Human Action Recognition in Video using Non-Hierarchical Approach

Nita Jadav[1] Khushali Raval[2]

[1]U.G Student [2]Assistant Professor

[1,2]Shantilal Shah Engineering College, Bhavnagar

Abstract— The activity recognition and understanding in video by computer is defined as gaining knowledge about what is going on in video by analyzing the frames of video. The analysis of such kind by computer can be used for human – computer interaction, video monitoring, Identification based on behaviors and patients monitoring system. The main challenges faced for activity recognition are scale variation, occlusion, pause variation, anthropometric variation, execution rate. There are different methods available for each step of recognizing human activity coping with different challenges. The complex activity recognition methods also rely on the performance of primitive action recognition. For recognizing primitive action, the non-hierarchical approach is used. This approach represents the video as Bag of Words (BoW) model. We identify the limitations of this model and propose good practices for enhancing performance of the model.

Key words: Activity Recognition, STIP, Spatio-Temporal Structure, BoF

I. INTRODUCTION

Recently the researchers of computer vision and image processing are leaning towards the composing of artificial intelligence in computer vision system. The ultimate goal is to make the computer understand and label the activity performed in video sequences. The human activity recognition has very wide area of applications. Many researchers proposed efficient methods for recognizing action. Though there are many challenges that are not yet worked out.

Human activity basically can be categorized into two parts as: 1) Simple action 2) Complex activity [1]. The simple actions include basic actions performed by single human. It can include gestures like hand waving and periodic activity like hand clapping. The complex activity is composed of primitive actions. The process of analyzing human action requires several numbers of steps. First of all, the human body part which is responsible to perform the human action is detected then the motion pattern is described and then video representation is generated to classify the human action. For every step, there are various methods available giving different results and constraints.

The non hierarchical approach for human action recognition concerns with recognizing primitive actions performed by human. In this approach the action is recognized step by step. The steps are entitled as shown in figure 1. The first step is to extract low level features from video sequences. As we know the pixel of image carries lot information. Videos are nothing but the ordered sequence of images. Hence the low level features of videos can be extracted by the space time point. Among the whole set of space time points important points that provide fruitful information about video can be identified and extracted in this step. The selected points are also called as spatiotemporal interest points (STIP) [2]. The second step of the framework is vocabulary building and feature encoding. The detected low level features are described and grouped together to make video representation that is understandable and used as training data for classifier.

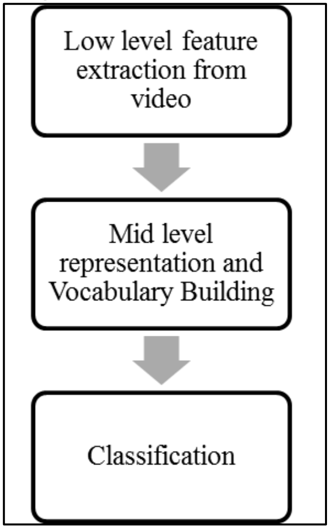
Low level feature extraction from video
Mid level representation and Vocabulary Building
Classification

Fig. 1: Steps for human action recognition approach

The visual vocabulary represents the action in terms of visual words from video. The feature encoding technique makes the representation of low level features in the form such that it can be classified into features of various action categories. The final step is to classify the feature representation. It uses various classifiers such as SVM [3].

The low level video features are represented as Bag of Words (BoW) approach. This approach has several drawbacks. It considers only frequency of words. Hence the information concerning temporal ordering and spatial structure is lost in BoW model [4]. To address this issue, we propose a method that represents the features maintaining the spatial relation.

