Journal of Electronic Imaging

JElectronicImaging.org

Criminal action recognition using spatiotemporal human motion acceleration descriptor

Abinta Mehmood Mir
Muhammad Haroon Yousaf
Hassan Dawood

Abinta Mehmood Mir, Muhammad Haroon Yousaf, Hassan Dawood, "Criminal action recognition using spatiotemporal human motion acceleration descriptor," *J. Electron. Imaging* **27**(6), 063016 (2018), doi: 10.1117/1.JEI.27.6.063016.

Criminal action recognition using spatiotemporal human motion acceleration descriptor

Abinta Mehmood Mir,[a] Muhammad Haroon Yousaf,[a,*] and Hassan Dawood[b]
[a]University of Engineering and Technology, Centre for Computer Vision Research (C2VR), Department of Computer Engineering, Taxila, Pakistan
[b]University of Engineering and Technology, Department of Software Engineering, Taxila, Pakistan

Abstract. Video surveillance systems have become one of the most useful entities in our routine life. Surveillance videos contain plenty of visual information about criminal actions happening in the field-of-view. With the increase of criminal activities, it is mandatory to develop the accurate criminal recognition system. Our paper aims to propose and evaluate action recognition system for the recognition of criminal actions. First, a descriptor is proposed as spatiotemporal human motion acceleration (ST-HMA) over improved dense trajectories (IDT) framework to correctly recognize the criminal actions. Second, a hybrid dataset is developed by the combination of criminal activities, e.g., fight, kick, push, punch, shoot gun, and sword fighting collected from state-of-the-art datasets named as hybrid criminal action (HCA) dataset. The dataset covers the common on-street criminal action poses. We have also evaluated different descriptors over the IDT framework. The achieved accuracies per class are 92.85%, 92.85%, 93.33%, 96.16% for kick, push, punch and fight actions, respectively. Experimental results show that ST-HMA on IDT framework gives better results than HMA descriptor in edge trajectory framework. The proposed framework also achieved high average accuracy rate of 80.89% for ST-HMA descriptor over IDT. Different descriptor applied over IDT also shows good action recognition accuracy for HCA dataset. [DOI: 10.1117/1.JEI.27.6.063016]

Keywords: action recognition; dense trajectories; criminal actions; spatiotemporal human motion acceleration descriptor.

Paper 180363 received Apr. 24, 2018; accepted for publication Nov. 13, 2018; published online Dec. 8, 2018.

1 Introduction

For security reasons, surveillance cameras are installed on public places, homes, educational institutes, markets, hospitals, etc. to capture real-time events. Action recognition literature mostly revolves around detecting and recognizing simple human actions occurring in daily life scenarios like clapping, walking, jogging, meeting, etc.[1] In the last few years, researchers paid attention toward human actions/activities in complex scenarios, e.g., sports, patient-care, etc. Actions related to any criminal activity are somehow different from above-mentioned actions as crime is defined as an act, which is harmful to an individual, a community, or society. Criminal actions/activities are in many diversified forms, including domestic homicide, robbery, burglary, cybercrime, etc., which led to the development of safety and anticrime systems.[2] Criminal actions/activities recognition is comparatively less studied in the literature, and we can hardly find any action recognition dataset having significant criminal actions. These actions have key importance among captured visual data. Criminal action is said to occur between a potential victim and a motivated offender when they meet in the absence of a guardian. The motivation to offend decreases when the offender is aware that they may be watched.[3] Therefore, the nature of human gestures, pose, and actions is clearly different as compared to other kinds of actions, which makes this task different and challenging as well. Crime rate is increasing at an alarming rate throughout the world and demand for autonomous recognition system of such activities is increasing day by day. Criminal actions revolve around humans, actions, and behaviors. Human action recognition has been extensively studied for many years due to its potential applications, such as human–computer interaction, security-related surveillance systems, sudden event analysis, sports video analyses, automatic retrieval, and many more.[4–6]

In recent years, trajectory-based methods[7–10] have gained much attention for human action recognition due to long-term motion information captured by them. Motion information is extracted over spatiotemporal trajectories and then described by using trajectory descriptors. Many descriptors have been proposed for dense trajectory (DT) and improved dense trajectory (IDT) framework like histogram of oriented gradient (HOG),[11] histogram of optical flow (HOF),[11] motion boundary histogram (MBH) with respect to x-axis MBHx,[8] MBH with respect to y-axis MBHy,[8] co-occurrence histogram of oriented gradient (CoHOG),[6] histogram of spatial gradient of information,[12] etc., to characterize trajectory shape, motion, and appearance information. Wang and Qi[13] extracted spatiotemporal edge trajectories and introduce descriptor human motion acceleration (HMA) to encode relative motion in temporal domain of human action. The drawback of using edge trajectories is that camera motion causes false edge localization in edge-based trajectories that affect the accuracy of action recognition system.

In human action recognition domain, significant progress has been made with respect to testing of action recognition algorithms on state-of-the-art datasets. Many of the datasets used for action recognition are captured in simple and constrained environments. Higher accuracy has been achieved over these datasets, e.g., KTH,[14] Weizmen,[15] etc.

*Address all correspondence to Muhammad Haroon Yousaf, E-mail: haroon.yousaf@uettaxila.edu.pk

However, datasets containing more complex and realistic actions, e.g., Hollywood,[11] HMDB51,[16] and YouTube,[11] are still considered challenging datasets.[17] Recognition accuracy achieved so far in literature for these datasets is not too high, thus there is a demand for more exploration for better recognition of high level complex action analysis. There are few datasets having violent or criminal actions like hockey fight events[1] and real life fight events.[16] However, few datasets contain video sequences of criminal actions that usually happen in outdoor environment like on-roads, etc. Descriptors characterizing shape and appearance lack motion information. For complex set of datasets, these descriptors are used in combination with motion descriptors for better recognition rates. Motion descriptors often increase the recognition rates, but recognition rates are affected due to camera motion. Datasets having videos taken from YouTube, movies, tv broadcast, etc., contain challenges like camera motion, viewpoint changes, resolution changes, and there is no single solution that can be applied to any given dataset. DT and IDT are giving competitive results in action recognition literature but still recognition results are low for complex action datasets. There is still a room for exploration of different descriptors over IDT framework especially for complex datasets with low recognition rates. The main contributions are as follows.

