Event detection: BJTU-SED at Trecvid 2012

Qiang Zhang, Wanru Xu, Qiang Zhai, Jie Yang, Zhenjiang Miao
Institute of Information Science, Beijing Jiaotong University
{11112066, 11112063,11120385,12120354,zjmiao}@bjtu.edu.cn

Abstract:

In trecvid 2012, our team takes part in 2 event detection competition including embrace and pointing. We build two systems to recognize these events separately. For embracing, we use a probability accumulated method. For pointing, we use texture and silhouette. Different from the former works, the two systems are interactive systems and feedback strategy is used in the detection of events. In the experiment, both two actions of our work obtain good performance.

1. Introduction

Human action recognition is one of the most challenging problems in computer vision. The focus of this problem is mainly reliability and effectiveness. However, in Trecvid dataset, it is more challenging than any other datasets, because the number of people in the scene and occlusion. Until now, many approaches have been presented for human action recognition.

One of the main approaches of recognition is dynamic models. Yamato et al. [1] used the Hidden Markov Models (HMM) as recognition model for human action recognition. Laxton et al. [2] used a Dynamic Bayesian Network to recognize human action.

Another main approach of recognition is spatio-temporal template. Bobick and Davis [3] introduced Motion-Energy-Image (MEI) and Motion-History-Image (MHI) templates for recognizing different motions. From then on, spatio-temporal templates were made famous on human action recognition. Efros et al. [4] used a motion descriptor based on optical flow measurements in a spatio-temporal volume to represent actions and used nearest-neighbor to classify actions. Blank et al. [5] defined actions as space-time shapes, and used Poisson distribution to represent the details of such shapes. Jhuang et al. [6] applied biological model of motion processing

for action recognition using optical flow and space-time gradient feature.

In recent years, space-time interest points feature and "bag of words" model are widely used in human action recognition studies. Laptev et al. [7] first introduced the notion of "space-time interest points". Piotr Dollar et al. [8] used 2-D Gauss filter and 1-D Gabor filter to extract space-time interest points for human action recognition. Popular topic models include pLSA [9], LDA [10]. Juan Carlos Niebles et al. [11] extracted space-time interest points feature and they perform unsupervised learning of action categories using pLSA model and LDA model separately. Yang Wang and Greg Mori [12] used optical flow method to extract motion feature and used latent topic models to do recognition. However, extracting space-time interest points need much computation time and topic models ignore the spatial and temporal information.

In trecvid 2012, our team takes part in 2 event detection competition including embrace and pointing. We build two systems to recognize these events separately. For embracing, we use a probability accumulated method. For pointing, we use texture and silhouette. Different from the former works, the two systems are interactive systems and feedback strategy is used in the detection of events.

The rest parts of this paper are organized as follow: the Section 2 introduces of two event detection systems. The conclusions are given are given in section 3.

2. Our approaches

Before event detection, we must detect and track people, In this study, we use trecvid dataset to train a HOG-SVM [13] head detection and people detection model.

2.1 Pointing

The flowchart of pointing event recognition is shown in figure.1. Through observation on the referenced pointing events in training videos, we found it is quite hard to find a specific feature to characterize this event as it takes place variously. But we found statistically a large percent of people pointing in the training dataset keep a relatively unified pointing pose. Then we define acting pointing event as people

keeping in a pointing pose for certain period of time. The recognition of pointing event turns into recognition of pointing pose.

