ATLANTIS PRESS

Advances in Engineering Research (AER), volume 142

International Conference for Phoenixes on Emerging Current Trends in Engineering and Management (PECTEAM 2018)

MANAGING INTERCLASS VARIATION IN HUMAN ACTION RECOGNITION

[1]Akila K, [2]Chitrakala S

[1]Assistant Professor, Dept. of CSE, R.M.K College of Engineering & Technology Chennai, India

akilmoorthy@yahoo.com

[2]Associate Professor, Dept of CSE, Anna University, Chennai,India

au.chitras@gmail.com

***Abstract*--Background: Human action recognition encompasses a scope for an automatic analysis of current events from video and has varied applications in multi-various fields. Recognizing and understanding of human actions from videos still remains a difficult downside as a result of the massive variations in human look, posture and body size inside identical category. Objective: This paper focuses on a specific issue related to inter-class variation in Human Action Recognition. To discriminate the human actions among the category, the poses of body parts are estimated and there by trailing its motion sporadically with geometric joints feature. Analysis: Example actions are listed to illustrate the similarity between the actions and steps to emulate the enhancement to discriminative the power of recognizing the similar actions. Conclusion: Experiments results have shown that the proposed approach is discriminative for similar human action recognition and well adapted to the inter-class variation.**

Keywords: **video processing, Feature Descriptor, Spatio-temporal, Action Recognition**

I. INTRODUCTION

Amazing growth of digital video, increases the demand in video analysis, creates a human into a tedious task. The human perceptual system has the inconceivable capability to flawlessly and quickly process visual knowledge thereby interpreting and recognizing thousands of objects in the environment. Hence, making a machine to understand a video becomes significant one and challenging too as there exists a gap between low-level features and high-level semantic content. Thus, Visual Recognition problem turns out to be a central one to Computer Vision Research (CVR). Many desired applications demand the ability to identify and localize categories, places, and objects. Specifically in surveillance systems , it is difficult for a manpower to intensive monitor the data collected from various cameras continuously. This gives rise to the necessity for automatic understanding of human actions and building a higher level knowledge of the events occurring in the scene by the computer vision system. These systems require cognitive vision techniques for analyzing videos which in real life scenarios. An automated system is required, to continuously monitor and process the input video to sense any unusual findings, which they can report to the supervisor for their alertness. Our ultimate aim is to develop a low cost automated vision based framework over the existing expensive sensor based systems.

Over the last two decade, researchers have explored application scenarios for HAR systems but not limited to, surveillance, healthcare, sports broadcast, machine/robotic control, machine-human / human-machine interaction, video retrieval and much other. Apart from the potential application scenarios, the task of recognizing action is much more difficult due to dynamic backgrounds, camera movements, and occlusion of scene surroundings, illumination variation, varying camera view point, lack of depth information, occlusion, overlapping objects, shadowed region and etc., The stupendous growth in interest in typifying human actions is partly due to the mounting number of real-world applications such as action centric video indexing and reclamation, human-computer interface, activity supervising in investigation set-up, scrutiny of sports videos, the expansion of smart surroundings, and so on. Recent technology and market trends have demanded the significant need for feasible solutions to video/camera systems and analytics. The emergence of multimedia systems and computer vision bring the challenge of accurate action recognition among similar class of actions.

II. RELATED WORK

Several approaches for human action recognition have been proposed. A survey on HAR can be found at [1] . A variety of approaches use features which describe the motion and/or shape of the entire human body figure to perform human action recognition. In the spatial domain, points with a significant local variation of image intensities have been extensively investigated in the past few decades. Such image points are frequently referred to as Interest points and are attractive due to their high information content and relative

stability with respect to perspective transformations of the data. In this paper Human Action Recognition(HAR) for both DEPTH sequence as well as RGB video sequence are analysed. In depth based HAR - the depth information of video sequence is used whereas in RGB based HAR - the extended notion of interest points into the spatio-temporal domain is used for a compact representation of video data as well as for interpretation of spatio-temporal events. Latent Dirichlet Allocation(LDA) is then used for modeling the human action [2].