This paper proposes a scheme for criminal action recognition by using spatiotemporal human motion acceleration (ST-HMA) descriptor on IDT framework. As there are no particular datasets for criminal action video sequences, a dataset with criminal activities is developed by getting the different criminal scenes from different video sequences named as hybrid criminal action (HCA) dataset. For HCA dataset, we collected video sequences of different actions from state-of-the-art action datasets, which are considered as criminal activity, e.g., fight, kick, push, punch, shoot gun, and sword fighting. Selecting action videos from multiple different datasets made the problem more challenging due to variation with respect to objects appearance, background, illumination, and view-point. The proposed ST-HMA descriptor with IDT framework is evaluated over HCA dataset along with other standard descriptors like HOG,[11] HOF,[11] MBHx,[8] MBHy,[8] and trajectory.[7] The rest of the paper is organized as follows: Sec. 2 provides the detail on related work, Sec. 3 presents an overview of the proposed methodology, Sec. 4 is focused on experimental results and discussions in detail, and finally, Sec. 5 concludes the paper.

2 Related Work

Low-level local features have been discussed a lot in action recognition literature because these features are robust to background clutter, illumination changes, etc. These local features have two parts, detector and descriptor: detector for detecting the local region and descriptor for describing detected region. Many feature detectors have been proposed so far, such as 3-D Harris,[18] 3-D Hessian,[19] cuboid,[20] DT,[7] IDT,[9] and space–time pairwise trajectories.[10] The 3-D Hessian is a spatiotemporal extension of Hessian saliency measure used for blob detection in image and cuboid descriptors relied on temporal Gabor filter. In Ref. 21, a set of similar key points-based trajectories are grouped and mid-level components are generated, and in Ref. 22, structure similarity between features extracted is measured for action recognition. In Ref. 23, a descriptor is proposed to enhance the local intensity order pattern feature descriptor by using global matching. Among local features, dense sampling has shown better performance over sparse interest point for image classification.[24,25] Inspired by this success of dense sampling, Wang et al.[7] proposed the concept of DT. Dense points are sampled and tracked using state-of-the-art optical flow algorithm. Similarly, many local descriptors have been proposed to represent volume extracted along these detectors, such as HOG[11] descriptor encodes static appearance, 3-D histogram of oriented gradient (HOG3D),[26] extended speeded-up robust features (ESURF),[19] HOF characterize local motion information,[11] and MBH descriptor is robust to camera motion and characterize relative motion information.[8] These descriptors represent a certain visual pattern like appearance, motion, and motion boundary. To estimate camera motion and to improve results of DT over more challenging datasets, IDT method has been proposed. Trajectories consistent with estimated camera motion are removed to significantly improve the result of motion-based descriptors. IDT method gained the performance of 3% on HMDB51 with 57.2% accuracy.[9] One concept is to establish space–time relationship between different trajectories to exploit local spatial and structure information around trajectories. This method gained an average accuracy of 60.57% on HMDB51 dataset.[10] Among all these, spatiotemporal interest points (STIPs)[18] and IDT[9] are widely used because of their good performance. STIPs result in a set of sparse interest points and extract two kinds of descriptors, HOG and HOF. IDTs integrate much richer set of low-level visual cues compared with STIPs. Local spatiotemporal features[7,11,18] with BOVW framework have been used a lot and much progress has been made over many state-of-the-art datasets like KTH,[27] Weizman,[15] HMDB51,[16] etc.

After the success of DT-based methods, many surveys related to main representations and aggregated methods along dense trajectories are published.[5,20,28] Similarly, evaluation of different descriptors and their fusion over different datasets have been studied a lot. Kataoka et al.[6] evaluated 13 descriptors characterizing shape, motion, texture, trajectory, and co-occurrence. They found that co-occurrence based features characterize a significant representation over DT. They also tested results of feature concatenation and found that concatenation of highly selected features gives better results than integration of all 13. In Ref. 13, a descriptor named HMA has been proposed over edge-based dense trajectories. They tracked edge points and only considered trajectories related to the boundary of the action related area. Edge-based trajectory methods give comparable results with an HMA descriptor than DT method,[7] but on complex datasets like HMDB51 and IDT methods[9] are still better than edged-trajectories with HMA descriptor. HMA has been tested on seven datasets including simple and complex action datasets. Chen and Corso[29] exploited implicit intentional movement by analyzing the properties of IDT. They extracted space–time graphs from IDT. Action proposals have been computed by clustering the graphs to extract implicit intentional movement. They performed recognition on these action proposals for testing their approach. They achieved state-of-the-art performance on MSR-II multiaction benchmark.

Much progress has been made over datasets containing simple actions far from realistic scenarios like KTH,[14]

CASIA, Weizman,[15] or IXMAX dataset.[30] Recognition rates on these datasets are very high. In Ref. 14, HMDB51 dataset containing real world complexity has been designed. Similarly, Hollywood2 and UCF50[1,31] are realistic datasets taken from YouTube and real movies. In action recognition, community violent actions and aggressive behaviors have been less studied. For evaluation of such type of actions in Ref. 31, the new dataset proposed containing fight events taken from hockey games of the National Hockey League and from action movies. They tested bag of visual words (BOVW) framework with two available descriptors STIP and MoSIFT and obtained 90% accuracy for fight recognition. In Ref. 32, real-life events dyadic interaction dataset (Re-DID) collected from YouTube containing urban fight situations has proposed. They used DT framework and introduced proxemics information by using interpersonal space to detect interaction and to discriminate different types of behaviors. Recognition rate for Re-DID dataset is up to 73.96%. In Ref. 33, unconstrained dataset WEB-interaction has proposed to represent more realistic scenes with more challenging environment and to evaluate methods. The best average precision result on this dataset is 44.2%. For these realistic datasets, there is still a room for improvement.

Success of IDT-based methods demand for evaluation of more set of descriptors to encode characteristics of trajectory in a better way. Study of these techniques and evaluation of descriptors for a complex set of datasets are very challenging tasks and demand for more exploration.

3 Methodology

A general overview of the proposed human action recognition system is shown in Fig. 1. In the training phase, features are extracted from training videos. For feature extraction, we have used IDT.[9] The descriptors are computed within the space–time volume aligned with trajectories. Then, ST-HMA descriptor was tested over IDT. Low-level extracted features have high dimensions, which results in a great challenge for the subsequent step of clustering, so PCA is used for feature reduction. BOVW model is used for video representation. Features are encoded, and a classifier is trained using support vector machine (SVM) for training and testing. Features are encoded based on codebook construction in training phase and a classifier is used to obtain testing videos label.

3.1 *Trajectory Extraction*

To obtain dense trajectories, points are densely sampled in each frame of a video and these dense points are tracked in dense optical flow field. Dense sampling guarantees a good combination of feature points and dense optical flow improves the characteristics of trajectories to increase the efficiency. We have evaluated IDT-based features. Steps involved for IDT extraction are as follows.

