Action Recognition Using HOG Feature in Different Resolution Video Sequences

Yuanyuan Huang, Haomiao Yang
School of Computer Science and Engineering
University of Electronic Science and Technology of China, Chengdu, 611731, People's Republic of China
Email :{yyhuang, haomyang}@uestc.edu.cn

Ping Huang
College of Mathematics and Software Science
Sichuan Normal University, Chengdu, 610068, People's Republic of China
Email:ihuangping@163.com

Abstract— **This paper summarizes the recent development of action recognition at first. Then based on Hierarchical Filtered Motion model and Nearest Neighbor classifier, we do action recognition using HOG feature in video sequences of different resolutions. Here we use KTH dataset for training and MSR action dataset II for testing. The experiment demonstrates that the new feature extraction process is effective and has better performance in the cross-dataset action recognition.**

Keywords- HOG; action recognition; Motion History Image

I. INTRODUCTION

In recently years, action recognition has become an active topic in computer vision and video understanding. It has broadly applications including biometrics security, video surveillance, human-computer interaction (HCI) [1]. Some related survey papers have appeared over the years [2]. We noticed that the feature representation and the classification are the two most important key steps in action recognition, and many new methods have been proposed [2].

However, there are many open problems in this area that have not been solved. For example, the authors in [6] successfully recognize actions in dynamic and crowded videos by the use of the action models that are trained on dataset with clean background. In our research, we noticed that the classification result on high resolution video (320x 240 spatial resolutions) is not as good as that on low resolution video (160x 120 spatial resolutions). The reason that the result on high resolution video gets worse than on low resolution video is because the mismatch of the scales between the testing data and training data. Thus we divided the feature extraction process sequences into two steps: In the first step, we do feature extraction on the downsampled video sequences. In the second step, we do feature extraction on the high resolution video sequences. For each interest point, we compute its feature descriptor, HOG [5] and HOG-MHI [6], per interest point by randomly selecting a single scale. The union of the feature vectors obtained from the two steps is used as the final feature vectors for each video sequence. The experimental results show that the new feature extraction process is better than the old process.

The paper is organized as follows: In the next section, we briefly describe some key technologies used in our action recognition framework; experiment results are summarized in Section 3 followed by discussions and conclusions in Section 4.

II. SOME KEY TECHNOLOGIES

The action recognition framework in this paper is mainly based on former work in [6-7]. So we briefly introduce several key technologies below, including the interest point detection in Motion History Image (MHI), Hierarchical Filtered Motion Model, and HOG features and its application in different resolution videos.

A. Motion History Image

The Motion History Image (MHI) [3] is constructed by successively layering selected image regions over time using a simple update rule, that is, to generate a MHI, at location (x, y) and time t, the intensity of MHI is calculated as :

$$MHI_{\tau}(x,y,t)=\begin{cases}\tau & if\ D(x,y,t)=1\\ \max(0, MHI_{\tau}(x,y,t-1)\text{-}1), & otherwise\end{cases} \quad (1)$$

where is a binary image of differences between frames and is the maximum duration of motion. Then the MHI image is scaled to a grayscale image [6].

B. Interest Point Detection in MHI using Hierarchical Filtered Motion Model

The point of interest detection usually is a key step in action recognition. One of the most successful sparse selections of interest points is STIP [4]. However, the main drawback of STIP detector is that on some low-resolution videos (such as MSR action dataset II with downsampled 160x120 spatial resolutions), it cannot generate enough interest points. This has been observed by many researchers. Although Harris corner detection is not scale invariant compared with STIP detector, it has the advantage that it can produce more interest points. Unfortunately, not all the detected points are good. Thus we use a model named HFM [6] to filter out some bad points, such as those without motions. Here MHI is used as motion mask to remove the corners in the static background. A global spatial motion smoothing filter and a local motion field filter in [6] is applied to enhance the motion. The result of point of interest detection in MHI is illustrated in fig. 1 as below.

