Improved Discriminative Model for View-Invariant Human Action Recognition

M.Niresh Kumar
M.Tech student Department of CSE
JNTUA College of Engineering
Anantapur
Email id: nireshkumar.m@gmail.com

Dr. K.Madhavi,
Assistant Professor
Department of CSE
JNTUA College of Engineering, Anantapur
Email id: kasamadhavi@yahoo.com

Abstract - Recognizing human actions play an important role in applications like video surveillance. The recent past has witnessed an increasing research on view-invariant action recognition. Huang et al. proposed a framework based on discriminative model for human action recognition. This model uses STIP (Space – Time Interest Point) to extract motion features and view invariants. Then a discriminative model is used to model known as hidden Conditional Random Fields (HCRF) for conditional probability estimation for action recognition. They focused on five classes of actions namely climb, jump, run, swing and walk. In this paper we extend the discriminative model proposed by Huang et al. to explore more classes of actions. We built a prototype application to demonstrate human action recognition. The experimental results revealed that the application is capable of accurately recognizing human actions.

Keywords –Human action recognition, motion detection, view invariants

I. INTRODUCTION

Video surveillance and computer vision are the applications that became popular as they can be used to monitor environments. In such applications human action recognition is very important aspect as that is required by the real world applications. For instance a video surveillance application can monitor premises of a shop. There might be restricted areas towards which humans are not allowed. In this context, when humans move into the restricted area, the surveillance application can detect it and raise an alarm to alert people concerned. Thus automatic human action recognition has attracted considerable interest in academic and scientific circles. There are many problems involved in human action recognition. They include robust feature extraction, image data acquisition, and classifier with discriminative capabilities. Many researchers worked in the area of computer vision and tried to address these challenges. Appearance based methods are used by some researchers for action recognition. Motion features are used for action recognition. The features include trajectory and optical flow. Speed and position information is used to identify human actions. In [1], [2], [3], [4], [5], [6], [7] these features are used. However, in this paper view invariants are used as motion features give much performance. Contours of human body can be obtained using tracking and classification. In [8] human silhouette is used with R transform which resulted in low computational cost. Unstable contour features are also handed here. In [9] the R-transform has been extended to support temporal domain that can exhibit different contours. Later on skeleton modeling [10], [11] was introduced that can describe human interactions in terms of joints of human body.There are many traditional methods for the purpose. They are known as template matching methods. Later on HMMs (Hidden Markov Models) which are known as parametric models came into existence for action recognition. For contour feature extraction also Yamato et al. [12] used HMM. In involves Expectation Maximization (EM) in order to find transitions among hidden states. Coupled HMMs were also used in order to extract complex interactions between human beings. This proved to be robust to complex situations [3].

Though HMM model is robust, it has limitations such as conditional independence among observations. Other limitations include having prior knowledge, local optimization and data distribution. For better discrimination in video surveillance applications various methods are used. They include conditional random fields (CRFs), bag of words, SVMs (Support Vector Machines). Along with these approaches various learning models are used [4], [5], [6]. With the support of large training set, these models achieve highlevel of recognition rate. As public security is given much importance, many algorithms came into existence on intelligent video surveillance. When compared to traditional surveillance methods, the human action recognition is difficult under various backgrounds and constrained scenes. In this paper we throw light into view-invariant action recognition that can detect human actions in real time applications.

There are many methods pertaining to action recognition using view invariance. They are of three types such as3-D recognition techniques, epipolar geometric relations and machine learning techniques. The 3-D reconstruction techniques were capable of providing more reliable functionality. This concept is used in [7] and [8] where multiple cameras are used to obtain 3-D visuals in various poses. These methods have limitations as they are computationally expensive. The second approach uses geometric relations. It is explored in [9] where ratios of matrix are used. In this approach the drawback is that it needs manual labeling of joints in the human body. The third approach is the mapping between view invariant patterns and pattern recognition. View invariant features are not taken by these methods. Instead they assume that the method satisfies the underlying models implicitly. In spatiotemporal curvature of 2-D trajectory of human limbs is captured from video. Though this feature is view-invariant it degrades STNR as the curve is not consistent. Another method explored in makes an assumption that when body's joints are coplanar, there is a moment in action. However, the application has limitation as detecting canonical pose is not easy. Huang et al. [12] proposed a framework for human action recognition which makes use of a discriminative model. This paper extends the framework to support recognition of more classes of human actions.