3.1.1 *Dense sampling*

Dense sampled points on a grid spaced by W pixels represent the feature points. The dense points are sampled on each spatial scale of video frames. For scaling, maximum eight scales are chosen depending on the resolution of video. Each frame is scaled by a factor of $1/\sqrt{2}$ as in Ref. 7. These dense points are tracked on each spatial scale separately. The points are removed from homogeneous image area without any structure. Sample dense points are shown in red in Fig. 2 for an example video frame of kick action.

3.1.2 *Dense trajectories*

Extracted feature points are tracked by using optical flow computation between two consecutive frames It and It+1.

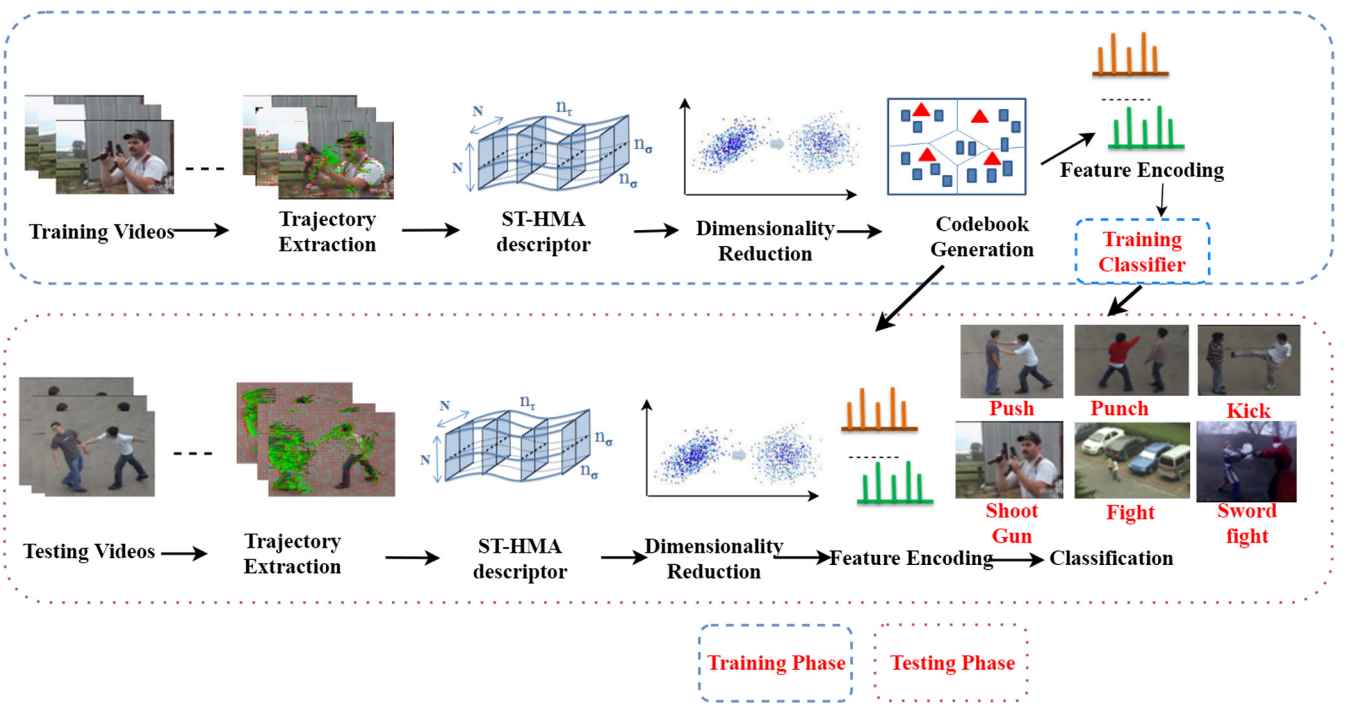

Fig. 1 Overview of general action recognition framework proposed.

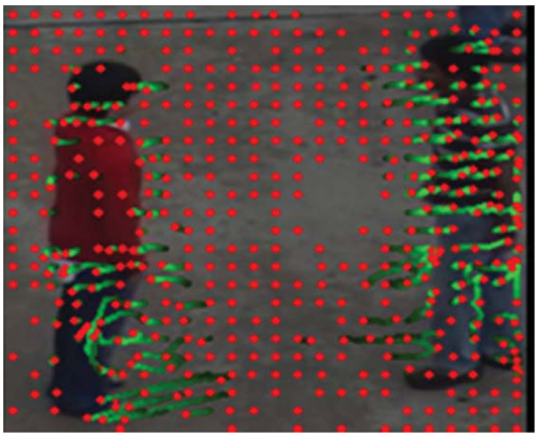

Fig. 2 Dense sampling for an example video frame from kick action.

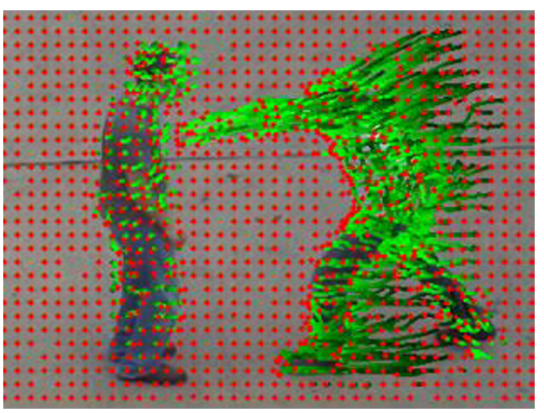

Fig. 3 Dense trajectories extraction for an example frame from kick action.

For a point Pt in video frame It, its tracked position in the next frame is estimated as follows:

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * w_t), \tag{1}$$

where $wt = (ut, vt)$ is dense optical flow computed with respect to the next frame $\mathrm{I}t + 1$, and ut and vt are horizontal and vertical components of optical flow. M is the kernel function of size 3×3. These points of consecutive frames are concatenated to represent trajectories: $\mathrm{Pt}, \mathrm{Pt} + 1$, $\mathrm{Pt} + 2, \ldots$. The length of tracked trajectories is limited to L number of frames because trajectories can drift from their initial location during the tracking process. Static trajectories are removed because they do not provide any useful information. Extracted dense trajectories for kick action are shown in green in Fig. 3.

