Human Interaction Recognition Using Improved Spatio-Temporal Features

M. Sivarathinabala and S. Abirami

Abstract Human Interaction Recognition (HIR) plays a major role in building intelligent video surveillance systems. In this paper, a new interaction recognition mechanism has been proposed to recognize the activity/interaction of the person with improved spatio-temporal feature extraction techniques robust against occlusion. In order to identify the interaction between two persons, tracking is necessary step to track the movement of the person. Next to tracking, local spatio temporal interest points have been detected using corner detector and the motion of the each corner points have been analysed using optical flow. Feature descriptor provides the motion information and the location of the body parts where the motion is exhibited in the blobs. Action has been predicted from the pose information and the temporal information from the optical flow. Hierarchical SVM (H-SVM) has been used to recognize the interaction and Occlusion of blobs gets determined based on the intersection of the region lying in that path. Performance of this system has been tested over different data sets and results seem to be promising.

Keywords Video surveillance · Blob tracking · Spatio temporal features · Interaction recognition

1 Introduction

Video Surveillance is one of the major research fields in video analytics and this has been mainly used for security purpose. Human Interaction Recognition becomes a key step towards understanding the human behavior with respect to the scenes.

M. Sivarathinabala (✉) · S. Abirami
Department of Information Science and Technology, College of Engineering,
Anna University, Chennai, India
e-mail: sivarathinabala@gmail.com

S. Abirami
e-mail: abirami_mr@yahoo.com

A. Nagar et al. (eds.), *Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics*, Smart Innovation, Systems and Technologies 43, DOI 10.1007/978-81-322-2538-6_20

Activity/Interaction recognition from the surveillance videos has been considered as challenging task. The situations such as background clutter, occlusion and illumination changes may cause difficulty in recognizing the activity of the person. Recognizing human interactions can be considered as the extension of single person action. In literatures [1–6], Action refers to the single person movement composed of multiple gestures such as arm/leg motion and torso motion, Activity refers to the combination of single or multiple people movement. Interaction may also refer to the activity that happens between two persons.

In videos, Action can be represented using global features as well as local features. Global features such as features considered from the entire image frame and local features have been considered from the local portion from the image frame. In literatures, Activity recognition approach relies on global feature representation or fusing all the local features or by computing the histogram. These approaches limit the performance of activity recognition under occlusion. To recognize interaction, body parts location and its movement during occlusion is still an open challenge. Thus we are motivated to propose a new middle level features to identify the activity/interaction even under occlusion.

Our Contribution lies in threefold: a new interaction recognition approach has been introduced to recognize/identify the activity of the person whenever there is crossover also. In feature extraction phase, Middle level features have been extracted to analyze the spatial and temporal relationships between two persons. The Hierarchical SVM classifier has been used to classify the interactions between two persons.

2 Related Works

Major works in the field of video analytics have been devoted in the object tracking and activity recognition phase and they are addressed in this section. Tracking the particular person/object under illumination conditions, occlusion and dynamic environments is still a challenging research. In general, Occlusion [2] can be classified into three categories: self occlusion, partial occlusion and total occlusion. In the previous work [7], to handle self and partial occlusion problems, a combination of blob tracking method and particle filter approach, has been employed by using Contour as shape representation and color as feature for tracking. In addition to this, blob splitting and merging approach has been attempted to identify occlusion.

Activity/Interaction of the tracked person has to be identified in order to increase the security in the environment. Human Interaction Recognition is crucial phase to understand from the nature of the moving persons. Arash Vahdat et al. [8] modelled activity with a sequence of key poses performed by the actors. Spatial arrangements between the actors are included in the model, for temporal ordering of the key poses. Chen et al. [9] proposed an automated reasoning based hierarchical framework for human activity recognition. This approach constructs a hierarchical structure for representing the composite activity. This structure is then transformed into logical formulas and rules, based on which the resolution based automated reasoning is applied to recognize the composite activity. There comes uncertainty in temporal and spatial relationship. This problem can be solved using mid level spatio-temporal features. Patron-Perez et al. [10] develop a per-person descriptor that uses head orientation and the local spatial and temporal context in a neighbourhood of each detected person. They also employed structured learning to capture spatial relationships between interacting individuals.

