Event detection: BJTU-SED at Trecvid 2011

Yuan Shen, Qiang Zhang, Xu Zhang, Liang Liang, Zhenjiang Miao
Institute of Information Science, Beijing Jiaotong University
{08112074, 10120411, 10120412, 10120392, zjmiao}@bjtu.edu.cn

Abstract:

In trecvid 2011, our team takes part in 2 event detection competition including pointing and cell to ear. We build two systems to recognize these events separately. For pointing, we extract texture and silhouette to recognize this action. And for cell to ear, we use orientation of optical flow and SVM classifier to finish this work.

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 2011, our team takes part in 2 event detection competition including pointing and cell to ear. We build two systems to recognize these events separately. For pointing, we extract texture and silhouette to recognize this action. And for cell to ear, we use orientation of optical flow and SVM classifier to finish this work.

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

2. Our approaches

Before event detection, we must detect and track human. In this study, we use trecvid dataset to train a HOG-SVM [13] human detection model and use mean-shift [14] to track people.

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.

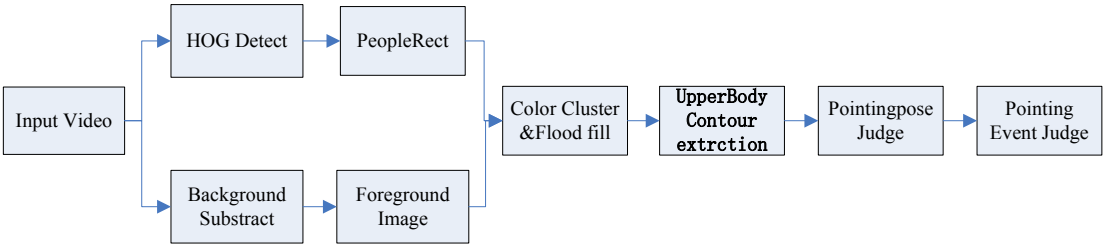

Figure 1. Flowchart of pointing event recognition

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 and 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.

Figure 2. Workflow of Upper-Body contour extraction

Given a rectangular box for each person by human 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. In the crowd airport environment we have to do more to extract the contour as the rectangle is with more than one people inside that the foreground image is in rather a mess. We use K-means color cluster method and flood-fill method to get the main people's body contour .Experiment results are shown in figure3.