3.1.3 *Improved dense trajectories*

Dense trajectories are improved by removing irrelevant trajectories due to camera motion. In realistic videos, many irrelevant trajectories are available in the background due to camera motion. The main purpose is to remove these irrelevant trajectories and keep only the object of interest. We estimated camera motion, and trajectories consistent with camera motion are removed as in Ref. 9. To estimate the global background motion homography, the same approach is used as in Ref. 34. For homography computation, the first step is to find correspondence between two frames. Correspondence is found by combining two approaches. First approach is to extract SURF[35] features and match them based on nearest neighbor rule then second approach is sampling motion vectors from optical flow field. Then, homography is estimated using RANSAC[36] to remove background camera motion. To enhance the estimation of homography, matches due to human motion are removed by using state-of-the-art human detector.[37]

3.2 *Trajectory Descriptor*

To characterize shape, appearance, and motion information, we have computed six descriptors aligned within space–time volume along trajectories: HOG, HOF, MBHx, MBHy, and trajectory shape descriptor, as in Ref. 9. We have proposed modified ST-HMA descriptor. ST-HMA descriptor is primarily based upon HMA.[13] In Ref. 13, edge points across video frames were tracked to extract edge-based trajectories and HMA was extracted as temporal relative motion of actions. HMA was computed by taking temporal derivative of optical flow for all points in spatial neighborhood only. Spatial neighborhood was divided into 4×4 cells. Then, orientation histograms of these cells were concatenated. For HMA descriptor of trajectory, a summation of all trajectory points was obtained. We deal with the spatiotemporal characteristic of trajectory-based descriptors and embed the structural information by diving in temporal domain, which was not discussed considered in the descriptor presented in Ref. 13. We have proposed ST-HMA descriptor, as shown in Fig. 4. ST-HMA is computed by taking temporal derivative of optical flow for spatiotemporal volume centered along trajectory.

As shown in Fig. 4, video frames for a trajectory are represented as $It, It + 1, \ldots, It + L$. Optical flow is computed for video frames. Optical flow is represented as horizontal component "u" and vertical component "v." Horizontal component of optical flow is represented in second row of Fig. 4 as $ut, ut + 1, \ldots, ut + L$. ST-HMA descriptor is obtained from spatiotemporal volume centered along trajectory. For every point in the spatial neighborhood of trajectory point, temporal derivative of optical flow is calculated. In Fig. 4, red square represents spatial neighborhood along trajectory point (xt, yt). For every point in neighborhood, its temporal derivative for horizontal component of optical flow is computed as follows:

$$\Delta u_t = u_{t+1} - u_t \tag{2}$$

Similarly, for vertical component:

$$\Delta v_t = v_{t+1} - v_t. \tag{3}$$

We embed structure information by dividing volume centered along trajectory into spatiotemporal grid of size $n\sigma \times n\sigma \times n\tau$. These cells are represented in the third row of Fig. 4. ST-HMA is computed for each cell of grid. Orientations are quantized into 8 bins in terms of magnitude for histogram computation. Final descriptor is concatenation of these descriptors. Size of STHMA descriptor is $n\sigma \times n\sigma \times n\tau \times 8$, where $n\sigma = 2$ and $n\tau = 3$ as in IDT framework.[9]

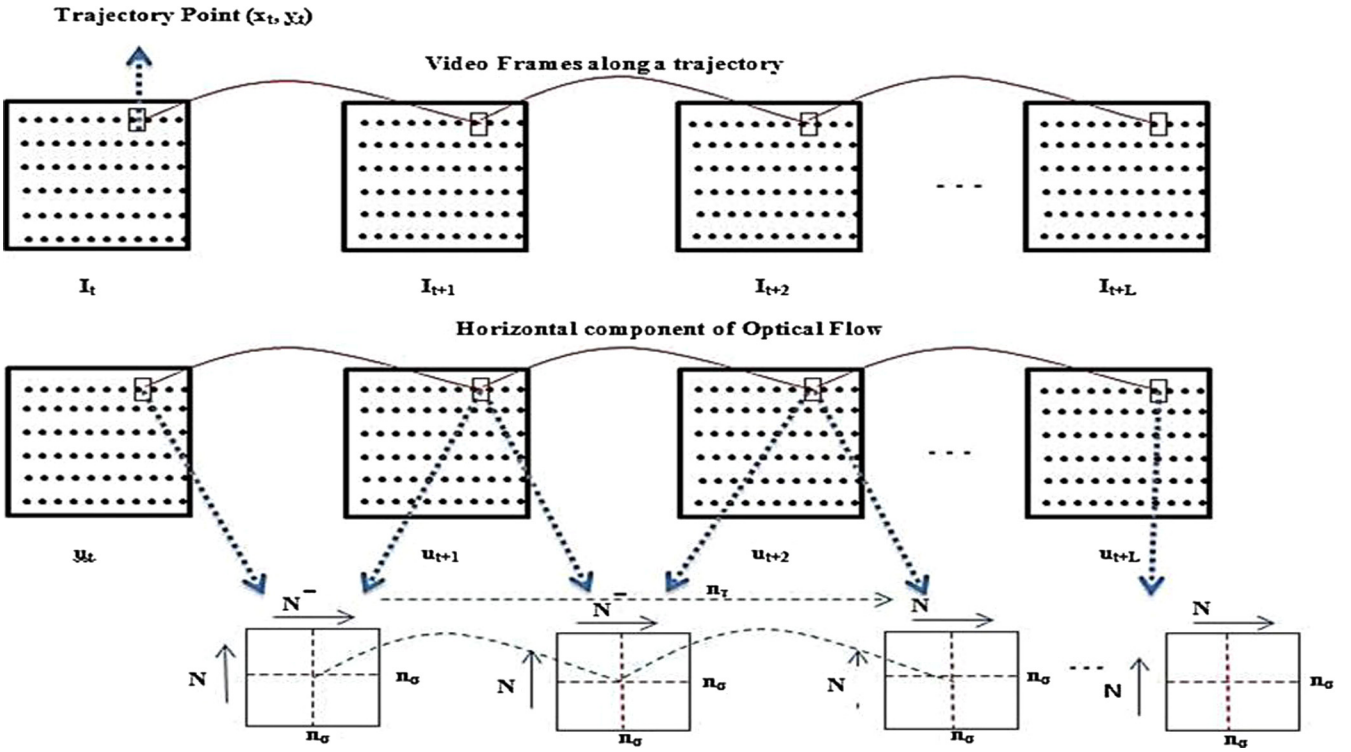

Fig. 4 STHMA descriptor calculation along space–time volume aligned within a trajectory.

3.3 *Dimensionality Reduction*

We have used PCA for feature's dimensionality reduction,[38] which is used to reduce data from n-dimensions to k-dimensional vector. Dimensionality of each extracted descriptor is reduced to half. For HOG, MBHx, MBHy, and ST-HMA, dimensionality is reduced from 96 to 48, and, for HOF, dimensionality is reduced from 108 to 54.

