Detection is an important step in this system. As there are many occlusions in the TRECVid corpus, parts or

even the whole body of the pedestrians are frequently unseen. For this reason, we apply head-shoulder detection instead of human body detection.

In [2], Dalal and Triggs proved that Histograms of Oriented Gradients are powerful for pedestrian detection. In order to speed up, Zhu et al. [3] combined the cascaded rejection approach with HOG feature. They used AdaBoost to select the best features and constructed the rejection-based cascade.

In our system, we use HOG feature to represent head-shoulder samples, piece-wise function to construct weak classifiers, and apply Multiple Pose Learning-based RealBoost for classifier learning. Multiple Pose Learning [4] is used to deal with large intra-class variety within the pedestrian's samples of the TRECVid corpus. The framework is presented in Fig. 2.

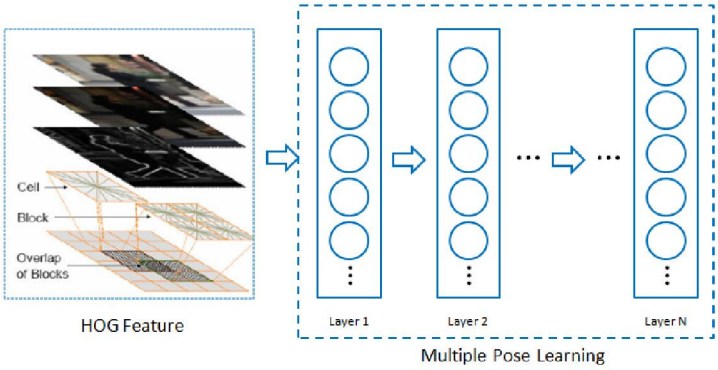

Fig.2 Detection module

The Multiple Pose Learning-based boosting used in this work is described as below.

Given n samples $x_i \in X$ and n corresponding labels $y_i \in \{-1, +1\}$, we assume, however, that there are K latent variables $y_i^k \in \{-1, +1\}$ associated with each sample. Each latent variable defines membership to one of the K groups. A sample is considered positive if it belongs to at least one of these groups, which can be expressed as $y_i = \max_k \{ y_i^k \}$. Our goal is to simultaneously split the positive data into K groups and train K classifiers $H^1, H^2, \ldots, H^K$, one per group, so that $\max_k(H^k(x_i)) = y_i$. The algorithm is summarized as below:

FOR t= 1 TO DO
  FOR k=1 TO K DO
    Compute weights $\omega_i^k = -\frac{\partial \varsigma}{\partial H^k(x_i)}$
    Train the best weak classifier with the current weights
      $h_t^k = \arg\min_h \sum_i 1(h(x_i) \neq y_i) |\omega_i^k|$
    Find α_t^k via line search to minimize cost
      $\alpha_t^k = \arg\min_\alpha \varsigma(\ldots, H^k(x) + \alpha h_t^k, \ldots)$
    Update strong classifier
      $H^k(x) \leftarrow H^k(x) + \alpha_t^k h_t^k$
  END FOR
END FOR

Fig.3 Flowchart of Multiple Pose Learning algorithm.

Some other cues are used for making the detection process more efficient. With the coarse foreground regions extracted by the background subtraction module, candidate sub-windows with sparse foreground can be neglected immediately. We can also estimate the reasonable sub-window size of head-shoulder appeared in all positions for each scene. In addition, regions those have low possibility of events are pruned in the searching process.

In practice, we labeled about 5000 head-shoulders as positive training samples, and collected hundreds of images without head-shoulders as the source to extract negative training samples.

4.2 Tracking

In the TRECVid corpus, target appearance always changes significantly. The same as last year, we use an adaptive Online

Boosting framework for tracking process as described by Helmut Grabner [5].

In camera 3 and 5, the head-shoulder of pedestrians are mostly small and blur, so we extend the head-shoulder detection result proportionally down to use the whole body for tracking instead. Another method is applied to deal with drifting. Dominant color similarity between corresponding object in two frames give a score to evaluate the matching. And Online Boosting tracking also provides a matching score. We combine these two scores to get the final tracked position of an object in each next frame.

5. Pair-Wise Events detection

To detect the pair-wise events in this year's SED task, the interactive events, such as PeopleMeet, PeopleSplitUp, and Embrace, are considered as a time-variant holistic pattern, and proper sequential model and structural classifier are introduced to serve the detection task.