Aggarwal and Ryoo [3] divides the recogntion methodologies into two major categories: single-layered approaches and hierarchical approaches. As such, single-layered approaches mainly recognize common actions and these recognized simple primitive actions can be employed to detect more complex action recognition using hierarchical conbinations, The methods are characterized by the activities to be recognized directly from the raw video data instead of primitive sub-actions or sub-activities. Various researchers tried to incorporate person models such as silhouettes or skeletons for action recognition. Ikizler and Duygulu [4] proposed a new pose descriptor called histogram of oriented rectangles(HOR) for action recognition. Kim and Cipolla [5]extended Canonical Correlation Analysis (CCA) to measure video-to-video similarity. Wang et al. [6] proposed an approach to describe videos by dense trajectories. They sampled dense points from each frame and tracked them based on displacement information from a dense optical flow field. Local descriptors of HOG, HOF and MBH (motion boundary histogram) around interest points were computed. Standard hidden Markov models have been widely used for state model-based approaches in [7]. HMMs are also extended to CHSMMs to model duration of human activities [8]. In previous work context-free grammers (CFGs), based on syntactic approaches, have been studied and applied in human action recognition. Several probabilistic extension of CFGs { stochastic context-free grammers(SCFGs) { are introduced in [9], [10]. Generally two-layer frameworks are proposed; the lower layer mostly functions to recognize atomic or low-level actions and the higher layer uses parsing techniques for the high-level activity recognition. Another limitation is that user must provide a set of production rules and in order to overcome such limitations Kitani et al. [11] introduced an algorithm to automatically learn rules from observations.

A. SIGNIFICANCE OF FEATURE DESCRIPTOR

The Feature descriptors should be a discriminativeness so that it discriminate the features of two different regions are different. Also these features should be affine invariant. On the other hand we come across at various examples of images where within the same class the same local feature can have much larger variance. In such case these feature descriptor should be able to lodge the intraclass variation but also give good inter class discriminativeness. Spatial features comprises of motion and shape information from a single frame. Spatio-temporal descriptor pattersn are formed to improve the accuracy of spatial features. One of the most popular approaches to interest point detection in the spatial domain is based on the detection of corners, such as Corners are defined as regions where the local gradient vectors point in orthogonal directions The gradient vectors are obtained by taking the first order derivatives of a smoothed image. [12] SIFT interest point detection on the first frame to identify candidate features and STIP space Time Interest Points effectively captures the local structure in spatio temporal dimensions of the video sequence are widely used descriptor in human action recognition. MoSIFT interest point detector is to treat appearance and motion separately, and to explicitly identify spatially-distinctive regions in a frame that exhibit sufficient motion at a variety of spatial scales. The most common descriptions are scale-invariant feature transform (SIFT), speeded-up robust features (SURF), which have advantages of scale, affine, view and rotation invariance [13]. Histogram of Oriented Gradient (HOG) and Histogram of Optical Flow (HOF) descriptors which encodes appearance and motion information of action in the video respectively [14]. In order to better represent the speed and movement characteristics of human actions, two novel descriptors are proposed, so-called the Histogram of Changing Points (HCP) and Average Speed (AS) , [15]which have recently been used in action recognition Spatio-temporal motion descriptor. Summary of various feature descriptor is given in table1.

TABLE 1
VARIOUS FEATURE DESCRIPTOR

Features	Description	Constraints
Space-time volumes (STV)	Features such as space-time saliency, action dynamics, shape structure and orientation	Global features are sensitive to noise, occlusion and variation of viewpoint.
Discrete Fourier transform (DFT)	Geometric structure (shape) in the spatial domain with image intensity variation	
Scale-invariant feature transform (SIFT) features	High dimensionality, not sufficiently discriminative	Local features are designed to be more robust to noise and occlusion, and possibly to rotation and scale.
Histogram of oriented gradient (HOG) features	extracted at a fixed scale	
Nonparametric weighted feature extraction (NWFE) features	Relies on accurate human body silhouette and contour, and ignore the color appearance	
Lucas-Kanade-Tomasi (LKT) features	Track human body joints in key frames and actual frames	
Shape-based features	Need an accurate silhouette segmentation	
Appearance-based features	Sensitive to clothing and illumination changes	
Human body including simple blobs, 2D/3D body modeling	Body modeling requires the 2D/3D pose estimation problem	Low dimensional or more discriminative features.

III. HUMAN ACTION RECOGNITION PROCESS:

Major process of Human Action Recognition is depicted in fig 1 which involves Pre-processing, Feature Extraction, Feature Selection / Descriptor, Human Detection, Action Recognition and Classification. Frames are extracted from a video on which initial pre-processing and segmentation is done like background subtraction, clustering etc and once the core or required part is obtained either in patches or as a whole, feature extraction is done where HOG features are extracted. High level feature extraction can further be done which involves PCA for dimensionality reduction. Then comes the classification which labels the class into particular category based on the available ones.