The rest of the document is structured into some sections. Section II provides information on the proposed extension to the model explored in [12]. Section III provides experimental results while section IV concludes the paper.

II. PROPOSED ACTION RECOGNITION FRAMEWORK

The proposed human action recognition framework is an extension to the model proposed by Huang et al. [12]. The extension is in terms of supporting more classes of actions. In [12] the actions recognized are climb, jump, run, swing and walk. In this paper we explore recognition of other actions such as clapping and jagging. The basic framework of Huang et al. has two phases. They are motion detection and human action detection using discriminative model. The framework is as shown in fig. 1.

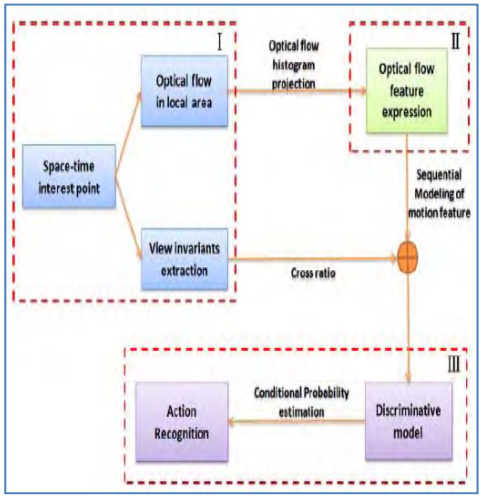

Fig. 1 –Action recognition framework (excerpt from [12])

As can be seen in fig. 1 it is evident that the framework takes video as input and recognizes actions such as clapping, jagging etc. As soon as input is given the framework identified the space- time interesting points. From STIPs it extracts optical flow in local area and also views invariants. In the second phase, optical flow feature extraction is done which results in sequential modeling of motion feature. At the same time the phase 1 returns cross ration through view invariants extraction. Both of these are taken as input in the next phase which will recognize human actions. The discriminative model estimates probability of actions which enables action recognition eventually.

A. STIP Detection

Space – Time Interesting Points from video are detected using the approach proposed in [5]. Generally object detection is done in the 2-D image plane using gray gradient. The detection starts from the corner. Measurement of corners is required in order to detect such regions. Laplacian of Gaussian (LOG) is used to measure it. In this corner detection approach the response function is computed as follows.

$$H = \det(\mu) - k \cdot \text{trace}^2(\mu)$$
$$= \lambda_1\lambda_2 - k(\lambda_1 + \lambda_2)^2$$

The STIP detection technique which is given in [25] results in the detection of space time interesting points. The sample detection is as shown in fig. 2.

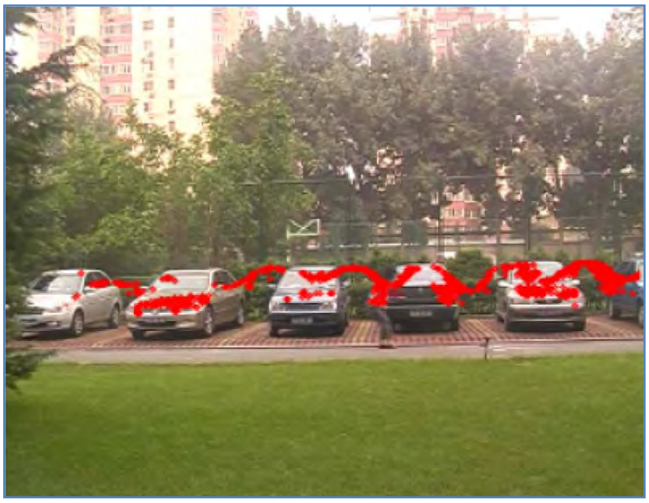

Fig. 2 – Illustrates STIP detection (excerpt from [12])

As can be seen in fig. 2, it is evident that the STIP has been detected. As shown in fig. 2 when a person is jumping, the red points illustrate the detected STIPs. These STIPs are further used in the process of motion detection.

