International Journal of Research in Advent Technology, Vol.3, No.2, February 2015
E-ISSN: 2321-9637

A Survey on Human Action Recognition using Scale Invariant Feature Transformation

Yogesh J. Wanare [1], Dr. P. R. Deshmukh [2]
M.E (Pursuing) [1], Prof. and Head of Dept. Amravati [2]
Computer Science and Engineering [1]
Sipna C .O. E. T, Amravati, India
Email: wanareyogesh@gmail.com [1], pr_deshmukh@yahoo.com [2]

Abstract- Human action recognition in videos is an important problem in computer vision, but it is very challenging sometimes especially when recognizing a large number of human actions. There are two main challenges with human action recognition, first it is difficult to capture the crucial motion patterns that discriminate among these actions. Second is the method should be scalable for large datasets because more training examples are often collected for more action classes. In this proposed work, we show latent models to capture the crucial motion patterns, and we propose an effective algorithm that will help us to efficiently address large datasets.

Index Terms- Scale Invariant Feature Transformation, Action recognition, keypoints, Spatio-Temporal domain.

1. INTRODUCTION

Human action recognition is the developing area in the computer vision. The aim of the human action recognition is to recognize the action performed in a video automatically. Numbers of algorithms have been used till date for the human action recognition. Human action recognition is still a challenging problem due to difficulties such as intraclass variation, scaling, occlusion, and cluttering. Motion and structure are the main important factors of the action occurring in a video sequence. Motion information is nothing but the body/body part movements and position translations, while structure information includes body poses, their occurring orders and relative positions. To obtain an informative representation, one needs to effectively encode both of them to characterize an action.

From human actions recognition we can get the wide range of applications, such as behavioural biometrics, content-based video analysis, and surveillance [1, 2, and 3]. As a result, a large number of studies have been conducted on some popular data sets, such as the Kungliga Tekniska Hogskolan (KTH) data set [4] and the Weizmann data set [5], where actions were recorded in well controlled settings such as clean background and fixed camera positions. However, how to recognize human actions with real-world scenarios is still remains a very challenging problem due to a number of issues such as dynamic backgrounds, camera movements, occlusion of scene surroundings, and illumination variations.

2. LITERATURE REVIEW AND RELATED WORK

Recently, local features have received a great deal of attention in video-processing applications. They are extended to take into account the spatio-temporal nature of video data. SIFT extension can be categorized into three groups: (1) extension of the descriptor part only, combined with 2D detectors, (2) a full 3D spatial extension, and (3) a combination of different approaches to separately describing motion and appearance.

P. Turaga, et al. [1] past decade has witnessed a rapid increase of video cameras in all walks of life and has resulted in a tremendous ignition of video content. Many applications such as content based video notation and retrieval highlight extraction and video summarization that require recognition of the activities occurring in the video. In this paper, the analysis of human activities in videos is an area with increasingly important consequences from security and surveillance to entertainment and personal storage. In this review paper, they present a comprehensive survey of efforts in the past couple of decades to address the problems of representation, recognition, and learning of human activities from video and related applications. They observe the problem at two major levels of complexity: 1) "actions" and 2) "activities." "Actions" are nothing but characterized by simple motion patterns typically executed by a single human. "Activities" are more complex and shows coordinated actions among a small number of humans

R. Poppe, [2] in this survey, they explicitly address these challenges and shows a detailed overview of current advances in the field. Image representations and the successive classification process are discussed separately to concentrate on the novelties of recent research. Moreover, they discuss

limitations of the state of the art and outline promising directions of research.

C. Schuldt, I. Laptev, and B. Caputo [4, 7] local space-time features capture local events in video and can be modify to the size, the frequency and the velocity of moving patterns. In this paper author demonstrate how such features can be used for recognizing complex motion patterns. They construct video representations in terms of local space-time features and integrate such representations with SVM classification schemes for recognition.

I. Laptev, T. Lindeberg [6] in this paper, they propose to extend the notion of spatial interest points into the spatio-temporal domain and show how the resulting features often reflect interesting events that can be used for a compact representation of video data as well as for its interpretation.

P. Dollar et al. [8] a common trend in object recognition is to detect and leverage the use of sparse, informative feature points of object. The use of such features makes the problem more manageable while providing increased robustness to noise and pose variation. In this work they develop an extension of these ideas to the spatio-temporal case.

P. Scovanner, S. Ali, M. Shah [9] introduce a 3-dimensional (3D) SIFT descriptor for video or 3D imagery such as MRI data. They also show how this new descriptor is able to better represent the 3D nature of video data in the application of action recognition. This paper will show that how 3D SIFT is able to outperform previously used description methods in an elegant and efficient manner.