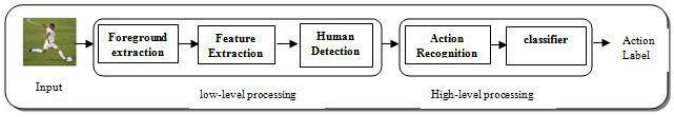

Fig.1 General Architecture of Human Action Recognition process

A. Spatial Binning/ Orientation:

In order to account for changes in illumination and contrast, the gradient strengths must be locally normalized, which requires grouping the cells together into larger, spatially connected blocks. The HOG descriptor is then the vector of the components of the normalized cell histograms from all of the block regions. These blocks typically overlap, meaning that each cell contributes more than once to the final descriptor. Each HOG feature vector computes edge orientation histogram and is accumulated into orientation bins over object regions i.e. spatial cells. Cells are rectangular blocks and the orientation bins are evenly spaced over 0◦–180◦ as an unsigned gradient with the step size of 20◦ to 40◦. The process of Human Action Recognition is clearly depicted in the functional architecture fig 2.

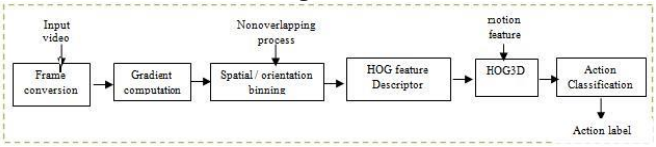

Fig.2 Functional of Human Action Recognition using HOG and motion feature

B. Motion Characterization

Motion pattern is a good feature to discriminate and but it is a weak classifier on its own. Combining motion with appearance or shape makes a strong classifier. Calculate optical flows and compute differential optical flow to remove camera motions, then combine the gradient and motions feature descriptors of HOG from spatial and orientation cells. the optical flow vectors calculated between two frames indicate the directions of motion as shown in fig 3. For better invariance to illumination and noise, a normalization step is usually used after calculating the histogram vectors. Four different normalization schemes have been proposed: L2-norm, L2-Hys, L1-sqrt, and L1-norm. This analysis used the L2-norm scheme due to its better performance: v v /($\| v \|_2^2 + \varepsilon^2$)^0.5 Where ε is a small positive value used for some regularization when an empty cell is taken into account and v stands for the characterization vector

Fig 3. (a) Gradient value

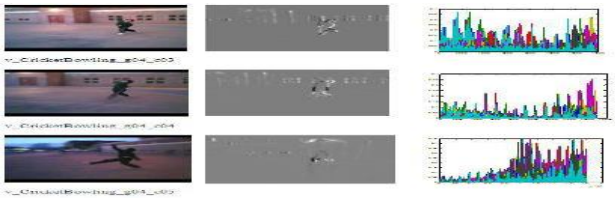

Fig 3. (b) optical flow.

C. Classification

In the last stage, the hypotheses are verified with computationally efficient SVM classification mechanism. A linear SVM was chosen because it has historically shown very good performance in lots of real world classification problems and also can deal with very high dimensional feature vectors. It uses a different set of indexed features extracted from the previous stage. The training part is done off -line to get SVM parameters. The classification part is just a inner product between SVM parameters and motion feature obtained from testing video. While the size of the feature vector is lesser, the classification performance of SVM classifier is more accurate, since the number of false positives is low.

IV. EXPERIMENTAL RESULTS

Our feature space consists of scale invariant HOG and HOF descriptors, from the decomposed spatial blacks. For HOF, optical flow is computed using Horn Schunk and the resultant vector is quantized using 9 bins. Here we use SVMs for auto video classification to evaluate the efficiency/accuracy trade-off by selecting key shots from video frames in the feature extraction process. Based on our experimental results we proved that histogram gradient and flow vectors are adequate for accurate video classification. Finally, we give a comparison with the state-of-the-art.

A. Compilation of common difficulties in Human Action Recognition Process

Each and every phase of HAR process finds some complexity due to some issues which are listed out in the table 2. Basic method to extract foreground object from the video is Background subtraction.

