Human Activity Recognition using Optical Flow based Feature Set

S.Santhosh Kumar and Mala John
Department of Electronics Engineering,
Madras Institute of Technology, Anna University,
Chennai, India.
santhosh.mitece@gmail.com, malajohnmit@gmail.com

***Abstract*— An optical flow based approach for recognizing human actions and human-human interactions in video sequences has been addressed in this paper. We propose a local descriptor built by optical flow vectors along the edges of the action performer(s). By using the proposed feature descriptor with multi-class SVM classifier, recognition rates as high as 95.69% and 94.62% have been achieved for Weizmann action dataset and KTH action dataset respectively. The recognition rate achieved is 92.7% for UT interaction Set_1, 90.21% for UT interaction Set_2. The results demonstrate that the method is simple and efficient.**

Keywords*— optical flow; feature descriptor; support vector machine; classification; human activity recognition

I. INTRODUCTION

The ever-increasing curiosity in characterizing human actions is fuelled, in part, by the rising number of real-world applications such as activity monitoring in video surveillance, elderly assistance in smart home, human-computer interaction, etc [1, 2, 3]. The performance of the recognition task depends on the efficiency of segmentation of the region of interest, extraction of image features and classification process. Wide range of techniques for activity recognition has been reported in the recent literature. An approach based on optical flow feature set is proposed herein to classify different human actions and human-human interactions using multi-class Support Vector Machine (SVM).

The organization of the paper is as follows. Section 2 gives the recent literature review. In Sections 3, the proposed approach for human action and interaction recognition using optical flow based feature set is discussed. The experimental result on standard datasets is provided in Section 4 followed by conclusions in Section 5.

II. RELATED WORK

Diverse methods have been proposed for identifying various human activities due to the dynamics in the underlying representation of human body. Recently, there has been significant interest in approaches that address human action and interaction recognition. Y. S. Sefidgar et al. [4] have developed a structured activity recognition model by incorporating temporal and spatial information obtained for the duration of human-human interaction in the form of a series of key-components. K. N. El Houda Slimani et al. [5] have extended the paradigm of bag-of-words meant for action recognition and proposed a framework with co-occurring visual words generated via 3D spatio-temporal volume information to describe the interactions between several persons. Y. Kong et al. [6] have used 3D interest points within the boundary of the action regions and constructed bag of words model to represent individual persons and max-margin formulation has been used for interaction recognition. M. Raptis et al. [7] have presented an action model with histogram of oriented gradient and bag of words based human poselets to recognize interactions in partial video observations. K. Yun et al. [8] have presented geometric relational body pose features built using color and depth information and tackled unrelated actions in entire sequence during real time interaction classification via Multiple Instance Learning approach. K. G. Derpanis et al. [9] have proposed spatiotemporal orientation based local descriptor that confines essential pattern dynamics to recognize actions in the given video using SVM. I. Everts et al. [10] have recognized realistic human actions in videos by means of color spatio-temporal interest points based representation of actions formulated by incorporating multiple photometric channels in addition to image intensities. B. Z. Yao et al. [11] have detected dissimilar actions from cluttered scenes in videos using shape and motion information in order to structure sequence of pose templates associated towards each action. C. Li et al. [12] have proposed motion energy oriented histogram based local action descriptor around the multi-velocity spatio temporal interest points and recognized human actions by means of bag-of-words framework. M. J. Roshtkhari et al. [13] have presented a probability based hierarchical codebook model for action recognition by calculating similarity between several spatio temporal video volumes.

The novel optical flow based feature extraction process proposed in this paper efficiently captures the variation in silhouette with time and the velocity vectors along the silhouette. An SVM classifier is used for classification. The next section describes the proposed method.

978-1-5090-1072-1/16/$31.00 ©2016 IEEE

III. Human Activity Recognition Framework

The human activity recognition framework is depicted in Fig. 1. In general, the video frames are subjected to foreground extraction in order to segment the region of interest. The set of widely used methods for foreground extraction includes Gaussian Mixture Model (GMM) based method [14] and optical flow based technique [15]. The extracted foreground of interest is subjected to feature extraction process. In this paper, we have proposed a novel, optical flow based feature extraction process. Optical flow feature set is used to represent the segmented object(s). The optical flow based feature vectors are computed along the boundary and the feature set incorporates the shape and instantaneous velocity information extracted along the boundaries of the action performers. The extracted optical flow based features are fed to a classifier. Classification algorithm working on the proposed feature set is the widely used multi-class SVM.