It is comprehensible that the discriminative patterns for these three events in video sequences are inherently time sequential. However, most pervious activity recognition methods did not handle this properly with only modeling the patterns in single frames or simply concatenating them together. In our solution, the event is considered as a whole sequence and described by the stochastic sequential model and classified using support vector machines. Specifically, we employ the Markov Support Vector Machine proposed in [6]. This method handles dependencies between neighboring frames using Viterbi-like decoding and the learning procedure is based on a maximum margin criterion. With the sequential learning method, the temporal correlations between different stages of the event are properly considered, and decisions based on integrated event sequences are reliable and semantically reasonable.

As shown in Fig.4, features are extracted based on the motion trajectories generated by human detecting and tracking module mentioned in previous sections. According to the locations of every person in a frame, we calculate the absolute velocity, the acceleration, the distance between each pair of people and the angular separation of moving directions as the raw features. Then the extracted raw features from the same video clips (ground truth event samples for training and test samples for detecting) are transformed to structural sequence feature. Some statistics of raw features are also included into the reformed features to explicitly employ the information of the temporal dependencies over adjacent frames.

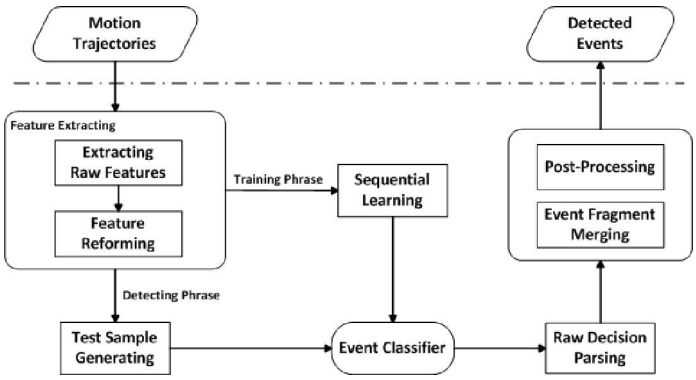

Fig.4 Flowchart of sequential learning based event detection

With the structural features, an appropriate implementation of Hidden Markov Support Vector Machine, SVM-HMM [7], is applied to train events classifiers and make decisions. It learns a hidden Markov model from training samples for each event category and makes sequence decisions for testing samples. As the raw decision is a sequence of binary decisions for each frame in a testing sample, we need to parse it into a single decision for the testing sample with the strategy like voting. As the detection task is actually transformed to a classification problem by using sliding window method to generate testing samples, the original results would be fragmental. So in the post-processing phrase, we merge the preliminary detections and introduce some prior knowledge based rules to filter out incredible detections. These rules are usually empirical restrictions such as a distance threshold between persons before "PeopleSplitUp" or after "PeopleMeet".

6. Experiment and results

Our team submitted four versions of results, which are obtained by using different human detection, tracking and events detection modules.

(a) (b (c)

Fig.5 Background subtraction results. (a) video frame (b)result with classic eigenbackground (c)result with proposed method

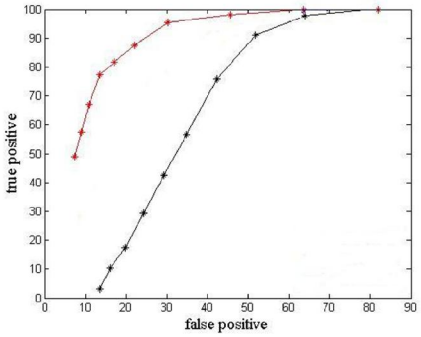

Fig.6 ROC analysis. Black line: classic eigenbackground; Red line: proposed method

Figure. 5 and 6 give the comparison results between the classic eigenbackground method and our proposed method for background subtraction. It can be observed the false alarms and the miss detections are significantly lowered by our selective eigenbackground method.

Table 1 Head-shoulder detection results of this year and last year

Camera1	**Recall**	**Precision**	**F-score**	**Camera2**	**Recall**	**Precision**	**F-score**
Last Year	0.335	0.888	0.4734	Last Year	0.243	0.816	0.3745
This Year	0.539	0.796	0.6429	This Year	0.560	0.773	0.6495
Camera3	**Recall**	**Precision**	**F-score**	**Camera5**	**Recall**	**Precision**	**F-score**
Last Year	0.305	0.728	0.4299	Last Year	0.385	0.662	0.4869
This Year	0.429	0.667	0.5222	This Year	0.468	0.757	0.5783