TABLE 2
COMMON DIFFICULTIES IN HAR PROCESS

Module-wise	Issues	Inferences
Background Subtraction	♦ Background clutter ♦ Dynamic illumination ♦ Camera movements ♦ Dynamic background	♦ Hierarchical method ♦ Temporal differencing ♦ Wrapping method ♦ Adaptive background model
Feature Extraction	♦ Presence of shadow ♦ appearance of noise	♦ color constancy model ♦ Gabor filtering
Human-object Detection	♦ view invariance ♦ change in style ♦ change in anthropometry ♦ change in dressing ♦ change in motion speed	♦ 3D optical flow ♦ Super-quadries 3D body model ♦ Harmonic Motion Context (HMC)
Tracking	♦ Full / partial occlusion	♦ Kernel density function
Action Detection / Classification	♦ Intra-class variation ♦ Inter-class similarity	♦ modelling spatio-temporal features ♦ highly supervised training based on context
Action Recognition	♦ performance is based choice of features for action representation	♦ The accuracy and efficiency of an action recognition method critically depends on the model and representation of actions

B. Analysis of challenges in Recognition of similar actions

Actions which are seem to be similar, for example action of interest involves a different object with same motion pattern actions that involve the same object have distinct spatial relation between the object and the human actions that involve the same object, with less spatial changes over time.

Fig 4. Example Actions having similar motion pattern and pose

Fig 4. shows a sample set of similar class of actions are taken to learn and analyze the features to carry out accurate action recognition. These set of action classes are really said to be challenging one, as it endowed with confusing and disambigious features. This challenging task can be resolved by various methods which are given in table 3 with the specific area of domain of applications.

Running and walking : the position of the hands and the legs differs for running and walking , while running the hand is placed close to chest while walking it is let loose . while running the front heel is rested on ground first and while walking the back heel is rested on the ground.

Drinking and smoking: 1) by the position of the fingers holding the object. In the case of smoking the cigar is holed usually with 2 fingers, whereas on the other case the glass object must be hold with a hand. 2) the second way is by means of identification of the object used in the respective pictures.

smile and laugh : smiling and laughing can be distinguished by the lips stretch and the eyes of the humans differs while smiling eyes are wide open and while laughing it is minimised.

tennis server - forehand: 1) while serving the ball is placed in the server hand and the angle of projection of the ball is created when it leaves the hand of the server and hits the bat, the bat is positioned in a vertical position. 2) the racquet is placed with dominant hand while the other hand holds your racquet at its throat. As the ball comes over the net and approaches your wing, open the shoulders by turning and taking the racquet back.

clapping and waving: while clapping the two hands are joint together to make sound and while waving the hands are not in contact with each other i.e there is no point of contact between the hands.

hand-shake and kicking: the position and extension of the arms will be able to differentiate and while hand shaking the arms of both persons are enclosed**.**

dialling and answering the phone: the position and placement of the mobile phone helps to differentiate it , while answering the phone is placed in close to ears.

TABLE 3
DISCRIMINATIONS OF SIMILAR ACTIONS

Similar Actions	Discriminated by	Finds Application in
Running and walking	Periodic motion of Body-parts	Video surveillance - to find the stranger
Drinking and smoking	Spatio-temporal features	Activity analysis
smile and laugh	Spatial feature	Sentiment Analysis - to study about emotion
tennis serve - forehand	Pose-estimation & Human -object interaction	Sports Analysis - to provide the score
clapping and waving	hand-hand contact	Event Detection
hand-shake and kicking	Tracking the movement of hands to generate spatio-temporal description of movement	Patrolling- to detect usual / unusual event

C. Data set and evaluation measures

Here we perform video retrieval experiments on three different datasets namely KTH, UCF50 and UCF sports data sets. The UCF50 dataset contains 50 different video categories with overall 6600 realistic videos from Youtube which includes variations in camera motion and illumination conditions etc. UCF sports data set contains videos with different frame rate and resolutions. The frames of the videos are 320 by 240 pixels. For state-of-the-art comparison, we validate the accuracy level of our proposed model using KTH pubic datasets.. The computational efficiency and accuracy measure and state-of-the-art comparison results for UCF 50 and KTH dataset are specified in Table 4.

TABLE 4
PERFORMANCE METRIC EVALUATION OF PROPOSED REFINED HOG+HOF MODEL.

DATA set-used	Trade-off accuracy/efficiency
KTH public data set	83% 20 frames/sec
UCF 50	64% 20 frames/sec

Due to the limited number of samples (persons) in the dataset, the leave-one-out method has been adopted where each run uses 24 persons (videos) for clustering and training and one person for testing. Then the average is calculated to give the final recognition rate by using KTH dataset, which is one of the largest public human activity video dataset, it consists of six action class (boxing, hand clapping, hand waving, jogging, running and walking) each action is performed by 25 actors each of them in four different scenarios including indoor, outdoor, changes in clothing and variations in scale. An experimental setup is made with data of 24 persons for clustering and training, and one person for

testing then the average of the results is computed to be the final result.