In literatures, many attempts have done to improve the interaction recognition rate using new feature extraction techniques and using new learning methods robustness to several complex situations such as background clutter, illumination changes and occlusion conditions. Here in this work, we have been provided solution to recognize the interaction between two persons under various situations.

3 Human Interaction Recognition

In this research, a special attempt has been made to identify the interaction between two persons from the tracked blobs and improved spatio-temporal features. Middle-level features [11], which connect local features and global features, are apparently suitable to represent complex activities. Our approach relies on corner detector and HOG descriptor to describe the pose of the person. Temporal features have been analyzed using optical flow for every corner points and in each of the body part. Pose information for head, arm, torso and leg has been obtained separately and clustered, in addition with the temporal features provided by the optical flow bins and then the activity/Interaction between two persons have been recognized. Semantics has been added with actions to recognize interactions without any confusion. Hierarchical SVM classifier has been used to classify the interaction from pose and activity classifiers. In this framework, spatio-temporal relationship has been maintained by constructing 5 bins in the optical flow descriptor [12, 13]. Real time human interaction recognition framework has been shown in Fig. 1.

4 Middle Level Features

Video Sequence has been represented by middle level features and their spatio-temporal relationships. In videos, local spatial temporal feature extraction has been widely used that involves interest point detection and feature description. From the tracked persons, local spatio-temporal features have been extracted directly to provide representation with respect to spatio-temporal shifts and scales. The spatio-temporal features can provide information about multiple motions in the scene. Feature detector usually selects the spatio-temporal locations and scales in the video. In this work, Harris corner detector [5] has been used to detect the spatio-temporal

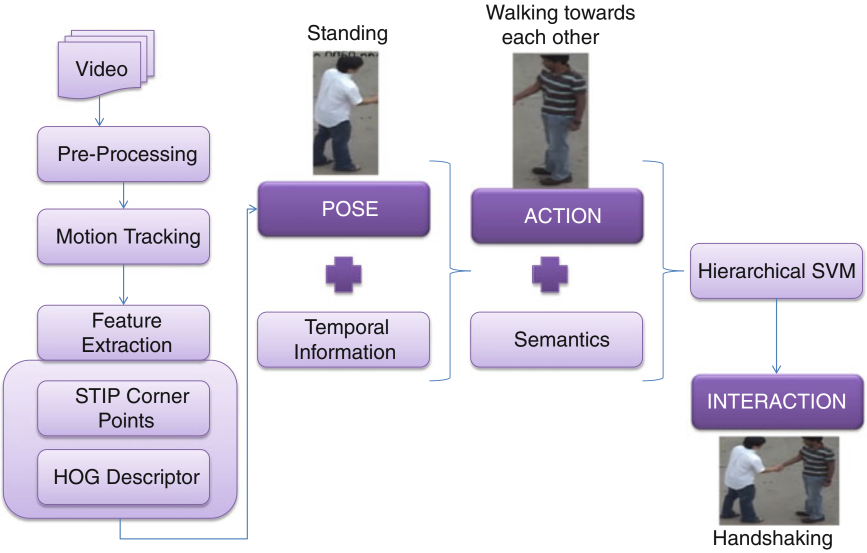

Fig. 1 Framework for human interaction recognition

points. Harris Corner detector, a simple detector that detects the interest points in the frame. Let the corner interest points be p_1, p_2, p_3, …, p_n described by the 2D co-ordinate. A densely sampled key point has been extracted that are stable in low resolution videos, clutter and also in fast motion and the motion has been analyzed from each corner points using optical flow algorithm. The distance between corresponding corner points of both the blobs makes separate spatial information. $d_t = [d[p_i] - d[p_j]]$ where i and j represents both the blobs, d_t represents the distance between the corresponding corner points in each frame at time t.

In general, detected interest points can be described using a suitable descriptor. Histogram of oriented gradients [10] that captures the spatial location of the body part and that is encoded by relative positions of HOG within the detector window. Motion information has been added with HOG i.e. a good feature combining this with the appearance (or) shape information makes a strong cue to represent the feature. Optical flow provides differential flow that will gives information about limb body relative motions. Temporal relations have been maintained throughout the video by analyzing optical flow from Hog descriptor. Optical flow motions has been differentiated into five different bins such as left, right, up, down and no motion.