II. RELATED WORK

The human action can be defined by the features it posses in video in terms of movements of body parts. The features can be categorized as global features and local features. Methods used in [5] describe the global features for action recognition. Global features are usually in the form of silhouettes [6], temporal pose let [7]. These features are very difficult to extract from realistic video as they need to be highly detailed. On the other hand local features based approaches are more successful for recognizing simple human action. According to survey these approaches are efficient for coping with camera motion and low resolution. The techniques based on these approaches do not require foreground extraction or tracking of human body parts or object. The approaches described in [8-10] are based on the local spatiotemporal interest points. The spatiotemporal points include points that are in corner, isolated forms where the intensity value is minimum or maximum.

The spatiotemporal interest points can be sampled from video sequence in two ways, either using sparse sampling or dense sampling. There are various methods like Harris3D [2], Cuboids detector [11], Hessian [12] detect spatiotemporal interest points sparsely. Dense sampling method is based on the idea of extracting features of video by dividing it into blocks at regular positions with consideration of space and time scales. Recently the trajectory approach is proposed by Wang et al [13]. Trajectories are formed by tracking space point over temporal domain. Dense trajectory provides comprehensive information about the video; they are redundant and low level representation of videos [14]. Dense trajectories with their descriptors prove to be very efficient method for recognizing human action. There are issues related to this representation as they are computationally very complex and have drifting problem [15]. The SIFT trajectories also proposed that is based on tracking key points instead interest points [16]. This method is not suitable for fast but small movements.

All these features are encoded by the Bag of Features (BoF) model using a vector quantization approach [17]. Fengsi et al. states that the BoW model ignores the spatial relation and temporal relation of the features [4]. Liu et al. [18] states that meaningful grouping of vision features within the original Bag of Words enhance the classification performance. Hence various methods were proposed to form meaningful grouping of words. Kardaris et al. proposed a method that introduces temporal structure between the visual words [19]. This method is based on the state of the art dense trajectories [14]. Dense trajectories are efficient but computationally very complex. Also they suffer from the drifting problem. The temporal relationships also introduced between the words using Allen’s temporal relations as proposed by Cheng et al [20]. This method is also based on the group of dense trajectories.

The most of the recent approaches try to make the representation of features such that it maintains its spatial relation and temporal relation. Human actions have specific spatial structure and temporal ordering that identifies them uniquely. Hence it is necessary to carry the information of them to recognize action efficiently reducing cost.

The following table compares the most effective methods for low level feature extraction.

	STIP	Dense Trajectory	SIFT Trajectory
Accuracy	moderate	high	high
Complexity	low	high	high
Robustness to scale variation	Satisfactory	Satisfactory	high
effectiveness for simple and fast motion	high	high	moderate
Drifting problem	NA	high	moderate

Table 1: Comparison of Low Level Feature Representation

III. Methodology

The proposed methodology is based on the space time interest point and its descriptors [2]. From the video key frames are extracted and space time interest points (STIP) are detected by analyzing the key frames.

The STIPs are those points which have non constant motion and are accelerating over space time. It is the extension of the 2-D Harris corners method [2].

A. *Interest Point Detection*

Interest points are basically local maxima of Harris corner function. The second moment matrix summarizes the gradient information of video which can be defined as, the Gaussian kernel.

$$\mu = \begin{bmatrix} \mu_x\mu_x & \mu_y\mu_x & \mu_t\mu_x \\ \mu_x\mu_y & \mu_y\mu_y & \mu_t\mu_y \\ \mu_x\mu_t & \mu_y\mu_t & \mu_t\mu_t \end{bmatrix} \qquad (3.1)$$

Where $\mu_i\mu_j$ = g (.)*(g (.)* I(x, y, t)), g(.) being the Gaussian kernel.

The Harris corner function can be given using the Eigen values ($\lambda_1, \lambda_2, \lambda_3$) of the second moment matrix as,

$$H = \lambda_1 \lambda_2 \lambda_3 \ - {}_k (\lambda 1 + \lambda_{2} {}_{+} \lambda_3)^3 \qquad (3.2)$$

Here the k is constant having value between 0.04 and 0.06.

The descriptor of the points is calculated as the gradient around the interest point. The interest points and descriptors together form the Bag of Words representation of the video.