W. Cheung, G. Hamarneh [10] propose the n-SIFT method for extracting and matching salient features from scalar images of arbitrary dimensionality, and compare the propose method's performance to other related features. The proposed features increase the concepts used for 2D scalar images in the computer vision SIFT technique for extracting and comparing distinctive scale invariant features and applying the technique to images of arbitrary dimensionality through the use of hyper spherical coordinates for gradients and multidimensional histograms to create the feature vectors.

M. y. Chen, A. Hauptmann [12] the goal of this paper is to build robust human action recognition for real world surveillance videos. Local spatio-temporal features throughout interest points provide compact but descriptive representations for video analysis and motion recognition. Current approaches tend to extend spatial descriptions by adding a temporal component for the appearance descriptor, which is only used to capture motion information. They propose an algorithm called MoSIFT, which calculate interest points and encodes not only their local appearance but also explicitly models local motion.

An empirical evaluation of the proposed methods has shown promising as below.

(Table 1 shows year wise representation of techniques used for human action recognition.)

Sr. No.	Authors	Year	Used Technique/ Approach
01	I. Laptev, T. Lindeberg	**2003**	Uses Space-time Interest Point method.
02	C. Schüldt, I. Laptev, and B. Caputo	**2004**	Propose A Local SVM Approach for Recognizing Human Actions.
03	M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri	**2005**	Actions as Space-Time Shapes Approach
04	P. Dollar, V. Rabaud, G. Cottrell, S. Belongie	**2005**	Propose Behavior Recognition via Sparse Spatio-Temporal Features
05	P. Scovanner, S. Ali, M. Shah	**2007**	Propose A 3-Dimensional SIFT Descriptor method
06	W. Cheung, G. Hamarneh	**2007**	Propose the *n*-SIFT method
07	P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea	**2008**	Discuss the problem at two major levels of complexity: 1) "actions" and 2) "activities."

08	S. Allaire, J. Kim, S. Breen, D. Jaffray, V. Pekar	**2008**	Propose 2D-to-3D extension SIFT Descriptor method.
09	M.y. Chen, A. Hauptmann	**2009**	Propose an algorithm called MoSIFT.
10	R. Poppe.	**2010**	A survey on vision-based human action recognition.

Table 1: Literature Review

3. ANALYSIS OF PROBLEM

Several numbers of algorithms have been used till date for the human action recognition. Some initial set of algorithms they are blob based, holistic and part-based representations methods were used for recognition. But they failed to produce accurate results under different variations. So there is need of an efficient algorithm which would produce accurate results even under such variations.

4. PROPOSED WORK

The proposed system is composed of three stages; initially the video or images containing the human actions are obtained. The obtained images divided into frames which are then input to the feature extraction algorithm for further processing. The key points are match with those in the database and the action performed is concluded.

The Fig. 1 shows the block diagram of the human action recognition using SIFT algorithm.

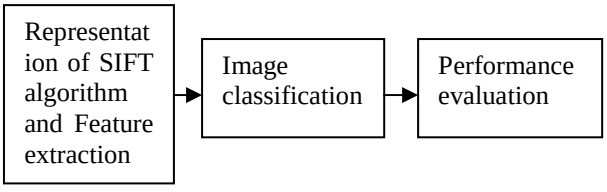

Fig. 1 Block diagram of human action recognition.

The proposed system is composed of three stages as given below.

- Representation of SIFT algorithm and Feature extraction
- Image classification
- Performance evaluation

4.1. Representation of SIFT algorithm and Feature extraction

The Fig. 2 shows the flow chart of proposed approach.