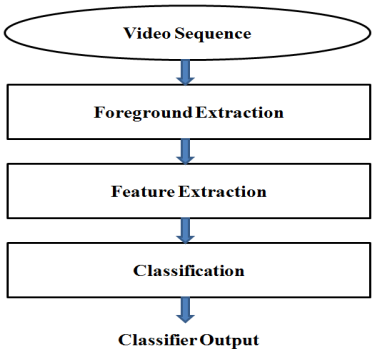

Fig. 1. Human activity recognition framework

The optical flow based feature extraction process that adds novelty to the proposed human activity recognition framework is discussed in detail in Section 3(A). However, the foreground extraction and classification process used are the popular ones reported in the literature. Therefore, the next section is devoted to the proposed optical flow based feature extraction algorithm.

A. Feature Descriptor

The spatio-temporal volume analysis used for human activity classification poses a lot of challenges. Thus, the key problem is to characterize actions and interactions within the action sequences by some kind of features. Feature design is a significant problem because informative features capture the essence of a behavior pattern while facilitating the classification task on a lower dimensional space.

As an example, the extracted optical flow vectors along the boundaries of the human action and interaction are shown in Fig. 2.

In the proposed technique, to start with, the centre of gravity (CG) of the foreground which includes human actions or human-human interactions is computed. Horn and Schunck algorithm [15] based optical flow extraction technique is used to compute the optical flow vectors along the boundaries. From the centroid, radial lines are drawn at $5k$ degrees (k is an integer) to intersect the boundary lines. The radial distances of the boundary points (72) that lie on these radial lines along with the optical flow vectors computed at these points are extracted and stored.

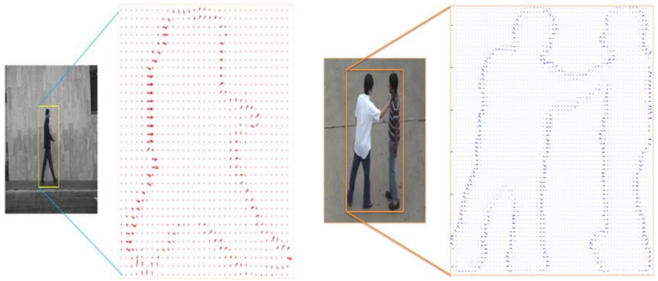

Fig. 2. Extracted foreground boundaries

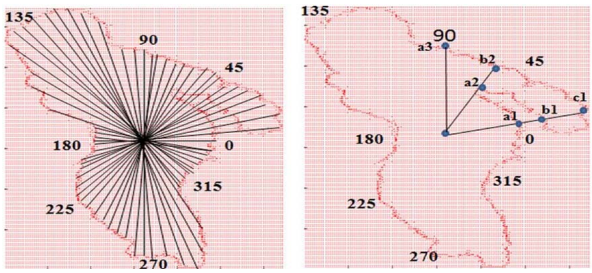

Fig. 3. Optical flow based feature extraction for human action

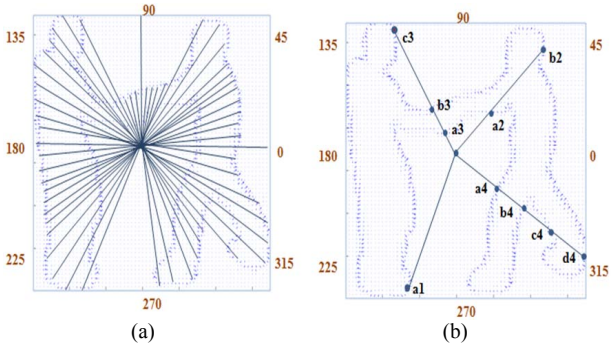

Fig. 4. Optical flow based feature extraction for human interaction

The number of points of intersection may differ from one radial line to the other (Fig. 3. and Fig. 4.). We consider a maximum of three such edge points starting from the innermost edge for human actions and four edge points for human-human interactions respectively. The radial distance (r) of the point of intersection and the magnitude of the velocities (u, v) are computed by considering the maximum number of edge points starting from the innermost edge, resulting in Z_k of dimension (1×9) for human actions and (1×12) for human-human interactions.