B. *Spatial Relation Among Interest Point*

The interest points are arranged sequentially according to the frame sequence. The temporal relation is identified based on the sequence of the features. For spatial relation the interest points are analyzed using its space location.

The space domains of the video sequence are divided into four regions as upper left, upper right, lower left and lower right with respect to human body. The interest points are categorized into these regions and the whole region is tracked through the temporal domain. The tracked regions can be shown as in figure 2.

$x_p < x_i$ & $y_p > y_i$ => Upper Left
$x_p > x_i$ & $y_p > y_i$ => Upper Right
$x_p < x_i$ & $y_p < y_i$ => Lower Left
$x_p > x_i$ & $y_p < y_i$ => Lower Right

where (x_p ,y_p) co-ordinates for key point p and (x_i ,y_i) co-ordinates for ideal point which can be chosen as median of all key points of the frame.

Now as we know the action performed by human hand has the action structure lies within the upper left and upper right region. Hence based on this statement we can represent the human hand movement video with the bag of feature representation of these two regions.

IV. Datasets, Experiments and Results

For the experiment we implement the original interest point based Bag of Word representation and gather result using k-NN classifier with one neighbor on the KTH dataset. Then the proposed method is implemented and compared the result with the previous one. We also analyze the result with very efficient methods based on dense trajectory [14] and SIFT trajectory [16].

A. Dataset

We use the standard KTH dataset. KTH dataset includes different human actions like hand clapping, hand waving, boxing, jogging, walking, etc. Each action is performed by 25 humans in four different scenarios: outdoors, outdoors with scale variation, outdoor with different clothes and indoor [15]. The dataset is collected with static background and using single static camera. All video sequences are in .avi format and has frame rate 25 frames/sec.

B. Experimental setup

The methodology is implemented on Matlab. For the dense trajectory, we use the online code given by Wang et al. [14]. The SIFT trajectory is implemented using the computer vision library that provides algorithm for extracting SIFT key points. The original bag of words for interest points can be implemented using algorithm given by Laptev [2].

We analyze the result for the hand clapping, boxing and hand waving. The training data is gathered using the video of different person of different action. The classification is tested against the stability of recognition during the scale variation and cloth variation. Hence we use 25 videos of hand clapping, boxing and hand waving performed by different person for training and 50 videos of three actions with scale variation and clothe variation for testing.

C. Result

In this section we present the important result. The figure 1 shows the detected space time interest points (STIP) with their neighborhood window. The descriptors are defined according to this window.

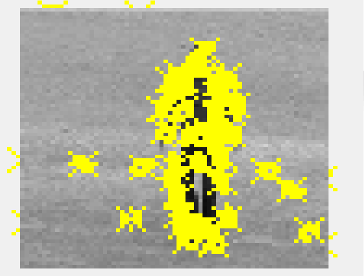

Fig. 2: The STIP feature for a frame containing a standing human

The figure 2 shows the different interest points groups according to their spatial relation.

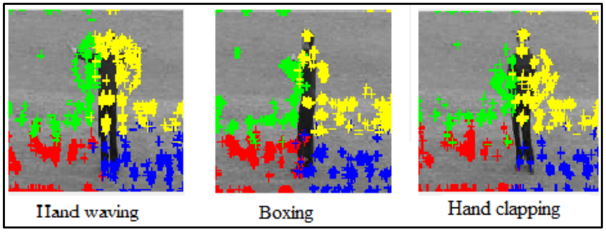

Fig 3: Illustration of spatial relation based grouping of features

The result of the proposed method with the original method of BoW model is analyzed. The dense trajectory gives the highest accuracy on KTH dataset but it is very complex and requires higher cost. The accuracy is more for boxing action while for hand waving and hand clapping is less that is due to having same semantics and spatial relation of patterns for those two actions. The proposed method provides reasonable accuracy with minimum cost. The implementation setup doesn't need any extra tools or library for support.