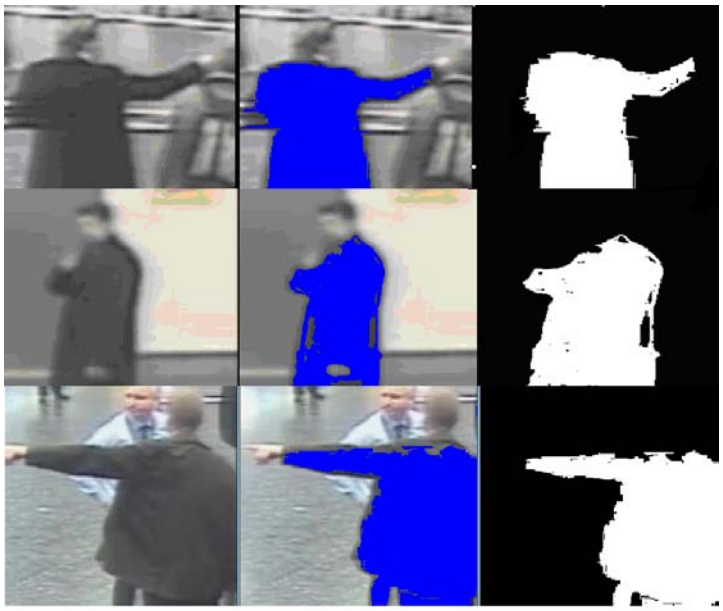

Figure 3. Result of Upper-body Extraction

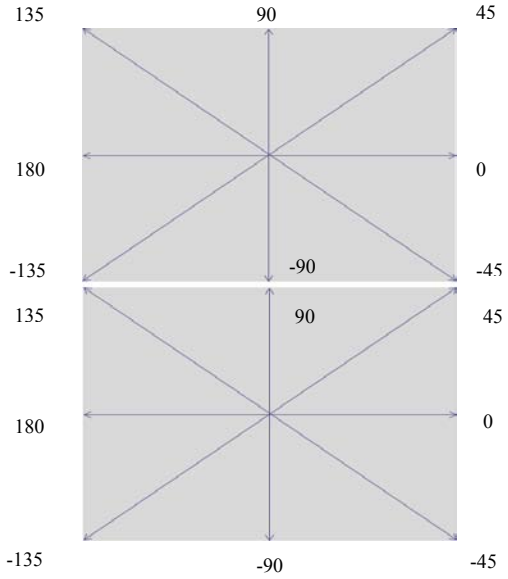

Figure 4. 16 dimensions vector

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.

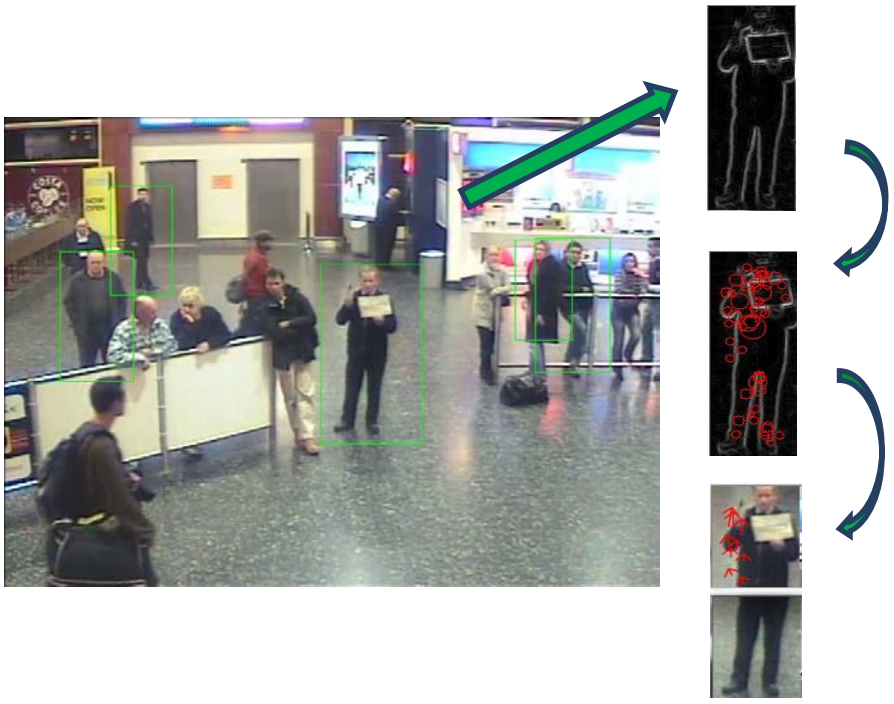

Figure 5. Flow chart

2.2 CellToEar

It's difficult to detect the mobile phone, so we try to detect the motion of the hands and the arms. We have tried the traditional Lucas- Kanade algorithm to calculate optical flow. We choose motion vector to represent the action.

First of all , we do a gamma compress transform with the image to reduce the effect of the light; There is much noise of the optical flow and the noise may influence the accuracy of the features. So we extract the surf points from the gradient image to calculate optical flow and it gets a better performance. The detected man are divided into 2 parts , the bottom part and the top part. And we cumulate the mode of 8 different directions in either part as a optical flow histogram (Figure 4). We get the maximal value of the modes as the denominator to normalize the histogram, then we get a 16 dimensions vector as a feature.

We Extract features from the positive data sets and negative data sets, and then we train a classify model with the SVM. At last we use this model to predict the features extracting from the test datasets. Figure 5 is the flow chart.

3. Conclusions

In trecvid 2011, our team takes part in 2 event detection competition including pointing and cell to ear. We build two systems to recognize these events separately. These two systems use different approaches to recognition actions. One uses texture and silhouette, the other uses SVM classifier with orientation of optical flow.

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.

[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 Evaluation 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] D.Comaniciu, V.Ramesh and P.Meer, “real-time tracking of non-rigid objects using mean shift”, IEEE Conference on Computer Vision and Pattern Recognition, (2000), 673-678.