978-0-7695-4639-1/12 $26.00 © 2012 IEEE
DOI 10.1109/CDCIEM.2012.27

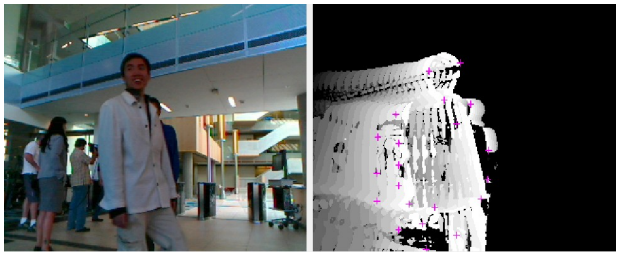

Figure 1. Point of interest detection in MHI

C. HOG and HOG-MHI feature extraction in different spatial resolution video

The Histogram of Oriented Gradients (HOG) feature descriptor [5], which is prevalent in object detection and classification, describe local object appearance and shape within an image by the distribution of intensity gradients or edge directions. The details of HOG calculation is described as below. Suppose denotes the pixel intensity at the position , and denote x and y components of the image gradient, respectively, are computed in (2).

$$\begin{cases} f_x(x,y) = I(x+1,y) - I(x-1,y) \forall x,y \\ f_y(x,y) = I(x,y+1) - I(x,y-1) \forall x,y \end{cases} \quad (2)$$

And the magnitude m(x, y) and orientation θ(x, y) are computed by (3).

$$\begin{cases} m(x,y) = \sqrt{f_x(x,y)^2 - f_y(x,y)^2} \\ \theta(x,y) = \tan^{-1}(f_y(x,y) / f_x(x,y)) \end{cases} \quad (3)$$

Then the gradient image is divided into cells. At each cell, the orientation θ(x, y) is quantized into orientation bins, weighted by its magnitude m(x, y) to make the histogram. After that block normalization is done to get the HOG features. To characterize the temporal features, we use Histograms of Oriented Gradient in Motion History Image (HOG-MHI) [6], while the spatial features are modeled by HOG in the intensity image as former work [5].

In our experiments, we choose KTH dataset, which is 160x120, for training [5-6]. The purpose of using KTH for training is to make use of existing labeled data since one general problem in machine learning is that, in practice, the testing data is always different from the training data. It's desirable if a classifier trained on one dataset also works on a different dataset. But the spatial resolution of MSR II dataset is different. We observed that the recognition rate is down if we use HOG features in 320x240 resolution videos. This is because of the mismatch of the scales between the test data and training data. So we wonder if it helps if we use two scales HOG features, that is, by combining the HOG features in 160x120 and 320x240. The number of feature points extracted on the two spatial resolution videos is similar but not necessarily equal to each other. Basically, the total number of feature vectors will roughly double compared to the feature vectors on 160x120 resolution video sequences only.

D. The Action recognition framework

Based on the above technologies, and to achieve the goal of focusing on the impact of the HOG in resolution videos, we simplify the classification process by using nearest neighbor classifier.

First, we detect 2D Harris corners as the point of interest in MHI. Then a hierarchical motion filter [5] is applied to enhance the points of motion and weaken the irrelevant and noisy points. Then, we characterize the spatial and temporal features by using HOG and HOG in MHI [6] and do feature extraction on low resolution video sequences (downsampled video sequences from MSR II dataset) and high resolution video sequences (MSR II dataset). We combine them to get the union of the feature vectors for each sequence. Finally, we choose the Nearest Neighbor classifier to get the recognition results. The action recognition process illustrated below.

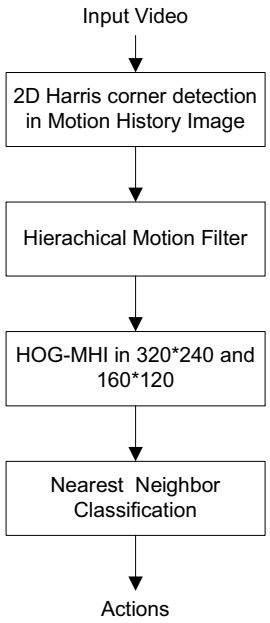

Figure 2. The framework of action recognition

III. EXPERIMENT AND PERFORMANCE EVALUATIONS

In our experiment, we choose KTH dataset for training and MSR action dataset II for testing.

A. KTH dataset

The KTH dataset contains six types of human actions: walking, jogging, running, boxing, hand waving and hand clapping, performed several times by 25 subjects in four different scenarios. Some samples are illustrated below [8].

Figure 3. Samples from KTH dataset

B. *MSR action dataset II*

The MSR action dataset II consists of 54 video sequences recorded in a crowded environment. The testing video resolution is 320x240. Each video sequence consists of multiple actions. There are three action types: hand waving, handclapping, and boxing [9].