Pose prediction is the first step in the interaction recognition and it can be done using HOG and SVM models. Posture has been estimated separately for the body parts such as head, torso, arm and leg. HOG descriptor builds a body part model pre trains the body parts into head, torso, arm and leg and SVM classification has been performed. Each forms a separate cluster and distance between the neighborhood points has been analyzed. $H_p = \{H_{1i}, H_{1j}, H_{2i}, H_{2j}, \ldots, H_{ni}, H_{nj}\}$; $A_p = \{A_{1i}, A_{1j},$

A_{2i}, A_{2j}, …, A_{ni}, A_{nj}}; T_p = {T_{1i}, T_{1j}, T_{2i}, T_{2j}, …, T_{ni}, T_{nj}}; L_p= {L_{1i}, L_{1j}, L_{2i}, L_{2j}, …, L_{ni}, L_{nj}} Where H_p, A_p, T_p and L_p represents the Head, Arm, Torso and Leg posture. $d_{Hp} = [d[H_{1i}]_t - d[H_{1i}]_{t+1}]$ where t and t + 1 represents the current frame and next frame. Similarly distance has been calculated and pose has been predicted. Next to the pose, actions have been analyzed from the spatial and temporal informations. Five classes of the actions such as walking, running, boxing, touching and hand waving that are trained and the actions have been classified.

5 Interaction Modeling

In each frame we have a set of human body detections X = [x … x_M]. Each detection xi = [l_x, l_y, s, q, v], has information about its corner location (l_x, l_y), scale (s), discrete body part orientation (q), and v represents SVM classification scores. Associated with each frame is a label Y = [y_1 … $y_{k,}$ y_c]. This label is formed by a class label y_i and K for each detection (where K is the number of interaction classes, with 0 representing the no-interaction class) and a configuration label y_c that serves as an index for one of the valid pairings of detections. (i, j) indicates that detection i is interacting with detection j and the 0 index means there is no interaction. The match between an input X and a labeling Y has been measured by the following cost function:

$$S(X,Y) = \sum\nolimits_i^m \alpha_{yiqi} v_{yi} + \sum \alpha_{yiqi} + \sum\nolimits_{(i,j)\in P_{yc}} \left(\delta_{ij}\beta_{yi}q_i + \delta_{ji}\beta_{yj}q_j\right) \tag{1}$$

where v_{yi} is the SVM classification score for class y_i of detection i, P_{yc} is the set of valid pairs defined by configuration index y_c, d_{ij} and d_{ji} are indicator vectors codifying the relative location of detection j with respect to detection i . y_iq_i are scalar weighting and bias parameters that measure the confidence that we have in the SVM score of class y_i when the discrete orientation of the body part is q_i. b_{yiqi} is a vector that weights each spatial configuration given a class label and discrete head orientation. Once the weights are learnt, we can find the label that maximizes the cost function by exhaustive search, which is possible given the small number of interaction classes and number of people in each frame. Interaction modeling has been shown in the Fig. 2a, b the handshake interaction model has been given.

The person 1 on the left side walks person 2 on the right, and they shaking their hands then rapidly depart. The poses such as arm stretch and arm stay and action such as walking has been correctly classified. From the optical flow bins, the directions has been identified such as the person moving right or left direction. A semantic description for the handshaking action has been shown in Fig. 2b. Along with the semantic descriptions, Interactions has been classified using hierarchical SVM. Another example is of kicking interaction has been shown in Fig. 3. Here in this case, leg posture has been considered. The poses such as leg stretch and leg stay and action such as walking has been correctly classified. From the optical flow bins, the directions has been identified such as the person moving right or left direction.