If the radial line intersects at only one edge, the radial distance and the velocity vector for the other two points in case of human actions and the other three points in case of human-human interactions are substituted as zero to create Z_k of dimension (1 × 9) and (1 × 12) respectively at every instance, independent of the number of points of intersections. For example, as shown in Fig. 3(b), the radial line at an angle of 90 degree intersects at one point viz., a_3.

The Z_k at this angle is given by

$$Z_k = (u_{a_3}, v_{a_3}, r_{a_3}, 0, 0, 0, 0, 0, 0) \quad (1)$$

In Fig. 4(b), the radial line at an angle of 315 degree intersects at four points viz., a_4, b_4, c_4 and d_4.

The Z_k at this angle is given by

$$Z_k = (u_{a_4}, v_{a_4}, r_{a_4}, u_{b_4}, v_{b_4}, r_{b_4}, u_{c_4}, v_{c_4}, r_{c_4}, u_{d_4}, v_{d_4}, r_{d_4}) \quad (2)$$

In action recognition, Z_k, the feature vector of the point placed at an angle of $5k$ degree, is of dimension (1 × 9). Therefore, the feature vector Z thus extracted is of dimension (72 × 9). Z is ordered into a row vector 'S' of dimension (1 × 648) to represent the optical flow based feature vector of the human action of a person in a given frame.

In interaction dataset, Z_k, the feature vector of the point placed at an angle of $5k$ degree, is of dimension (1 × 12). Therefore, the feature vector Z thus extracted is of dimension (72 × 12). Z is ordered into a row vector 'S' of dimension (1 × 864) to represent the optical flow based feature vector of the interacting persons in a given frame.

B. Human Activity Recognition using Multi-Class SVM

A spatiotemporal volume of 'L' frames is considered and the 'S' vectors of individual frames are concatenated to form the input to the SVM. Radial Basis Function (RBF) kernel based SVM classifier has been used herein. Since SVM classifier is a binary classifier, '*one-vs-all*' method has been used to train each classifier. With 'N' actions under consideration, 'N' SVM classifiers have been used as illustrated in Fig. 5.

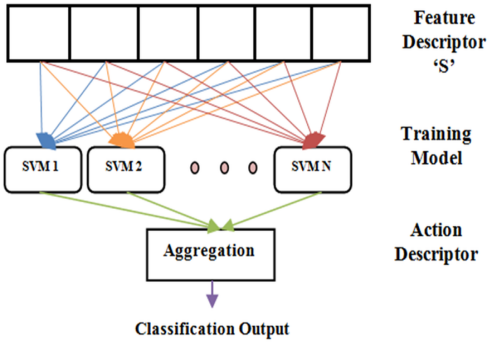

Fig. 5. Multi-class SVM model

IV. Experimental Results

The performance of the proposed method is evaluated based on the Weizmann, KTH and UT-interaction [16, 17, 18,] datasets. Weizmann and KTH datasets contain different human actions performed by individuals, whereas the UT-interaction dataset contains different interactions performed by two persons. Leave-One-Actor-Out approach (LOAO) has been used to investigate the efficiency of our proposed method with each of the datasets.

A. Human Action Recognition

Weizmann dataset consists of 90 video sequences, where 10 actions are performed by 9 different individuals. Representative examples of the 10 actions performed by different individuals in Weizmann dataset are shown in Fig. 6.

Fig. 6. Representative examples of actions included in Weizmann dataset

The video sequences in KTH dataset depict four different scenarios, consisting of 598 video sequences, where 6 different actions are performed by 25 individuals. Representative examples of the 6 actions performed by different individuals in KTH dataset are shown in Fig. 7.

The results on Weizmann and KTH datasets are shown in Table I and Table II respectively.