Table 2 Tracking results of this year and last year

Camera1	**MOTA**	**MOTP**	**Miss**	**FA**	**ID Switch**
Last Year	0.09	0.55	0.571	0.322	0.017
This Year	0.321	0.591	0.51	0.134	0.035
Camera3	**MOTA**	**MOTP**	**Miss**	**FA**	**ID Switch**
Last Year	-0.152	0.552	0.632	0.505	0.016
This Year	0.022	0.571	0.652	0.293	0.033
Camera5	**MOTA**	**MOTP**	**Miss**	**FA**	**ID Switch**
Last Year	-0.866	0.587	0.498	1.339	0.029
This Year	-0.002	0.602	0.537	0.44	0.025

Table 1 and 2 show the comparison detection and tracking results between the best outputs of our system this year and those of last year. It can be seen from the tables that detection result is improved greatly in recall with low or no decrease in the precision. Here we introduce Multiple Object Tracking Accuracy (MOTA) and Multiple Object Tracking Precision (MOTP)[8], metrics used in PETS 2009, to evaluate overall performance. These ID switches used in MOTA are calculated from the number of identity mismatches in a frame, from the mapped objects in its preceding frame. The MOTP is calculated from the spatiotemporal overlap between the ground truth tracks and the algorithm's output tracks. Conclusion can be drawn from table 2 that our performance is improved greatly.

Table 3 shows the comparison results between the best outputs of our system this year and those of last year. It can be seen from the table that our eSur system is greatly improved by detecting more correct events. The number of correctly detected PeopleMeet and PeopleSplitUp events is two times more than last year and that of Embrace are raised dramatically. Meanwhile, the false alarms do not rise too much and event dramatically decreased for PeopleSplitUp. This year we did not use any prior knowledge like last year, so it is believed that when prior knowledge is used, the performance can be further improved. It should be noticed that for the events PeopleSplitUp and Embrace, the NDCRs last year of our system are higher than 1.0 but we lowered them below zero this year, which verifies the effectiveness of our improvement methods.

According to the comparative results in the TRECVid SED formal evaluation, our experimental results are promising this year, especially for the events PeopleMeet and PeopleSplitUp where the NDCRs are the lowest among all the participants.

Table 3 Comparsion results between the best ourputs of eSur this year and last year

PeopleMeet	**#Ref**	**#Sys**	**#CorDet**	**#FA**	**#Miss**	**#F-score**	**Act.DCR**
eSur last year	449	125	7	118	442	0.9031	1.023
eSur this year	449	156	12	144	437	0.8570	1.02
PeopleSplitUp	**#Ref**	**#Sys**	**#CorDet**	**#FA**	**#Miss**	**#F-score**	**Act.DCR**
eSur last year	187	198	7	191	180	0.5864	1.025
eSur this year	187	167	16	136	171	0.6505	0.959
Embrace	**#Ref**	**#Sys**	**#CorDet**	**#FA**	**#Miss**	**#F-score**	**Act.DCR**
eSur last year	175	80	1	79	174	0.7932	1.02
eSur this year	175	925	6	71	169	0.8024	0.989

7. Conclusion

This year we improved our system significantly in background subtraction where selective eigenbackground method is proposed, head-shoulder detection where multi-pose learning based method is employed and event detection where SVM-HMM classifier is used for pair-wise events detection and a distance-based outlier detection method is employed to the single-actor event detection. The promising results of our system this year verify the effectiveness of these improvements. However, we believe there are still large improvement spaces for our system in exploring more effective and descriptive event models.

Reference

[1] Stauffer C, Grimson W. E. L. Adaptive background mixture models for real-time tracking. in Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149). IEEE Comput. Soc. Part Vol. 2, 1999.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. Coference on Computer Vision and Pattern Recognition (CVPR), 2005.

[3] Qiang Zhu, Mei-Chen Yeh, Kwang-Ting Cheng, Shai Avidan: Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. CVPR (2) 2006: 1491-1498

[4] Boris Babenko, Piotr Dollar, Zhuowen Tu, Serge Belongie. Simultaneous Learning and Alignment: Multi-Instance and Multi-Pose Learning, Euro. Conference of Computer Vision, 2008.

[5] H. Grabner, T.T. Nguyen, B. Gruber, H. Bischof. On-line boosting-based car detection from arial images. ISPRS Journal of Photogrammetry & Remote Sencing, 2007,63(3),pp.382-396.

[6] Yasemin Altun, Ioannis Tsochantaridis and Thomas Hofmann. Hidden Markov Support Vector Machines. International Conference on Machine Learning (ICML), 2003.

[7] Thorsten Joachims. SVMHMM tool package available at http://www.cs.cornell.edu/People/tj/svm_light/svm_hmm.html

[8] A. Ellis, A. Shahrokni, J. Ferryman. Overall Evaluation of the PETS 2009 Results. Performance Evaluation of Tracking and Surveillance Online Proceedings, pp: 117-123, 2009.