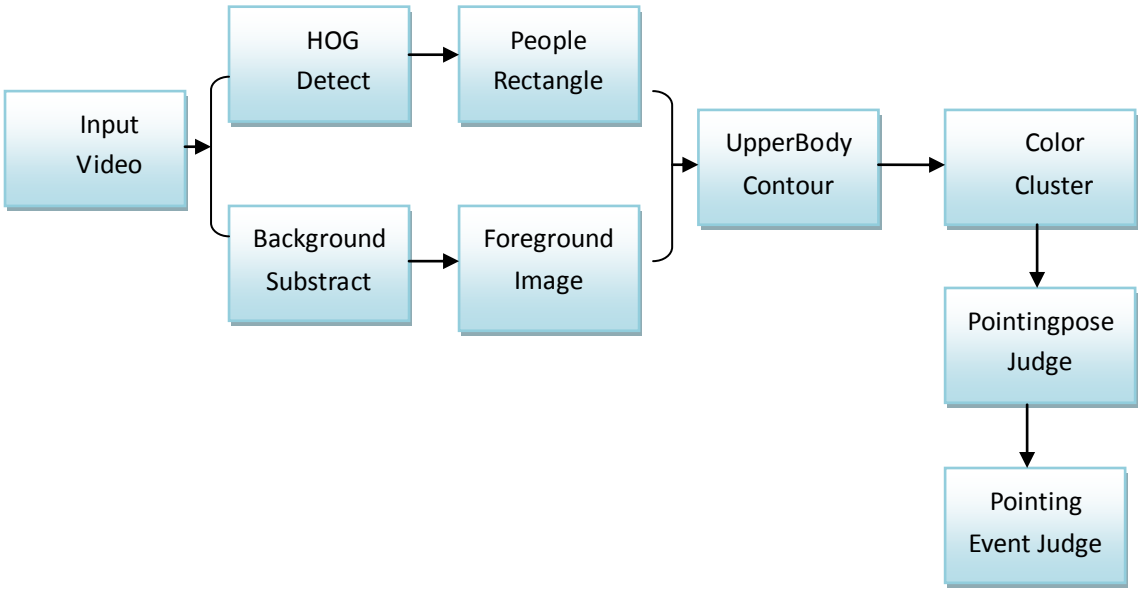

Figure.1 Flowchart of pointing event recognition

We define pointing pose as one of human's arms lift up in one side of his body. In order to recognize pointing pose, it is necessary to get body contour. Moreover, the pointing pose can be described using the relationship between arms and torso an d has little to do with lower body, it is enough that we just get the upper body to recognize pointing pose rather than the whole body.

It is really a great challenge to get body contour that we have read large number papers about body contour segmentation but none of them could give an effective solution to extract human body in such a crowd and complex video data as our airport video set. Then we tried to combine classical methods and use some new method to get a relatively precise upper body contour.

Given a rectangular box for each person by head detection and people detection using HOG (as shown in figure 2), we could know the location and size of people occurring in the frame. With background subtract method, we could get the foreground image. Indeed in the condition that the environment is less crowd, the foreground image can be regarded as the body contour. We use K- means color cluster method and floodfill method to get the main people's body contour.

Figure 2 . Workflow of Upper- Body contour extraction

After we get the main body contour, it becomes easy to judge whether the people is pointing or not. We use Projection Histogram method to do the Pointing- pose Judge. Next, we judge every people detected in pointing pose or not. We consider a people having pointing event from when he is in pointing pose until when he is not in pointing pose.

According to the interaction, we get our interactive results based on the retro one. we use the feedback of our retro-system, and delete the and unqualified results, the second step we add some qualified results that we get from watching the video rather than our retro-system, through those two steps we get our interaction result.

2.2 Embrace

Different from the former works, now it needs an interactive system to detect event. So feedback strategy is used in the detection of events. In our work, a core feedback [14] structure is proposed based on operation on different levels of detail for the event detection, which is relate to the different levels of features which we extracted. In short, feedback is interpreted as a coarse-to-fine strategy .Through the feedback, we reduce the computation complexity and get a progressed result.

For action embracing, the system contains three parts. The first is to find the space that it maybe occurs embracing action; the second is to determine the possible temporal interval , when the action happened and finished; last one is the feedback. The overall flow of this recognition is as shown in figure 3.

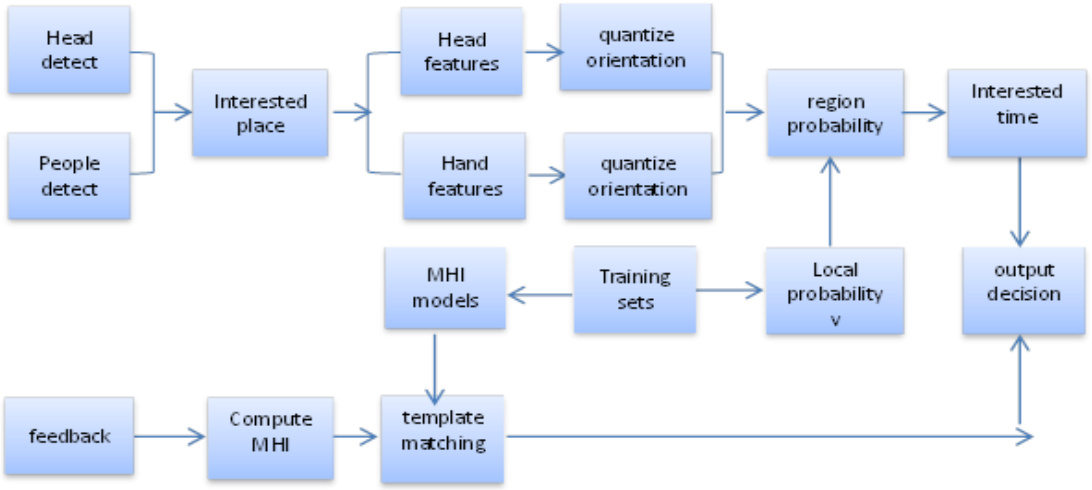

Figure 3

In part one, head detection and people detection are both used to get the interested place. The trecvid dataset is challenging, because it involves multiple co-occurring and complex human activities. Because of the occlusion, only use people detection it can become lacking check easily, while only use head detection it can get more error check. So combine them can get better performs. For action embracing, it needs two persons to interaction each other to complete this action. So when two persons are close to each other, we assume that these two persons maybe begin embracing and the place is our interested place.

After known the interested place, concerning about when the action happen and finished is also necessary. We do this job in part two. The first step is to compute optical flow features in the obtained the interested place. The question is the background is too clutter and the optical flow features we obtained can't precisely describe the action. We notice that the action embracing is mainly the action focus on arms and hands. So compute the optical flow features only in complexion areas. See figure 4, it actually reduces interference.