B. Feature Extraction and Representation

When a live video is running, it is challenging to extract features that can help human action recognition. Stability under different views is the characteristic of view invariance. Then the entropy is used to know the bulk of information being carried. The more entropy is the less is the information being carried. The combination of motion features and view invariants are used in the learning methods in order to avoid tradeoff between discrimination and motion invariance.

C. View Invariants and Cross Ratios

Cross ratio is the most common invariant. As seen in fig. 3, the four sets of collinear points are forming cross ratios with similar value. Cross ratios are used as invariants and for this purpose before projection the image has to be taken as 3-D space.

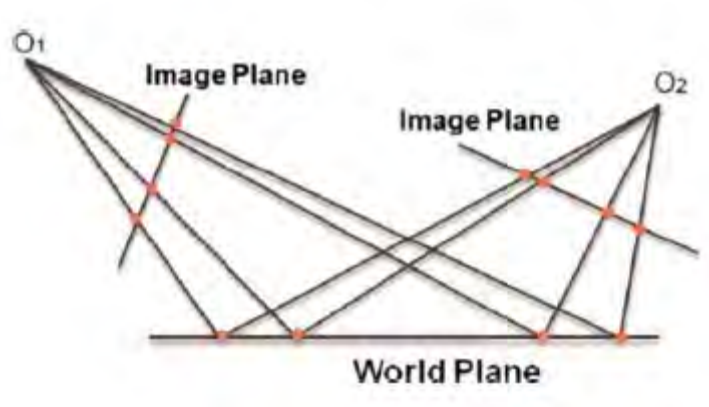

Fig. 3 – Illustrates sets of four points with similar cross ratios (excerpt from [12])

The computation of cross ratios is done as follows.

$$\mathrm{CR}_1(X_1, X_2, P, Q) = \frac{(|X_1X_4| + |X_4X_5| + |X_5X_1|)}{(|X_2X_4| + |X_4X_5| + |X_5X_2|)} \times \frac{(|X_2X_4| + |X_4X_3 + X_3X_2|}{(|X_1X_4| + |X_3X_4| + |X_3X_1|)}$$

$$\mathrm{CR}_2(X_5, X_4, P, Q) = \frac{(|X_5X_2| + |X_2X_1| + |X_1X_5|)}{(|X_5X_2| + |X_2X_3| + |X_3X_5|)} \times \frac{(|X_4X_2| + |X_2X_3 + X_3X_4|}{(|X_4X_2| + |X_2X_1| + |X_1X_4|)}$$

D. Action Modeling

After obtaining view invariants of interesting points, and motion feature description, the temporal information has to be extracted from sequential data and modeled. In order to model complex human action a discriminative model is used. This model is based on hCRF for representing motion features. In hCRF model, it is possible to fully utilize information from neighboring frames instead of a single frame. The structure of hCRF is as shown in fig. 3.