3.4 *Codebook Generation and Features Encoding*

After dimensionality reduction, training features are used to construct visual vocabulary. BOVW framework is adapted for codebook learning.[20] For codebook generation, k-mean clustering is widely used in literature.[7,9] Feature space is divided into k regions each of which is represented by its center called code-word. Nearest neighbor strategy is used to quantize the testing and training features by assigning to the closest visual word in the codebook, based on Euclidean distance between cluster center and input descriptor. After that, BOVW histogram is constructed by counting the frequency of each visual word. This results in a global video-level representation from local features. Dimension of this vector representation is equal to the number of visual words. In our case, codebook size is set to 2000.

3.5 *Classification*

For classification, multiclass linear SVM is used.[14] A linear classifier is selected because nonlinear SVM has more computational cost in training phase.

We have a trained classifier using quantized training features and training labels and then labels are predicted for quantized testing features.

For a training data $\{(z_i, y_i), y_i \in Y\{1, \ldots, L\}\}$, a linear SVM learns L linear functions $\{w_c^T z | c \in \mathcal{Y}\}$, where class for a test datum z is predicted as follows:

$$y = \max_{c \in \mathcal{Y}} w_c^T z. \tag{4}$$

One-against-all strategy is used to train L binary linear SVMs for the prediction of class labels.

4 Results and Discussions

In this section, we discussed about HCA dataset, experimental setup, and evaluation of the proposed descriptor on IDT framework.

4.1 *HCA Dataset*

The criminal action videos are collected from different state-of-the-art datasets to make HCA dataset. Criminal actions can be divided into different categories, i.e., offence against the person, violent offence, sexual offence, offence against property, etc.[39] Offence against a person is defined as physical harm directly to a person or force being applied to a person.[39] In this sense, we have included actions of kick, push, punch, and fight in this category. In violent actions, offender threatens to use force against victim. These actions may or may not include some weapons.[40] Gun shooting and violent fight actions are categorized as violent actions. All the action sequences have full or partial similarity to the nature of action in realistic scenario.

This dataset contains video sequences for six criminal sorts of actions. Each action category contains set of videos depicting that particular action. These videos are single action videos. We did not include multiple action videos in a single video sequence because temporal segmentation of action videos is itself a challenging task. Each video is recorded with different actors performing particular action and environmental conditions. There are in total 302 videos.

Details of selected criminal actions and number of videos are given in Table 1. These combined actions are taken from

Table 1 HCA dataset detail with challenges.

Actions	Dataset	Dataset challenges
Fight	CASIA[41]	Complex and nonstatic background with uncontrolled illumination conditions
Kick	UT-interaction[42]	Complex human activities with realistic settings
		Set 1 videos with static background
		Set 2 videos with camera jitter and background motion
Push	UT-interaction[42]	Complex human activities with realistic settings
		Set 1 videos with static background
		Set 2 videos with camera jitter and background motion
Punch	UT-interaction[42]	Complex human activities with realistic settings
		Set 1 videos with static background
		Set 2 videos with camera jitter and background motion
Gun shooting	HMDB51[16]	Videos ranging from digitized movies to YouTube
		Challenging dataset with large intra-class variability, camera motion, low quality, occlusion, nonstatic background, changes in position and view point
Sword fighting	HMDB51[16]	Videos ranging from digitized movies to YouTube
		Challenging dataset with large intraclass variability, camera motion, low quality, occlusion, nonstatic background, changes in position and view point

a diverse set of environments and contain a large set of challenges like intraclass variability, camera motion, low quality, occlusion, nonstatic background, changes in position and view point, etc. These criminal actions also contain large interclass variability because some actions are from simple background having less complexity, while some of these actions are from complex set of environments, as in HMDB51 action classes. These videos also contain a diverse set of actors forming actions as well. Dealing with such type of hybrid complex action dataset is very challenging task.

Sample video frames taken from all action classes for HCA dataset are shown in Fig. 5. Fight action is taken from CASIA dataset[41] with complex and nonstatic background. Kick, push, and punch are taken from UT-interaction. These actions are captured in a realistic environment; shoot gun and sword fighting is taken from HMDB51 dataset,[16] which is very complex dataset with real-world scenarios.

4.2 *Experimental Setup*

We have focused on default parameters for IDTs extraction as in Ref. 9. The length of trajectories is limited to $L = 15$ frames as in Ref. 9, to encode enough motion information because trajectories tend to drift from their initial location during tracking process. Spatial scale size is limited to $S = 8$ spatial scales to encode enough structural information as in Ref. 9. For dense sampling, $W = 5$ pixels spacing is selected because sampling space with $W = 5$ pixels offers good trade-off between speed and accuracy. Although increasing the sampling density increases the performance, it also increases the computational cost.[9] Descriptor is computed on default window size for IDT framework that is $N \times N = 32 \times 32$. To encode enough structural information and to represent descriptors with more detail, the window is divided into spatiotemporal cells. Default size is selected for spatiotemporal cells, i.e., $n_\sigma \times \mathrm{n}_\sigma \times n_\tau = 2 \times 2 \times 3$. Then, the cell descriptors are computed. Descriptors for each cell of spatiotemporal grid are computed. Final descriptor is obtained by the concatenation of these all cell descriptors. Descriptors are represented as orientation and magnitude histograms. Eight-bin histogram is obtained for descriptors like HOG, MBHx, MBHy, ST-HMA, where magnitude is used for weighting and orientation information if quantized to histogram. The size of descriptor is $n\sigma \times n\sigma \times n\tau \times 8$, i.e., $2 \times 2 \times 3 \times 8 = 96$. For HOF, descriptor size is $n\sigma \times n\sigma \times n\tau \times 9$, i.e., $2 \times 2 \times 3 \times 9 = 108$. To reduce the computational cost, PCA is used for dimensionality reduction. The dimensions are reduced to half from 96 to 48 and 108 to 54. To handle the variation of action classes taken from different state-of-the-art action recognition datasets, the proposed method is evaluated with two cross-validation schemes. BOVW framework is tested with different codebook sizes like 1000, 2000 to 4000. Linear SVM (SPM) kernel is used for classification. One-against-all approach is adapted with multiclass SVM classification.

4.3 *Evaluation Parameters and Results*

The proposed methodology is evaluated by average accuracy for all action classes. Also, per-class accuracy to evaluate individual action class is reported. Densely sampled feature points and IDT extracted for a sample frame are shown in Fig. 6. One of the most important steps in BOVW framework is dividing feature space into a group of similar clusters. K-mean clustering is the most popular way because of its simplicity and efficiency. The behavior of descriptors is evaluated with different codebook sizes. Choice of codebook size is very crucial. Increasing the codebook size may result in over partitioning of feature space. In our experiments, three different codebook sizes are used, i.e., 1000, 2000, and 4000. Results are shown in Fig. 7 on different codebook sizes.