❖ **Project flow**

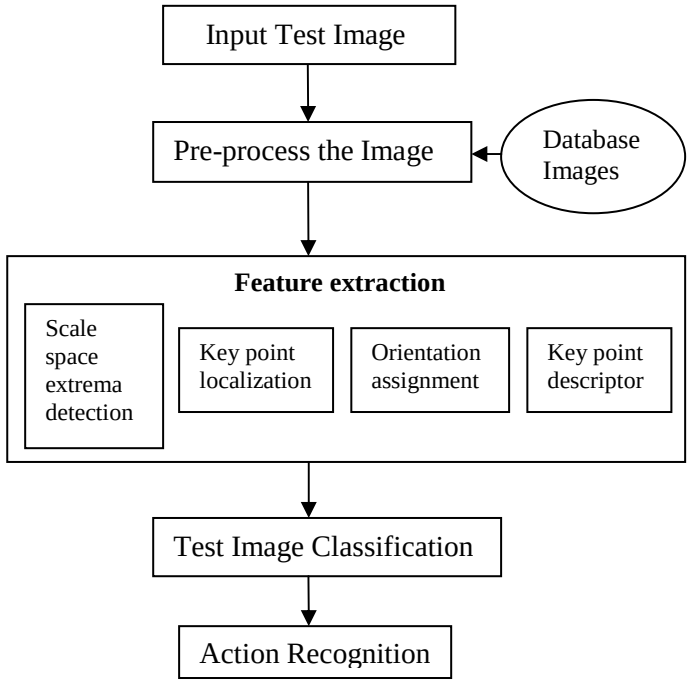

Fig. 2 Flow chart of proposed system.

1. Input Test Image

The Query Image which we have to recognize is given to the Input Test Image.

2. Pre-process the Image

The Query Image is now pre-processed with database images and then input to the feature extraction for recognition.

3. Feature extraction

Computation of SIFT image features is performed through the four consecutive phases which are described in the following.

3.1 Scale space extrema detection

This stage of the filtering attempts to identify those locations and scales those are identifiable from different views of the same object. This can be efficiently done by using a "scale space" function and it is based on the Gaussian function.

3.2 Key point localization

This stage attempts to eliminate more points from the list of keypoints by finding those that have low contrast or are poorly localised on an edge. This is done by calculating the Laplacian.

3.3 Orientation assignment

This step aims to assign a consistent orientation to the keypoints based on local properties of image. The keypoint descriptor can then be represented relative to this orientation and reach invariance to rotation.

3.4 Key point descriptor

The local image gradients are measured at the selected scale in the region around each keypoint. These are change into a representation that allows for significant levels of local shape distortion and change in illumination.

4. Test Image Classification

In Test Image classification the Query image is classified on basis of their extracted features which is input to action recognition.

5. Action Recognition

In order to get a reliable recognition, it is quite important that the features extracted from the training image are detectable even under changes in image scale, noise and illumination. Such points generally lie on high-contrast regions of the image.

4.2. Image classification

All the recognised images are need to be classified on basis of their actions.

4.3. Performance evaluation

In performance evaluation we will compare our calculated results with the different algorithms so that we can get the accuracy of our approach.

5. CONCLUSION

Recognition is one of the popular tasks in image processing, while to recognizing the human action is a critical one. In this seminar, we have surveyed the topic of human action recognition and we observe that some of the methods have provided efficient performance result. Even though the performances of these propose methods has yielded good quality results, we know that today's technology has been developing day by day, which shows for further improvement in the performance. So our approach will be to present the modifier to SIFT method for better human action recognition.

ACKNOWLEDGEMENT

It is a matter of great pleasure by getting the opportunity of highlighting fraction knowledge, I acquired during my technical education through this paper. This would not have been possible without the guidance and help of many people. This is the only page where I have opportunity of expressing my emotions and gratitude from the care of my heart to them. This paper would not have been successful without enlightened ideas; timely suggestion and keen interest of my respected Guide **Dr. P. R. Deshmukh** without his best guidance this would have been an impossible task to complete.

REFERENCES

[1] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1473–1488, Nov. 2008.

[2] R. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.*, vol. 28, no. 6, pp. 976–990, Jun. 2010.

[3] J. Aggarwal and M. Ryoo, "Human activity analysis: A review," *ACM Comput. Survey*, vol. 43, no. 3, p. 16, Apr. 2011.

[4] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. IEEE ICPR*, 2004, vol. 3, pp. 32–36.

[5] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space–time shapes," in *Proc. IEEE ICCV*, 2005, vol. 2, pp. 1395–1402.

[6] I. Laptev, T. Lindeberg, "Space-time interest points," *in International Conference on Computer Vision*. (2003)

[7] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. IEEE ICPR*, 2004, Vol. 3, (2004)

[8] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, "Behavior recognition via sparse spatio-temporal features," *in IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*. (2005)

[9] P. Scovanner, S. Ali, M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," *in International Conference on Multimedia*. (2007)

[10] W. Cheung, G. Hamarneh, "N-sift: N-dimensional scale invariant feature transform for matching medical images," *in International Symposium on Biomedical Imaging: From Nano to Macro*. (2007)

[11] S. Allaire, J. Kim, S. Breen, D. Jaffray, V. Pekar, " Full orientation invariance and improved feature selectivity of 3d sift with application to medical image analysis," *in Computer Vision and Pattern Recognition Workshops*. (2008)

[12] M. y. Chen, A. Hauptmann, "Mosift: Recognizing human actions in surveillance videos," Transform (2009)