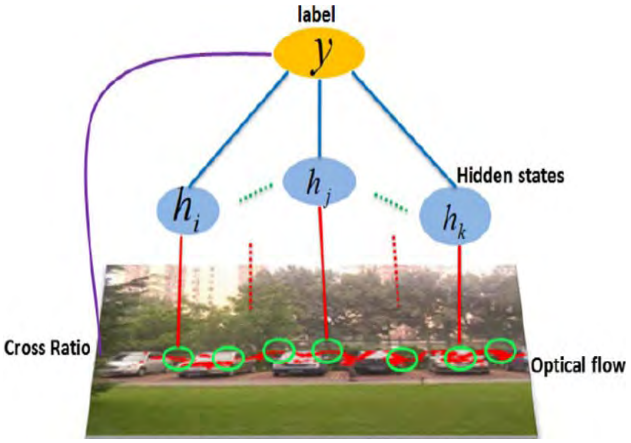

Fig. 3 – Illustrates structure of hCRF (excerpt from [12])

As can be seen in fig. 3, it is evident that the hidden states are represented as x. The x denotes observations which are in the form optical flow features. H represents hidden variable whereas y represents a class label for human actions. The overall conditional probability is computed as follows. More details of the motion detection approach can be found in [12].

$$p(y|x;\theta) = \sum_h p(y, h|x;\theta) = \frac{\sum_h \exp\left(\psi(y, h, x;\theta)\right)}{\sum_y \sum_h \exp\left(\psi(y, h, x;\theta)\right)}$$

III. PROTOTPYE APPLICATION

A prototype application is built using Java programming language. The environment used for the development is a PC with 4GB or RAM, Core 2 Dual processor and Window 7 operating system. The IDE (Integrated Development Environment) used is Net Beans. The main user interface of the application is as show in fig. 4.

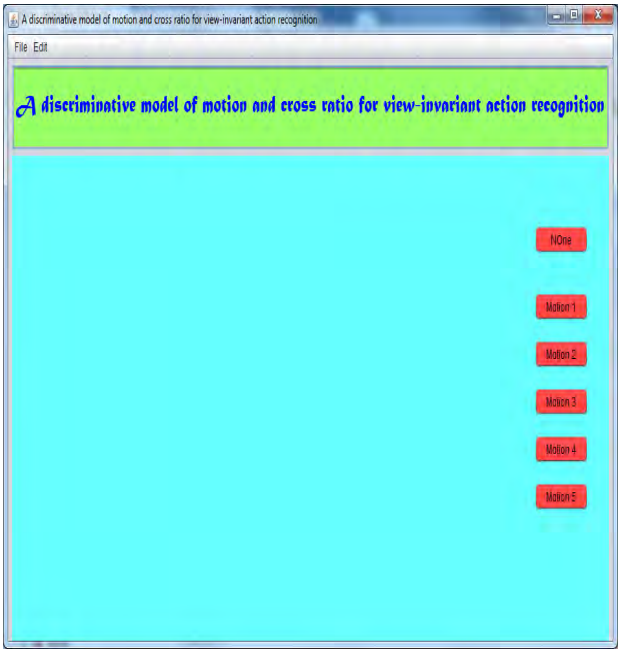

Fig. 4 – The main UI of the prototype application

As can be seen in fig. 4, the main user interface of the prototype application is shown. This will allow playing various videos and testing the human action recognition. The recognition of human actions such as clapping and jagging are shown in fig. 5 and 6.

Fig. 5 – Result of Clapping Action Recognition

As seen in fig. 5, the clapping action is detected ad STIP and then the action recognition framework is capable of recognizing clapping action.

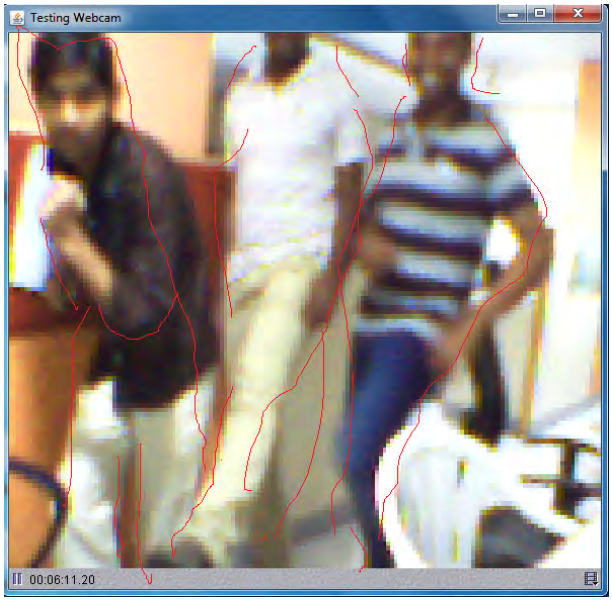

Fig. 6 – Result of Jagging Action Recognition

As seen in fig. 6, the clapping action is detected ad STIP and then the action recognition framework is capable of recognizing jagging action.