Difference in recognition accuracy for different codebook size is not too much for all descriptors. From 1000 to 2000 codebook size, recognition rates are increasing for all descriptors. From 2000 to 4000, on some descriptors, the results are improved. Best results are at $k = 2000$ for most of descriptors like HOF, HOG, STHMA, trajectory, MBHx. Increasing codebook size may result in over

Fig. 5 Example video frames from each action class in HCA dataset.

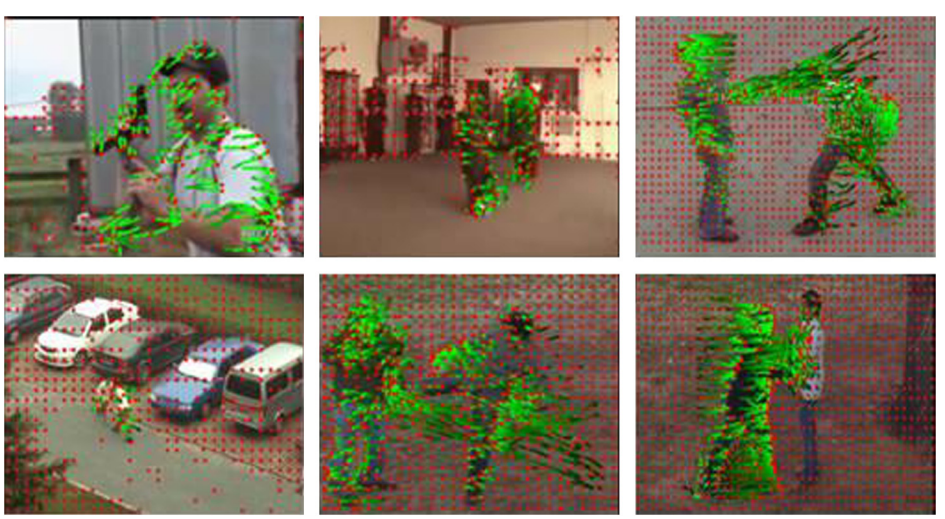

Fig. 6 Visualization of densely sampled points in red and IDTs for some example videos.

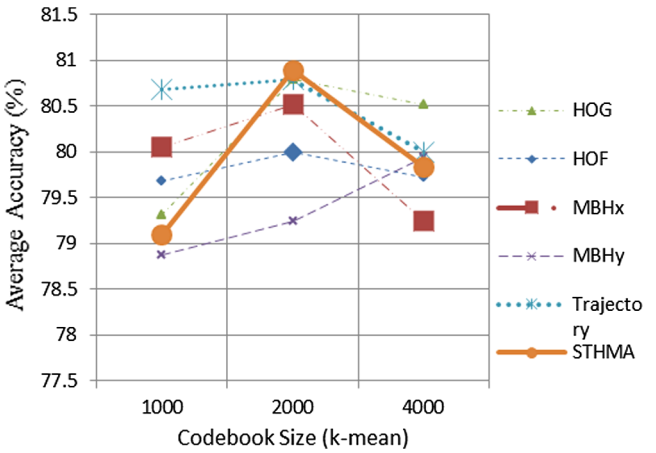

Fig. 7 Analysis of codebook size for k-mean clustering.

partitioning of feature space for some videos in case of HCA dataset. So, the rest of the evaluation is carried out with $k = 2000$.

Two types of cross validation schemes: randomly selecting training testing (tr-ts) splits and k-fold cross validation are tested over HCA dataset. Experimental results and discussions using two cross-validation schemes are discussed as follows.

4.3.1 *Randomly selecting tr-ts splits*

Videos are divided into training and testing splits randomly. To handle large variation among action classes, 30% of videos are used from each class for classifier training and 70% of videos from each class are used for testing. To verify the robustness of the proposed system, experiments are conducted three times and the average is taken. Average accuracy is reported in Table 2.

Six descriptors are evaluated on IDT framework for HCA datasets. Recognition rates for all descriptors are below 81% because HCA dataset contains large intraclass variation, camera motion, and videos taken from different environments. ST-HMA descriptor is compared with other state-of-the-art descriptors. For this type of cross-validation scheme, the highest recognition rate is 80.89% for ST-HMA descriptor because it encodes HMA and provides good recognition. HOG feature characterizes shape information and it is giving 80.79% recognition rate. IDT is an improved version of DT with camera motion correction and improvement of trajectories extracted. MBHx and MBHy descriptors are characterizing motion boundary information with respect to horizontal and vertical directions. MBH descriptor is robust to camera motion. MBHx attains 80.52% average accuracy rate better than MBHy because motion in vertical direction is more dominant for HCA dataset. HOF is giving average accuracy 80% that is less than other motion descriptors because HMA descriptor is robust to camera motion present in realistic set of actions. While HOF is not able to handle camera motion in a better way, trajectory descriptor is giving recognition results of 80.79 comparable to HOG and better than HOF descriptor. IDT improves motion descriptors by correcting camera motion, which is why recognition rate for STHMA descriptor is better as compared to other descriptors. HOG and trajectory descriptors are giving comparable results 80.79%. Mbhx is giving 80.52% average accuracy rate higher than Mbhy 79.25%. HOF descriptor is giving 80.00% recognition rate, which is less than other motion descriptors Mbhx and ST-HMA.

Table 2 Average accuracy for six descriptors, based on different cross-validation schemes.

	Cross validation schemes	
Descriptors over IDT framework	Randomly selecting training testing splits	k-fold cross validation
HOG[11]	80.79	79.80
HOF[11]	80.00	79.02
Mbhx[8]	80.52	77.39
Mbhy[8]	79.25	80.79
Trajectory[7]	80.79	80.89
Our STHMA	80.89	79.80

4.3.2 *K-fold cross validation scheme*

In this scheme, $k = 3$ folds cross-validation scheme is adapted. In threefold cross-validation, sample videos are divided into randomly partitioned three-equal sized subsamples. From the three subsamples, a single subsample is used for testing and remaining two are used for training. The process is repeated three times and average accuracy is reported in Table 2. It is found that trajectory descriptor is giving highest recognition rate of 80.89%. The proposed STHMA is giving 79.80% recognition rate lower than obtained from previous scheme. For this type of scheme, folds are randomly selected, and in previous scheme, splits are randomly selected from each class.

The proposed STHMA is evaluated for respective action class using per-class accuracy, as shown in Table 3. Based upon the observations from Table 2, this evaluation is carried out with randomly selection 30% training and 70% testing splits. For shoot-gun and sword fighting, per class accuracy is low 51.9% and 45.23%, respectively. These actions are taken from HMDB51 dataset having a large amount of camera motion, background motion, occlusion, and intraclass variance. HMDB51 videos are taken from action movies and YouTube having different resolutions, low quality, etc. For fight action taken form CASIA action dataset, per class accuracy is 96.19% higher than other classes because CASIA dataset contains videos taken from static camera with little background motion. Kick, push, and punch are taken from UT-interaction datasets with background movement in some videos.