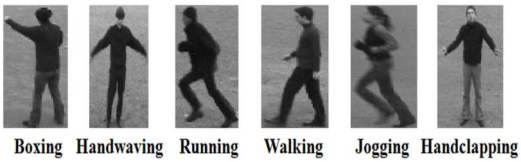

Fig. 7. Representative examples of actions included in KTH dataset

TABLE I. Performance of the Proposed Method on Weizmann Dataset

Methods	*Evaluation Approach*	*Year*	*Recognition Rate (%)*
Proposed method	Leave one out		95.69
C. Li et al. [12]	Leave one out	2014	98.89
M. J. Roshtkhari et al. [13]	Split	2013	98.7
Z. Zhang et al. [19]	Leave one out	2012	93.87
T. H. Thi et al. [20]	Split	2012	98.9
H. J. Seo et al. [21]	Split	2011	97.5

TABLE II. PERFORMANCE OF THE PROPOSED METHOD ON KTH DATASET

Methods	*Evaluation Approach*	*Year*	*Recognition Rate (%)*
Proposed method	Leave one out		94.62
C. Li et al. [12]	Leave one out	2014	95.48
M. J. Roshtkhari et al. [13]	Split	2013	95
T. H. Thi et al. [20]	Split	2012	94.67
Z. Zhang et al. [19]	Leave one out	2012	93.5
Z. Jiang et al. [22]	Leave one out	2012	95.77
H. J. Seo et al. [21]	Split	2011	95.1

B. Human Interaction Recognition

The UT-interaction dataset contains six classes of human-human interactions. The dataset is composed of 20 video sequences divided into two sets. Each of the UT interaction Sets (Set_1 and Set_2) consists of 10 video sequences. Fig. 8 and Fig. 9 shows the representative examples of human-human interactions performed in UT interaction Set_1 and UT interaction Set_2 respectively. The performance of the proposed algorithm on these datasets is given in Table III.

Fig. 8. Representative examples of human interactions in UT interaction Set_1

Fig. 9. Representative examples of human interactions in UT interaction Set_2

TABLE III. PERFORMANCE OF THE PROPOSED METHOD ON UT INTERACTION DATASET

Methods	*Year*	*Recognition Rate (%)*	
		UT Interaction Set_1	*UT Interaction Set_2*
Proposed method		92.7	90.21
G. Yu et al. [23]	2015	93.3	91.7
Y. S. Sefidgar et al. [4]	2015	93.3	90
S. Mukherjee et al. [24]	2014	91.66	81.66
M. Raptis et al. [7]	2013	93.3	-
Y. Zhang et al. [25]	2012	95	90

V. CONCLUSIONS

In this work, an efficient human action and interaction recognition method based on optical flow based feature set has been reported. By using the proposed feature descriptor with multi-class SVM classifier, recognition rates as high as 95.69% for Weizmann dataset and 94.62% for KTH dataset have been achieved for human action recognition. For UT interaction Set_1 and UT interaction Set_2, the recognition rates achieved are 92.7% and 90.21% respectively. The performance of the proposed method is comparable to that of the sophisticated algorithms reported in the recent literature. Optical flow is commonly used for foreground extraction. Therefore, the proposed recognition technique based on optical flow feature vector alleviates the computational complexity associated with the computation of an independent feature vector.

The results demonstrate that the method is simple but yet efficient. Therefore, the proposed technique can be regarded as a good choice for human action and interaction classification for video surveillance application.

References

[1] J.K. Aggarwal and M.S. Ryoo, "Human activity analysis - A review," ACM Comput. Surveys, vol. 43, pp. 1–47, April 2011.

[2] 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, pp. 1473 – 1488, September 2008.

[3] D. Weinland, R. Ronfard and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," Comput. Vis. Image Und., vol. 115, pp. 224-241, February 2011.

[4] Y.S. Sefidgar, A. Vahdat, S. Se and G. Mori, "Discriminative key-component models for interaction detection and recognition," Comput. Vis. Image Und., vol. 135, pp. 16–30, June 2015.

[5] K.N. El Houda Slimani, Y. Benezeth and F. Souami, "Human interaction recognition based on the co-occurence of visual words," in Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit. Workshops, Jun. 2014, pp. 461–466.