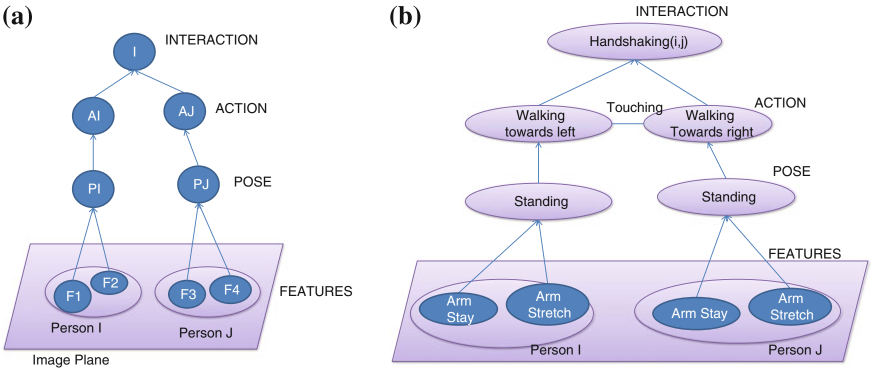

Fig. 2 **a** Interaction modeling and **b** handshaking interaction model

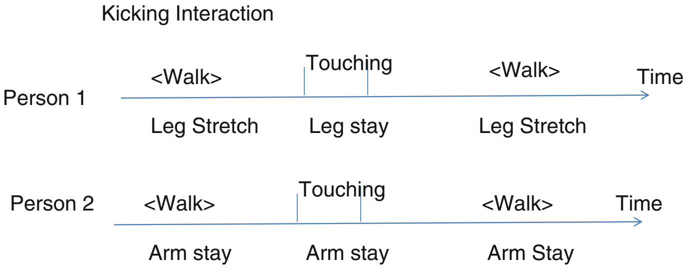

Fig. 3 Semantic descriptions of kicking interaction

Hierarchical Support Vector Machine (H-SVM) has been used to classify the interaction between two persons. To perform human interaction recognition, we fuse the features at two levels, (1) the output of the pose classifier and activity classifier has been concatenated and given as the input to classifier, and (2) the classifiers for the two sources are trained separately and classifier combination is performed subsequently to generate final result. To combine the classifier outputs of both the spatial and temporal features, classifier outputs have been interpreted as probability measure.

6 Results and Discussion

The implementation of this object tracking system has been done using MATLAB (Version2013a). MATLAB is a high performance language for technical computing. The input videos are taken from UT interaction dataset [14] and BIT interaction dataset [15]. The proposed algorithm have been applied and tested over many

Fig. 4 Sample frames from UT and BIT interaction datasets

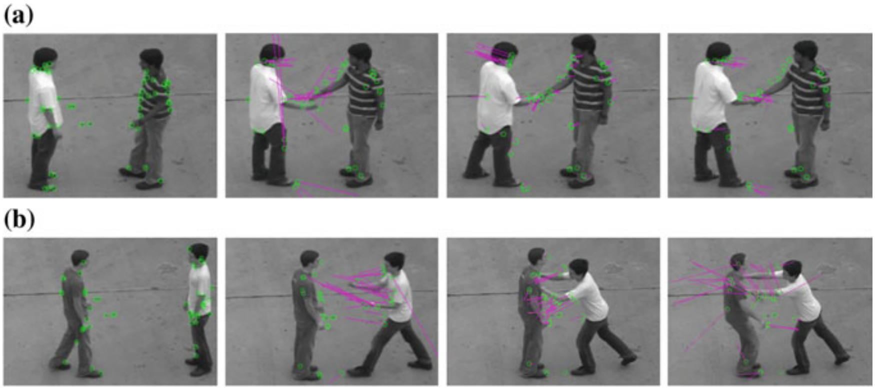

Fig. 5 **a** Handshaking—interaction. **b** Pushing—interaction

different test cases and two of the scenarios have been shown here. The sample frames from the datasets has been shown in Fig. 4.

Figure 5a, b represents the handshaking and pushing interaction. HS represents Hand Shake and HF represents High Five interaction. Table 1a, b shows the confusion matrix for the UT and BIT Interaction datasets respectively. It is evident from the table that, the seven interactions classes has been trained and tested using hierarchical SVM. Our method using midlevel features obtained the accuracy of 90.1 % in UT Interaction dataset (set1) and 88.9 % in BIT Interaction dataset. Our approach has been compared with the existing methods used in interaction modelling shown in Table 2.