Per class accuracies for these actions are compared with state-of-the-art techniques and it is found that STHMA on IDT provides better results than edge based trajectories, mid-level representation, space–time multiscale motion descriptor and spatiotemporal relationship match. In Ref. 13, edge trajectories are more likely to corrupted with camera motion than IDT because correct localization of edge points

Table 3 Comparison of STHMA with state-of-the-art methods.

Methods	Criminal actions per class accuracy (%)					
	Kick	Push	Punch	Shoot-gun	Sword fighting	Fight
Spatiotemporal relationship match[22]	75	75	50	—	—	—
Mid-level representation[21]	85	80	55	—	—	—
HMA on edge trajectories[13]	85	100	80	—	—	—
Space–time multiscale motion descriptor[23]	90 (set 1)	70 (set 1)	80 (set 1)	—	—	—
	80 (set 2)	60 (set 2)	80 (set 2)			
IDT[9]	—	—	—	57.2 Avg. Acc (HMDB51 dataset)	57.2 Avg. Acc (HMDB51 dataset)	—
Spatiotemporal pairwise trajectory representation[10]	—	—	—	60.57 Avg. Acc (HMDB51)	60.57 Avg. Acc (HMDB51)	—
STHMA on IDT	92.85	92.85	93.33	51.90	45.23	96.19

is challenging one in case of camera motion. For HMA descriptor on edge trajectories, average accuracy for push is 100%, however, for lower values for kick (85%) and punch (80%). The proposed STHMA descriptor is 92.85% for kick and 93.33% average accuracy is obtained for punch action classes. In Ref. 29, visual motion patterns are segmented, and middle level components are generated by grouping set of similar key points based trajectories. In this paper, per-class accuracy for UT-interaction dataset is reported, and it is found that for kick action recognition rate is 85% same as in Ref. 13, however, lower than STHMA descriptor with IDT framework. For punch and push, recognition rates are 55% and 80% higher than Ref. 22, however, lower than other two techniques, respectively.

In Ref. 22, structural similarity between features extracted from videos is measured and matched. This method attains lower recognition rate on UT-interaction dataset than other methods, as shown in Table 3. In Ref. 23, a descriptor is the proposed space–time multiscale motion descriptor from spatially constrained decomposition. Results are reported individually for set 1 and set 2 of UT-interaction dataset. It can be seen that for push action, results are 70% and 60%, respectively. These per-class accuracies are lower than other methods, as reported in Table 3.

For punch and kick actions, per-class accuracies are comparable with other methods. Average accuracy for HMDB51 dataset having actions like shoot-gun and sword fighting with 51 action classes is also shown in Table 3. Average accuracy is 57.2% in case of IDT-based methods and 60.57% in case of spatiotemporal pairwise trajectory representation, higher than our per class accuracy results for individual action. HMDB51 dataset contains simple daily life actions too, i.e., smile, talk, laugh, etc. These simple actions are easy to detect, and recognition rates are much higher than complex actions like sword fighting or gun-shooting, which is why recognition rate average is higher than percales accuracy. From Table 3, it is clearly visible that per-class accuracies for STHMA descriptor are better than other methods.

Computational costs for k-fold cross validation and randomly selecting training testing splits are shown in Table 4. This is the total time in minutes required for feature extraction, feature encoding, and classification. It is found that k-fold cross-validation with $k = 3$ has more computational cost than randomly selecting training testing splits with three rounds. Time for computing descriptor for six action classes is more than the rest of evaluation. Time required for descriptor computation is 206.28 min out of total 210.11 for k-fold and 209.86 for randomly selecting tr-ts split. Time required for descriptor calculation per action video is 0.66 min.

Table 4 Computational cost with ST-HMA descriptor over HCA dataset.

Cross-validation techniques	k-fold cross validation	Randomly selecting tr-ts splits
Computational cost in minutes	210.11 min	209.86 min
Computational cost per action video for ST_HMA	0.66 min	0.66 min

Table 5 Results on HCA and other action datasets.

Datasets	Cross validation schemes	
	Randomly selecting training testing splits	*k*-fold cross validation
Our HCA	80.79	79.80
UT-interaction	74.60	73.24
CASIA	72.59	75.33

4.3.3 *Evaluation and comparison with other action datasets*

The proposed descriptor ST-HMA is evaluated for two action datasets UT-Interaction[42] and CASIA.[41] UT-interaction dataset consists of six human–human interactions of hug, kick, pointing, punch, push, and shake-hand. CASIA dataset consists of seven type of interactions, including fight, follow-always, follow-together, meet-apart, meet-together, overtake, and rob. Results are shown in Table 5. For both types of cross validation schemes, the best results of 80.79% accuracy are for HCA dataset. On UT-interaction, 74.60% accuracy is achieved for randomly selecting training testing splits. On CASIA dataset, *k*-fold cross-validation scheme is giving good result of 75.33% accuracy as compared to randomly selecting tr-ts splits. We can observe that ST-HMA shows better recognition accuracy on HCA, which is primarily containing actions of criminal nature; however, on general actions, its accuracy is acceptable but not much better than criminal actions.

5 Conclusion

In this paper, ST-HMA descriptor is proposed, and evaluation is carried out in IDT framework. An HCA dataset having many challenges like intraclass variation, illumination changes, view point changes, background motion, and camera motion is developed. The behavior of different descriptors on HCA dataset is evaluated and found that ST-HMA on IDT framework provides promising results. Behaviors of the dataset for two cross-validation schemes are also reported. Moreover, individual action class is evaluated using per-class accuracy and compared with state-of-the-art results. Computational cost for ST-HMA descriptor with both cross-validation schemes is also evaluated.

In the future, HCA dataset can be used to further evaluate state-of-the-art or action recognition techniques for complex criminal action dataset. Similarly, in the future, more discriminative descriptors can be tested with IDT framework. These descriptor combinations can be used to eliminate or replace some other ones.