[6] Y. Kong, W. Liang, Z. Dong and Y. Jia, "Recognising human interaction from videos by a discriminative model," IET Comput. Vis., vol. 8, pp. 277 – 286, August 2014.

[7] M. Raptis and L. Sigal, "Poselet Key-Framing: A Model for Human Activity Recognition," in Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit., Jun. 2013, pp. 2650-2657.

[8] K. Yun, J. Honorio, D. Chattopadhyay, T.L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.Workshops, Jun. 2012, pp. 28–35.

[9] K.G. Derpanis, M. Sizintsev, K.J. Cannons and R.P. Wildes, "Action spotting and recognition based on a spatiotemporal orientation analysis," IEEE Trans. Patt. Anal. Mach. Intell., vol. 35, pp. 527–540, March 2013.

[10] I. Everts, J.C. van Gemert and T. Gevers, "Evaluation of color spatio-temporal interest points for human action recognition," IEEE Trans. Image Process., vol. 23, pp. 1569-1580, April 2014.

[11] B.Z. Yao, B.X. Nie, Z. Liu and S.C. Zhu, "Animated pose templates for modeling and detecting human actions," IEEE Trans. Patt. Anal. Mach. Intell., vol. 36, pp. 436-452, March 2014.

[12] C. Li, B. Su, J. Wang, H. Wang and Q. Zhang, "Human action recognition using multi-velocity STIPs and motion energy orientation histogram," J Inf. Sci. Eng., vol. 30, pp. 295–312, March 2014.

[13] M.J. Roshtkhari and M.D. Levine, "Human activity recognition in videos using a single example," Image Vis. Comput., vol. 31, pp. 864–876, November 2013.

[14] H. Li, A. Achim and D.R. Bull, "GMM-based efficient foreground detection with adaptive region update," in Proc. IEEE Int. Conf. Image Process., Nov. 2009, pp. 3181-3184.

[15] B.K. Horn and B.G. Schunck, "Determining optical flow," Artif. Intell., vol. 17, pp. 185–203, August1981.

[16] L. Gorelick, M. Blank, E. Shechtman, M. Irani and R. Basri, "Actions as space time shapes," IEEE Trans. Patt. Anal. Mach. Intell., vol. 29, pp. 2247-2253, December 2007.

[17] C. Schüldt, I. Laptev and B. Caputo, "Recognizing Human Actions: A Local SVM Approach," in Proc. IEEE Int. Conf. Pattern Recognit., vol.3, Aug. 2004, pp. 32-36.

[18] M.S. Ryoo and J.K. Aggarwal, "UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA)," in Proc. IEEE Int. Conf. Pattern Recognit. Workshops, vol. 2, Aug. 2010.

[19] Z. Zhang and D. Tao. " Slow feature analysis for human action recognition," IEEE Trans. Patt. Anal. Mach. Intell., vol. 34, pp. 436-450, March 2012.

[20] T.H. Thi, L. Cheng, J. Zhang, L. Wang, and S. Satoh, "Integrating local action elements for action analysis," Comput. Vis. Image Und., vol. 116, pp. 378–395, March 2012.

[21] H.J. Seo and P. Milanfar, "Action recognition from one example," IEEE Trans. Patt. Anal. Mach. Intell., vol. 33, pp. 867-882, May 2011.

[22] Z. Jiang, Z. Lin and L.S. Davis, "Recognizing human actions by learning and matching shape motion prototypes trees," IEEE Trans. Patt. Anal. Mach. Intell., vol. 34, pp. 533-547,March 2012.

[23] G. Yu, J. Yuan and Z. Liu, "Propagative hough voting for human activity detection and recognition," IEEE Trans. Circuits Syst. Video Technol., vol. 25, pp. 87-98, January 2015.

[24] S. Mukherjee, S.K. Biswas and D.P. Mukherjee, "Recognizing interactions between human performers by 'Dominating Pose Doublet'," Mach. Vision Appl., vol. 25, pp. 1033-1052, May 2014.

[25] Y. Zhang, X. Liu, M.C. Chang, W. Ge, and T. Chen, "Spatio-temporal phrases for activity recognition," in Proc. European Conf. Comput. Vis., Oct. 2012, pp. 707-721.