IV. EXPERIMENTAL RESULTS

This section provides the summary of experimental results. The experiments are done with various datasets such as Weizmann and KTH. The proposed approach is also compared with other approaches. The results of experiments are presented in a series of graphs presented in this section.

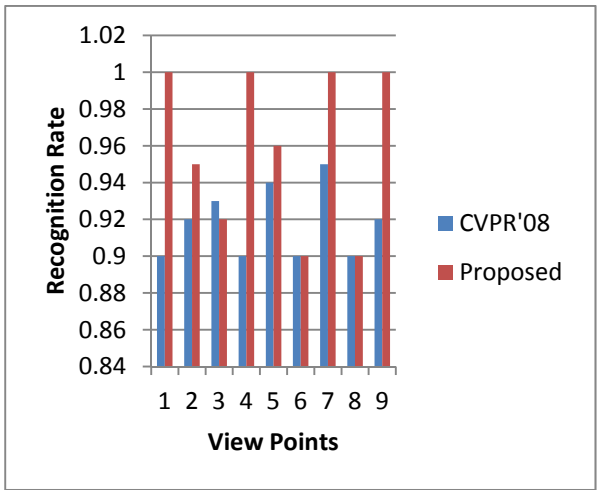

Fig. 7. Recognition rate in different views

As can be seen in fig. 7, the recognition rate in different views is presented. The horizontal axis represents view points and vertical axes represents .recognition rate. As seen in the results the recognition rate of proposed system is high.

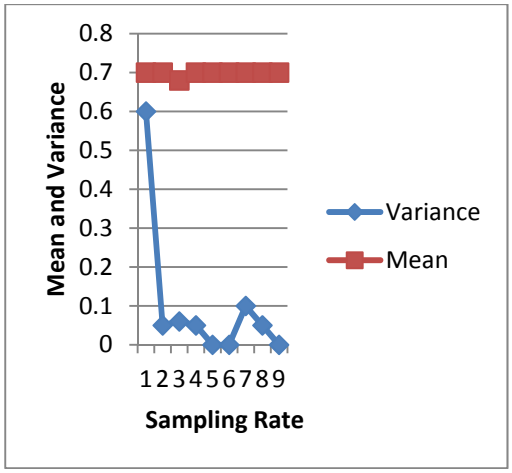

Fig. 8 - Mean and variance of CRs in different viewpoint frames

As can be seen in fig. 8 the mean and variance of cross ratios in different viewpoint frames are presented. The horizontal axis represents sampling rate while the vertical axes represents mean and variance.

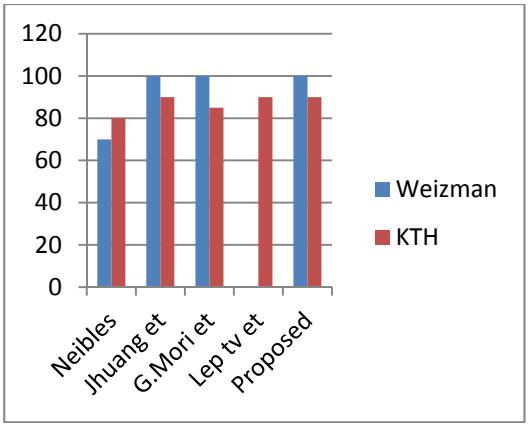

Fig. 9 - Comparison results on Weizmann and KTH action data sets

As can be seen in fig. 9 the recognition rate in different techniques is presented. The horizontal axis represents techniques while the vertical axis represents performance against Weizman and KTH datasets.