Figure 4

And then we quantize the orientation of each feature into eight bins as the hand

features. For the head features, relative movement of the two person's head is extracted. The recognition area is divided into nine blocks. Each block has a local probability which we train in the positive training sets. Based on the local probability in figure 5, we start to compute the region probability. For computing embracing starting, we only use the hand features whose orientations belong to [π/4, π/2, 3π/4] and the head features whose orientations is get close to. For computing embracing ending, we only use the hand features whose orientations belong to [5π/4, 3π/2, 7π/4] and the head features whose orientations is far away from. The region probability can be represented by equation (1), it is the probability of embracing start. p_i is the local probability, $N_{i_{hand\ up}}$ is the hand feature number whose orientations belong to [π/4, π/2, 3π/4] in the i_{th} block.

$$P_{region\ start} = a\frac{N_{head\ close}}{N_{head\ all}} + (1-a)\frac{\sum_{i=1}^{9} p_i \times N_{i_{hand\ up}}}{N_{hand\ all}} \quad (1)$$

1241	7908	1357
2133	10719	2077
1377	0	1335

0	3/9	0
1/9	4/9	1/9
0	0	0

Figure 5

The last part is the feedback. When we get the action start and finish time, we already have the output result. But the result is too coarse and at sometimes it is unlikelihood. It assumes that the event is accepted if its value is above 0.7, rejected if its value is below 0.3 and declared unknown in other case. We need feedback and do more work like extract higher level of features in the last case. In our work, Four MHI models of embrace (left, right, front, back) which are show in figure 6 is used to do farther classify by template matching. Also based on the rule, event is accepted or rejected. If there are still some uncertain, we can watch them by ourselves.

Figure 6

3. Conclusions

In trecvid 2012, we takes part in 2 event detection competition including embrace and pointing. We build two interactive systems to recognize these events separately by feedback strategy. For embracing, we use a probability accumulated method. For pointing, we use texture and silhouette. In the experiment, both two actions of our work obtain good performance.

Reference

[1] J.Yamato, J.Ohya, and K.Ishii. Recognizing human action in time-sequential images using hidden Markov model.
In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (Champaign
IL, June 1992). CVPR '92, 379 -385.
Figure 5. Flow chart
[2] B. Laxton, J.Lim, and D. Kriegman. Leveraging temporal, contextual and ordering constraints for recognizing
complex activities in video. In Proceedings of the IEEE International Conference on Computer Vision and Pattern
Recognition (Minneapolis MN,, June 2007). CVPR'07, 1 -8.
[3] A.F.Bobick and J.W.Davis. The recognition of human movement using temporal templates. IEEE Transactions on
Pattern Analysis and Machine Intelligence . 23, 3 (March 2001) 257-267.
[4] A.A.Efros, A.C.Berg, G.Mori, and J.Malik. Recognizing action at a distance. In Proceedings of the IEEE 9th
International Conference on Computer Vision (Nice France, 2003). ICCV'03, Vol.2, 726-733.
[5] M.Blank, L.Gorelick, E.Shechtman, M.Irani, and R.Basri. Actions as space-time shapes. In Proceedings of the

IEEE 10th International Conference on Computer Vision (Beijing, 2005). ICCV'05, Vol.2, 1395-1402.
[6] H.Jhuang, T.Serre, L.Wolf, and T.Poggio. A biologically inspired system for action recognition. In Proceedings of
the IEEE 11th International Conference on Computer Vision (Rio de Janeiro, October 2007). ICCV'07, 1 -8.
[7] I.Laptev and T.Lindeberg. Space-time interest points. In Proceedings of the IEEE 9th International Conference on
Computer Vision (Nice France, 432 -439, 2003). ICCV'03, Vol.1, 432-439.
[8] P.Dollar, V.Rabaud, G.Cottrell, and S.Belongie. Behavior recognition via sparse spatio-temporal features. In
Proceedings of the IEEE workshop on Visual Surveillance and Performance Evaluat ion of Tracking and
Surveillance. (October 2005). VS-PETS'05, 65-72.
[9] T.Hofmann. Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR
conference on research and development in information retrieval (California US, August 1999). ACM Press, New
York, NY, 50-57,
[10] D.M.Blei, A.Y.Ng, and M.I.Jordan. Latent dirichlet allocation. Journal of Machine Learning Research. 3 (2003)
993-1022.
[11] Juan Carlos Niebles, Hongcheng Wang, and Fei -Fei Li. Unsupervised learning of human action categories using
spatial -temporal words. International Journal of Computer Vision. 79, 3 (2008) 299 -318.
[12] Yang Wang and G.Mori. Human action recognition by semilatent topic models. IEEE Transactions on Pattern
Analysis and Machine Intelligence. 31, 10 (Oct. 2009) 1762-1774.
[13] N.Dalal and B.triggs. Histograms of oriented gradients for human detection, IEEE Conference on Computer
Vision and Pattern Recognition, (2005),1-8.
[14] J.C.SanMiguel J.M.Martinez. Use of feedback strategies in the detection of events for video surveillance. IETComputerVision