It is given a clear picture in the table 5, the majority of actions are correctly classified. An average accuracy is of 83% is achieved with our proposed method. The mistakes where confusions occur are only between jog and run actions and between wave and clap actions. This is also due to the high closeness or similarity among the actions in each pair of these. A sample set of similar class of actions are taken to learn and analyze the features to carry out accurate action recognition. These set of action classes are really said to be challenging one, as it endowed with confusing and disambigious features. Run and walk: the position of the hands and the legs differs for running and walking , while running the hand is placed close to chest while walking it is let loose . while running the front heel is rested on ground first and while walking the back heel is rested on the ground.

TABLE 5
CONFUSION MATRIX FOR THE CLASSIFICATION RESULT OF REFINED HOF DESCRIPTOR ON KTH DATASET

(%)	Box	Clap	Wave	Jog	Run	Walk
Box	**100**	0	0	0	0	0
Clap	0	**100**	0	0	0	0
Wave	þ	1.77	**98.23**	0	0	0
Jog	0	0	0	**95.4**	4.6	0
Run	0	0	0	16.47	**83.53**	0
Walk	0	0	0	2.77	0	**97.22**

V. CONCLUSION

This paper focused on a explicit issue related to inter-class variation in Human Action Recognition. To discriminate the human actions among the category, the poses of body parts are estimated and there by trailing its motion sporadically with geometric joints feature. Example actions are listed out to illustrate the similarity between the actions and established the steps to emulate to enhance the discriminative power of recognizing the similar actions.

REFERENCES

[1] G. Cheng, Y. Wan, A. N. Saudagar, K. Namuduri, and B. P. Buckles, "Advances in Human Action Recognition: A Survey," pp. 1– 30, 2015.

[2] M. M. Moussa, E. Hamayed, M. B. Fayek, and H. A. El Nemr, "An enhanced method for human action recognition," *J. Adv. Res.*, vol. 6, no. 2, pp. 163–169, 2015.

[3] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 1–43, 2011.

[4] N. Ikizler and P. Duygulu, "Histogram of oriented rectangles: A new pose descriptor for human action recognition," *Image Vis. Comput.*, vol.27, no. 10, pp. 1515–1526, 2009.

[5] G. Sasi and B. Rao, "Face Recognition Using Discriminate Analysis and Canonical Correlations," vol. 8491, pp. 1–4, 2015.

[6] H. Wang, A. Kl, C. Schmid, L. Cheng-lin, H. Wang, A. Kl, C. Schmid, L. C. Action, and A. Kl, "Action Recognition by Dense Trajectories To cite this version :," 2011.

[7] T. Ravet, Jo, #235, L. Tilmanne, and N. D'Alessandro, "Hidden Markov Model Based Real-Time Motion Recognition and Following," *Proc. 2014 Int. Work. Mov. Comput.*, pp. 82–87, 2014.

[8] P. Natarajan and R. Nevatia, "Coupled hidden semi Markov Models for activity recognition," *2007 IEEE Work. Motion Video Comput. WMVC 2007*, 2007.

[9] D. Moore, "Recognizing Multitasked Activities from Video using Stochastic Context-Free Grammar Introduction & Related Work Representation using SCFG The Earley-Stolcke Parsing AAAI-02," *ReCALL*, pp. 770–776, 2002.

[10] Y. a. Ivanov and A. F. Bobick, "Recognition of visual activities and interactions by stochastic\nparsing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 852–872, 2000.

[11] K. Lee, Y. Su, T. K. Kim, and Y. Demiris, "A syntactic approach to robot imitation learning using probabilistic activity grammars," *Rob. Auton. Syst.*, vol. 61, no. 12, pp. 1323–1334, 2013.

[12] D. Ta, W. C. Natasha, G. Kari, and P. Alto, "SURFTrac: Ef cient Tracking and Continuous Object Recognition using Local Feature Descriptors," *Cvpr*, pp. 2937–2944, 2009.

[13] L. Juan and O. Gwun, "A comparison of sift, pca-sift and surf," *Int. J. Image Process.*, vol. 3, no. 4, pp. 143–152, 2009.

[14] Y. Shi, Y. Tian, Y. Wang, and T. Huang, "Sequential Deep Trajectory Descriptor for Action Recognition with Three-stream CNN," vol. 15, no. 9, pp. 1–11, 2016.

[15] T. L. Vu, T. D. Do, C. Jin, S. Li, V. H. Nguyen, H. Kim, C. Lee, and I. Introduction, "Improvement of Accuracy for Human Action Recognition by Histogram of Changing Points and Average Speed Descriptors," vol. 9, no. 1, pp. 29–38, 2015.