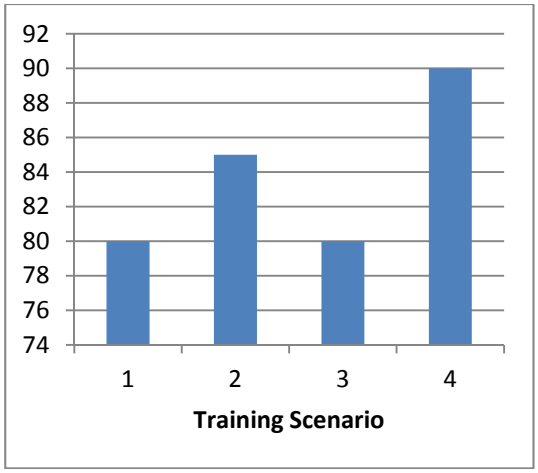

Fig. 10 - Results on different scenarios on KTH data set

As can be seen in fig. 10, the four scenarios are presented. The horizontal axis represents Training Scenario while the vertical axis represents performance against different scenarios. The results reveal the robustness of the proposed technique.

CONCLUSION

In this paper we have proposed an extension to the human action recognition framework proposed by Huang et al. [12] in terms of recognition of more action classes such as clapping, and jagging. The proposed framework makes use of both view invariants and also motion features in order to detection motion. Afterwards, it used discriminative model in order to recognize human actions. We also built a prototype application that takes live video or pre-existed video as input and demonstrates the efficiency of the proposed extended framework. The empirical results are encouraging.

REFERENCES

[1] F. I. Bashir, A. K. Ashfaq, and S. Dan, "View-invariant motion trajectory-based activity classification and recognition," *Multimedia Syst.*,vol. 12, no. 1, pp. 45–54, Aug. 2006.

[2] M. Ahmad and S.-W. Lee, "Human action recognition using shape and clog-motion flow from multi-view image sequences," *Pattern Recognit.*,vol. 41, no. 7, pp. 2237–2252, Jul. 2008.

[3] Y. Wang and G. Mori, "Learning a discriminative hidden part modelfor human action recognition," in *Proc. NIPS*, 2008, vol. 21, pp.1721–1728.

[4] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action ata distance," in *Proc. ICCV*, Nice, France, 2003, pp. 726–733.

[5] N. Johnson and D. Hogg, "Learning the distribution of object trajectoriesfor event recognition," in *Proc. 6th BMVC*, 1995, pp. 583–592.

[6] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity usingreal-time tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22,no. 8, pp. 747–757, Aug. 2000.

[7] K. Huang, D. Tao, Y. Yuan, X. Li, and T. Tan, "View independenthuman behavior analysis," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*,vol. 39, no. 4, pp. 1028–1035, Aug. 2009.

[8] Y.Wang, K. Huang, and T. Tan, "Human activity recognition based onr transform," in *Proc. IEEE CVPR*, 2007, pp. 1–8.

[9] R. Souvenir and K. Parrigan, "Viewpoint manifolds for action recognition,"*J. Image Video Process.*, vol. 1, pp. 1–13, 2009.

[10] A. J. Lipton, H. Fujiyoshi, and R. S. Patil, "Moving target classificationand tracking from real-time video," in *Proc. 4th IEEEWACV*, 1998, pp.8–14.

[11] H. Fujiyoshi and A. J. Lipton, "Real-time human motion analysis byimageskeletonization," in *Proc. 4th IEEE WACV*, 1998, pp. 15–21.

[12] Kaiqi Huang, Yeying Zhang and Tieniu Tan, "A Discriminative Model of Motion and Cross Ratio for View-Invariant Action Recognition". IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 21, NO. 4, APRIL 2012.