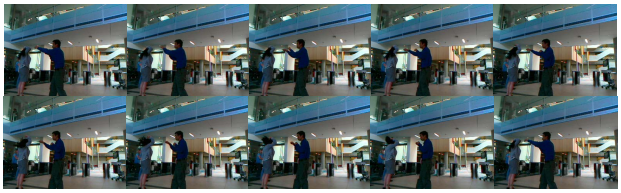

Figure 4. An action sequence in MSR action dataset II

C. *Experiment results*

After the cross-dataset action recognition process, we get the experimental results. The confusion matrix of action recognition on MSR action dataset II is illustrated in table I, table II and table III, and the action recognition rate is showed in table IV, respectively.

TABLE I. CONFUSION MATRIX OF ACTION RECOGNITION ON MSR ACTION DATASET II (HOG OF DOWNSAMPLED VIDEO OF 160X120 RESOLUTION)

	Boxing	Clapping	Waving
Boxing	43	32	6
Clapping	4	42	5
Waving	7	20	44

TABLE II. CONFUSION MATRIX OF ACTION RECOGNITION ON MSR ACTION DATASET II (HOG OF VIDEO OF 320X240 RESOLUTION)

	Boxing	Clapping	Waving
Boxing	49	20	12
Clapping	7	39	5
Waving	8	23	40

TABLE III. CONFUSION MATRIX OF ACTION RECOGNITION ON MSR ACTION DATASET II (COMBINE TWO RESOLUTION VIDEO FEATURE TOGETHER)

	Boxing	Clapping	Waving
Boxing	48	21	12
Clapping	5	40	6
Waving	8	21	42

TABLE IV. RECOGNITION RATE

Feature descriptor	Accuracy
HOG in 160 x 120	63.55%
HOG in 320x240	63.05%
HOG in 320x240 and 160 x 120	64.04%

The experimental results demonstrate that the performance could be improved by combine multi-resolution HOG features. We also notice that the impact of HOG feature descriptor is different for different actions.

IV. CONCLUSIONS AND FUTURE WORK

In this paper, we use a cross-dataset action recognition framework to investigate the HOG feature and the impact of HOG feature representation in different spatial resolution videos for action recognition. The experimental results indicate that it is a new approach to improve the effect of cross dataset action recognition. Besides, the impact of HOG feature is different for different actions. Since the research is very preliminary, we will continue to study the HOG feature descriptor and its impact on action recognition results when dealing with different resolution video sequences.

ACKNOWLEDGEMENTS

This work was supported by the National High Technology Research and Development Program of China (No: 2007AA01Z423), the National Natural Science Foundation of China (No: 60703113), the Fundamental Research Funds for the Central Universities (No: ZYGX2009J056), Sichuan Committee of Economics Project (No: 2008CD00053) and CSC scholarship.

REFERENCES

[1] Pavan K. Turaga, Rama Chellappa, V. S. Subrahmanian, Octavian Udrea, "Machine Recognition of Human Activities: A Survey," IEEE Transactions on Circuits and Systems for Video Technology, Vol.18, No.11, pp.71-86, 2008.

[2] Ronald Poppe, "A survey on vision-based human action recognition," Image and Vision Computing, Vol.28,No.6 pp.976-990,2010.

[3] Bobick A. F, Davis J. W. "The recognition of human movement using temporal templates," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.23, No.3, pp.257-267, 2001.

[4] I. Laptev, T. Lindeberg, "Space-Time Interest Points," International Conference on Computer Vision, pp.432-439, 2003.

[5] N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 886-893 2005.

[6] Y. Li. Tian, L. L. Cao, Z. C. Liu, Z. Y. Zhang, “Hierarchical Filtered Motion Field for Action Recognition in Crowded Videos,” IEEE Transactions on Systems, Man, and Cybernetic Part C, accepted.

[7] L. L. Cao, Z. C. Liu, Thomas S. Huang, “Cross-dataset Action Recognition,” IEEE Conference on Computer Vision and Pattern Recognition, pp.1998-2005, 2010.

[8] Christian Schuldt, Ivan Laptev, Barbara Caputo, “Recognizing Human Actions: A Local SVM Approach,” IEEE Conference on Computer Vision and Pattern Recognition, pp.32-36, 2004.

[9] J. S. Yuan, Z. C. Liu, Y. Wu, “Discriminative Subvolume Search for Efficient Action Detection,” IEEE Conference on Computer Vision and Pattern Recognition, pp.2442-2449, 2009.