References

1. E. B. Nievas et al., "Violence detection in video using computer vision techniques," *Lect. Notes Comput. Sci.* **6855**, 332–339 (2011).
2. S. Shah et al., "Automated vigilance assistance system with crime detection for upcoming smart cities," SAE Technical Paper 2017-01–1726, Warrendale, Pennsylvania (2017).
3. T. Lawson, R. Rogerson, and M. Barnacle, "A comparison between the cost effectiveness of CCTV and improved street lighting as a means of crime reduction," *Comput. Environ. Urban Syst.* **68**, 17–25 (2018).
4. M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "A review of human activity recognition methods," *Front. Rob. AI* **2**, 28 (2015).
5. G. Cheng et al., "Advances in human action recognition: a survey," arXiv:1501.05964 (2015).
6. H. Kataoka et al., "Evaluation of vision-based human activity recognition in dense trajectory framework," *Lect. Notes Comput. Sci.* **9474**, 634–646 (2015).
7. H. Wang et al., "Action recognition by dense trajectories," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3169–3176 (2011).
8. H. Wang et al., "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vision* **103**(1), 60–79 (2013).
9. H. Wang and C. Schmid, "Action recognition with improved trajectories," in *IEEE Int. Conf. on Computer Vision*, pp. 3551–3558 (2013).
10. Q. Huang, S. Sun, and F. Wang, "A compact pairwise trajectory representation for action recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1767–1771 (2017).
11. I. Laptev et al., "Learning realistic human actions from movies," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1–8 (2008).
12. A. Edison and C. V. Jiji, "HSGA: a novel acceleration descriptor for human action recognition," in *Fifth National Conf. on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, pp. 1–4 (2015).
13. X. Wang and C. Qi, "Action recognition using edge trajectories and motion acceleration descriptor," *Mach. Vision Appl.* **27**(6), 861–875 (2016).
14. C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proc. of the 17th Int. Conf. on Pattern Recognition*, Vol. 3, pp. 32–36 (2004).
15. L. Gorelick et al., "Actions as space-time shapes," *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(12), 2247–2253 (2007).
16. H. Kuehne et al., "HMDB51: a large video database for human motion recognition," in *High Performance Computing in Science and Engineering '12*, W. E. Nagel, D. H. Kröner, and M. M. Resch, Eds., Springer, Berlin, Heidelberg, pp. 571–582 (2013).
17. S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition," *Image Vision Comput.* **60**, 4–21 (2017).
18. I. Laptev, "On space-time interest points," *Int. J. Comput. Vision* **64**(2–3), 107–123 (2005).
19. G. Willems, T. Tuytelaars, and L. V. Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," *Lect. Notes Comput. Sci.* **5303**, 650–663 (2008).
20. X. Peng et al., "Bag of visual words and fusion methods for action recognition: comprehensive study and good practice," *Comput. Vision Image Understanding* **150**, 109–125 (2016).
21. F. Yuan, V. Prinet, and J. Yuan, "Middle-level representation for human activities recognition: the role of spatio-temporal relationships," *Lect. Notes Comput. Sci.* **6553**, 168–180 (2010).
22. M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: video structure comparison for recognition of complex human activities," in *IEEE 12th Int. Conf. on Computer Vision*, pp. 1593–1600 (2009).
23. F. Martínez, A. Manzanera, and E. Romero, "Spatio-temporal multiscale motion descriptor from a spatially-constrained decomposition for online action recognition," *IET Comput. Vision* **11**(7), 541–549 (2017).
24. L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR)*, Vol. 2, pp. 524–531 (2005).
25. E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," *Lect. Notes Comput. Sci.* **3954**, 490–503 (2006).
26. A. Klaeser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *19th British Machine Vision Conf.*, pp. 99.1–99.10 (2008).
27. C. Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag, New York (2006).
28. H. Xu et al., "A survey on aggregating methods for action recognition with dense trajectories," *Multimedia Tools Appl.* **75**(10), 5701–5717 (2016).
29. W. Chen and J. J. Corso, "Action detection by implicit intentional motion clustering," in *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, Washington, DC, pp. 3298–3306 (2015).
30. D. Weinland, E. Boyer, and R. Ronfard, "Action recognition from arbitrary views using 3D exemplars," in *IEEE 11th Int. Conf. on Computer Vision*, pp. 1–7 (2007).
31. K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Mach. Vis. Appl.* **24**(5), 971–981 (2013).
32. P. Rota et al., "Real-life violent social interaction detection," in *IEEE Int. Conf. on Image Processing (ICIP)*, pp. 3456–3460 (2015).
33. C. Gao et al., "From constrained to unconstrained datasets: an evaluation of local action descriptors and fusion strategies for interaction recognition," *World Wide Web* **19**(2), 265–276 (2016).
34. R. Szeliski, "Image alignment and stitching: a tutorial," *Found. Trends Comput. Graphics Vision* **2**(1), 1–104 (2006).
35. H. Bay et al., "Speeded-up robust features (SURF)," *Comput. Vision Image Understanding* **110**(3), 346–359 (2008).

36. M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM* **24**(6), 381–395 (1981).
37. A. Prest, C. Schmid, and V. Ferrari, "Weakly supervised learning of interactions between humans and objects," *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(3), 601–614 (2012).
38. C. Bishop, *Pattern Recognition and Machine Learning*, Springer, http://www.springer.com/gp/book/9780387310732 (22 July 2017).
39. E. Participation, "Offences Against the Person Act 1861," http://www.legislation.gov.uk/ukpga/Vict/24-25/100/contents (18 September 2018).
40. "Main Features–Introduction," http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/4530.0main+features100022012-13 (25 November 2018).
41. CBSR, "CASIA action database for recognition," http://www.cbsr.ia.ac.cn/english/Action%20Databases%20EN.asp (2 September 2016).
42. M. S. Ryoo et al., "An overview of contest on semantic description of human activities (SDHA) 2010," *Lect. Notes Comput. Sci.* **6388**, 270–285 (2010).

Abinta Mehmood Mir is an MSc research scholar at Computer Engineering Department, University of Engineering and Technology, Taxila. Her research area covers computer vision and action recognition. Currently, she is working on criminal action recognition from surveillance videos. She has completed her BSc degree in computer engineering from UET, Taxila, with honors.

Muhammad Haroon Yousaf received his BSc, MSc, and PhD degrees in computer engineering from the University of Engineering and Technology Taxila, Pakistan. He is currently serving as an associate professor and director postgraduate studies in Computer Engineering Department, University of Engineering and Technology Taxila, Pakistan. His research interests lie in image processing/computer vision. He has published a number of papers in international conferences and journals. He is also working as editor and reviewer for different journals. He has been the recipient of Best University Teacher Award (2012–2013) given by Higher Education Commission (HEC) of Pakistan.

Hassan Dawood is working as an assistant professor at Department of Software Engineering, University of Engineering and Technology, Taxila, Pakistan. His research interests include image restoration, feature extraction, and image classification. He has received his MS and PhD degrees in computer application technology from Beijing Normal University, Beijing, China, in 2012 and 2015, respectively.