Table 1 Confusion matrix of UT interaction dataset and BIT interaction respectively

a					b				
	HS	Hug	Push	Punch		HF	Kick	Push	Pat
HS	1	X	X	X	HF	0.94	X	0.06	X
Hug	X	1	X	X	Kick	X	0.95	0.02	0.03
Push	X	X	0.98	0.02	Push	X	0.06	0.92	0.02
Punch	X	X	0.05	0.95	Pat	0.02	0.01	0.08	0.89

Table 2 Comparison with other existing methods

Previous works	Dataset considered	Accuracy (%)
BOW [16]	UT interaction (set1)	58.20
Visual co-occurrence [17]	UT interaction (set1)	40.63
Our method	UT interaction (set1)	90.1
	BIT interaction	88.9

7 Conclusion

In this research, an automated interaction recognition system has been developed using new spatio-temporal features which are called as mid level features. The features have been considered from the tracked blob in the spatial and also in temporal domain. We have been integrated spatio-temporal relationship between every consecutive frame in the video sequence. The activities of the each person have been identified and the activities/Interaction that happened between two persons has been recognized through the midlevel features and high level semantic descriptions. The intersecting regions between the potential detects the occlusion states. The proposed algorithm has been tested over BIT interaction dataset and UT interaction dataset. This system has the ability to recognize the interaction of the person even if there is a person/object crossover also. In future, this system could be extended along with the detection of heavy occlusion and multiple objects tracking too.

References

1. Yilmaz, A., Javed, O., Shah, M.: Object tracking :a survey. ACM Comput. Surv. **38**(4) (2006) (Article no. 13)
2. Ryoo, Agarwal, J.K.: Human activity analysis: a survey. ACM Comput. Surv. **43**(3), 16:1–16:43 (2011)
3. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio temporal features. In: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72 (2005)
4. Gowsikhaa, D., Manjunath, Abirami, S.: Suspicious human activity detection from surveillance videos. Int. J. Internet Distrib. Comput. Syst. **2**(2), 141–149 (2012)
5. Gowshikaa, D., Abirami, S., Baskaran, R.: Automated human behaviour analysis from surveillance videos: a survey. Artif. Intell. Rev. (1046). doi:10.1007/s2-012-9341-3(2012)
6. Gowsikhaa, D., Abirami, S., Baskaran, R.: Construction of image ontology using low level features for image retrieval. In: Proceedings of the International Conference on Computer Communication and Informatics, pp. 129–134 (2012)
7. Sivarathinabala, M., Abirami, S.: Motion tracking of humans under occlusion using blobs. Advanced Computing, Networking and Informatics, vol 1. Smart Innovation, Systems and Technologies, vol. 27 , pp. 251–258 (2014)

8. Vahdat, A., Gao, B., Ranjbar, M., Mori, G.: A discriminative key pose sequence model for recognizing human interactions. In: Proceedings of the IEEE International Conference on Computer Vision (2011)
9. Chen, S., Liu, J., Wang, H.: A hierarchical human activity recognition framework based on automated reasoning. In: The Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, pp. 3495–3499 (2013)
10. Patron-Perez, A., Marszalek, M., Zisserman, A., Reid, I.: High five: recognising human interactions in TV shows. In: British Machine Vision Conference (2010)
11. Bruhn, A., weickert, J.: Lucas/Kanade meets Horn/Schunck: combining local and global optic flow methods. Int. J. Comput. Vision **61**(3), 211–231 (2005)
12. Jain, M., Jégou, H., Bouthemy, P.: Better exploiting motion for better action recognition. In: The Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 2555–2562 (2013)
13. Huang, K., Wang, S., Tan, T., Maybank, S.: Human behaviour analysis based on new motion descriptor. In: IEEE Transactions on Circuits and Systems for Video Technology (2009)
14. Ryoo, M.S., Aggarwal, J.K: UT Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA) (2010)
15. Yu Kong and Yunde Jia and Yun Fu, "Learning Human Interaction by Interactive Phrases", Book title,European Conference on Computer Vision,pp.300–313, vol.7572,2012
16. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: local SVM approach. In: The Proceedings of ICPR, Cambridge (2004)
17. Nour el houda Slimani, K., benezeth, Y., Souami, F.: Human interaction recognition based on the co-occurence of visual words. In: The Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 455–460 (2014)