V. Conclusions

This paper describes the simple yet effective method proposed by us. This method models the spatial relationship between feature points and based on that relationship forms the BoW model. We have analyzed the result of original BoW model and BoW model with proposed method. Hence the spatial relationship provides the complementary information for the BoW model.

References

[1] Sarvesh Vishwakarma, Anupam Agrawal: ”A survey on activity recognition and behaviour understanding in video surveillance” Springer 2012

[2] Ivan Laptev, Tony Lindeberg: “Space-time Interest Points” IEEE International Conference on Computer Vision 2003

[3] Christian Schuldt, Ivan Laptev, Barbara Caputo: “Recognizing Human Actions: A Local SVM Approach” IEEE International Conference on Pattern Recognition 2004

[4] Feng Shi: “Local Part Model for Action Recognition in Realistic Videos” University of Ottawa, Canada 2014

[5] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri: “Actions as space-time shapes,” IEEE Transaction Pattern Analysis Machine Intelligence 2007

[6] Roberto Lublinerman, Necmiye Ozay, Dimitrios Zarpalas, Octavia Camps: “Activity Recognition from Silhouettes using Linear Systems and Model validation Techniques” IEEE International Conference on Pattern Recognition 2006

[7] Moin Nabi, Alessio Del Bue, Vittorio Murino: “Temporal Poselets for Collective Activity Detection and Recognition” IEEE International Conference on Computer Vision Workshops 2013

[8] P. Doll´ar, V. Rabaud, G. Cottrell, S. Belongie: “Behavior recognition via sparse spatio-temporal features” IEEE International Workshop 2005

[9] J. C. Niebles, H. Wang, L. Fei-fei: “Unsupervised learning of human action categories using spatial-temporal word” International Journal of Computer Vision 2008

[10] Ivan Laptev, Marcin Marszałek, Cordelia Schmid, Benjamin Rozenfeld: ”Learning realistic human actions from movies” IEEE Conference on Computer Vision & Pattern Recognition 2008

[11] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie: “Behavior recognition via sparse spatio-temporal features” IEEE 2nd Joint International Workshop 2005

[12] G. Willems, T. Tuytelaars, L. Van Gool: “An efficient dense and scale-invariant spatio-temporal interest point detector” Computer Vision ECCV 2008

[13] P. Matikainen, M. Hebert, R. Sukthankar: “Trajectory: Action recognition through the motion analysis of tracked features” ICCV Workshops on Video-Oriented Object and Event Classification 2009

[14] Heng Wang, Alexander Klaser, Cordelia Schmid, Liu Cheng-Lin: “Action Recognition by Dense

Trajectories" IEEE Conference on Computer Vision & Pattern Recognition 2011

[15] Laptev: " Local Spatio-Temporal Image Features for Motion Interpretation" KTH Department of Numerical Analysis and Computer Science 2004

[16] Jia-Tao Zhang, Ah-Chung Tsoi, Sio-Long Lo: "Scale Invariant Feature Transform Flow Trajectory Approach with Applications to Human Action Recognition" IEEE International Joint Conference on Neural Networks 2014

[17] Xiaojiang Peng, Limin Wang, Xingxing Wang, Yu Qiao: " Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice" Elsevier Computer Vision and Image Understanding 2016

[18] J. Liu, S. Ali, M. Shah: "Recognizing human actions using multiple features" IEEE Conference on Computer Vision & Pattern Recognition 2008

[19] N. Kardaris, V. Pitsikalis, E. Mavroudi, P. Maragos: "Introducing temporal order of dominant visual word sub sequence for human action recognition" IEEE International Conference on Image Processing 2016

[20] Guangchun Cheng, Yiwen Wan, Wasana Santiteerakul, Shijun Tang, Bill P Buckles: "Action Recognition with Temporal Relationships" IEEE Conference on Computer Vision and Pattern Recognition Workshops 2013.