Vis Comput
DOI 10.1007/s00371-015-1066-2

SURVEY

A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector

Debapratim Das Dawn · Soharab Hossain Shaikh

Abstract Over the past two decades, human action recognition from video has been an important area of research in computer vision. Its applications include surveillance systems, human–computer interactions and various real-world applications where one of the actor is a human being. A number of review works have been done by several researchers in the context of human action recognition. However, it is found that there is a gap in literature when it comes to methodologies of STIP-based detector for human action recognition. This paper presents a comprehensive review on STIP-based methods for human action recognition. STIP-based detectors are robust in detecting interest points from video in spatio-temporal domain. This paper also summarizes related public datasets useful for comparing performances of various techniques.

Keywords Human action recognition · Human activity recognition · Spatio-temporal interest point · STIP

1 Introduction

Over the last two decades [1–3], human action recognition from video generated lots of momentum among the researchers of computer vision community. This field is closely related to other field of studies like motion analysis [4] and action recognition [5].

This paper focuses on the method based on STIP-based human action recognition. However, at the onset, it will be logical to briefly talk about different types of human activities researchers are dealing with. Analyzing and identification of different types of human activity from an unknown video sequences are the main objectives of human activity recognition. The types of human activity are classified under four different categories depending on complexity of actions and number of body parts involved in the action; gestures, actions, interactions, and group activities are the four different types of human activities [6].

Gestures It is a collection of movements, made with hands, head or face to show a particular meaning [6]. The 'Arm stretching', 'head shaking' , 'facial expression', and 'leg rising' are good examples of human gestures.

Actions It is a collection of multiple gestures performed by a single person [6]. The 'walking', 'waving', 'running', 'jogging', and 'punching' are examples of human action categories.

Interactions It is a collection of human actions of maximum two actors. One actor must be a human being and other one may be a person or an object [6]. Also, this section is classified as human–human interactions and human–object interactions. For human–human interactions, two actors are human beings. In human–object interactions, one actor must be a human being and other one is an object. 'Talking between two persons', 'fighting between two persons', 'hand shaking', and 'welcoming each other' are the examples of human–human interaction, and 'ATM theft' and 'doing work in front of a computer' are the examples of human–object interaction.

Group activities It is a combination of gestures, actions or interactions where the number of actors is more than two and there may be single or multiple interactive objects [6]. 'Two groups playing some games or involving some activity',

D. Das Dawn (✉)
University of Calcutta, Kolkata, India
e-mail: debapratimdd@gmail.com

S. H. Shaikh
Faculty Member, University of Calcutta, Kolkata, India
e-mail: soharab.h.shaikh@ieee.org

Published online: 10 March 2015

‘marches group of people’, ‘group meeting’ and ‘fighting between two groups’ are the examples of group activity.

The methods of human action recognition from image frames or video sequences are broadly classified as template-based approach (emphasis on collecting low- and mid-level features) and model-based approach (emphasis on feature for high-level interaction) [7]. A motion analysis system generates information regarding video data by processing consecutive image frames of a video. The “optic flow and feature tracking” are two traditional approaches of motion analysis. However, ‘Optic flow’ approaches are unable to capture variable motion information and ‘Feature Tracking’ approaches fail when appearance changes suddenly such that merging or splitting of two objects.

A STIP-based detector captures interest points from a video in spatio-temporal domain. An image I(x, y) in spatial domain is represented as image stack I(x, y, t) in spatio-temporal domain. An interest point can be robustly detected by STIP-based detector. For example, an interest point may be a corner point or an isolated point (where intensity is maximum or minimum) or end point of a line or point of a curve (where curvature is maximum). In the context of human action recognition, generally corner is used as a interest point. Some popular corner detection algorithms are Moravec [8], Harris [9], Forstner [10], etc. However, it is found that there is a gap in literature when it comes to methodologies of STIP-based detector. So, this survey work has concentrated on spatio-temporal-based human action recognition.

From literature review [6,11–13], it is well known that due to the sparsity and good performance, the STIP detector-based techniques are very effective in recognizing human actions from an unknown video. The STIP-based methods have more power to select interest points in the dynamic contents of a video [11].

Over the past one decade, a lot of STIP-based techniques [12–15] are useful for recognizing action of human being. Spatio-temporal interest points are detected from input video using STIP-based detector. Subsequently, features are extracted from the interest points and vocabulary of features are build from feature representer. Finally, action classification is performed using suitable classifier. Figure 1 shows a schematic overview of STIP-based action recognition system. In real-life implementation of human action recognition, the STIP-based approaches are more suitable for obtaining the desired result. The STIP-based approaches are prevailed to handle real-life scenario such as clutter background, illumination, variation of contrast and brightness more robustly.

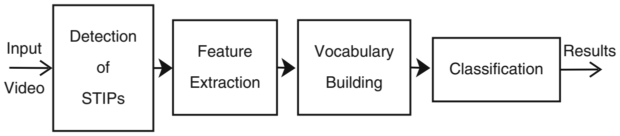

Fig. 1 Flowchart of STIP-based human action recognition

Different applications of human action recognition include surveillance [16–18], human–computer interactions (HCI) [17–20], content-based summarization and indexing [16,17, 20–22], and health care system [23]. In the context of surveillance, recognition of human action can play an important role. As for example, a typical action of a human being (i.e., sudden change from walking to running) in a restricted area may lead to a secure thread to the system. Content-based video summarization and indexing are another application areas in human action recognition. In this context, a summarized video can be formed by considering only the frame of typical action (e.g., players running with a football in a football match). An action recognition-based health care system can support patient rehabilitation, behavior monitoring and detection of abnormal activity, etc.

2 Related work

This section summarizes some of the potential review works on the field of human action recognition based on different methodologies apart from STIP. A lot of survey papers on motion analysis, human action recognition and activity recognition have been proposed by several researchers.

Generally, a motion-based recognition system concerns with motion information and extracts from image sequences. Cedras et al. [24] made an extensive survey on motion-based recognition of human movements. The motion information of a video was classified as trajectory-based feature (e.g., velocity, direction, spatio-temporal curvature), optical flow (e.g., average flow, flow statistics) and region-based feature (e.g., mesh feature). Aggarwal et al. [4], classified the methods of human activity recognition into template matching (i.e., points, meshes) and state space based (i.e., points, lines, blobs). It focused on human movement recognition and motion analysis of high-level tracking. Moeslund et al. [25] covered a considerable survey on the basis of human motions from the year 2000 to 2006 and focused on automatic tracking, reconstruction and recognition of human motion. The survey work put emphasis on function-based taxonomy such as initialization, tracking, pose estimation and recognition.

In computer vision, “Looking at people” means analysis of human movements with the help of a machine. Gavrila [26] made a comprehensive survey on the movement of hands and motion of whole body parts. The methods of human movement analysis were classified as 2D approaches (i.e., with and without explicit shape models) and 3D approaches.

The study of action recognition consists of three scientific disciplines: computer vision, robotics and artificial intelligence (AI). Kruger et al. [27] made up a meaningful survey on action recognition at different levels of complexities. It adapted an action hierarchy of each action recognition system such as action/motor primitives, actions and activities. Besides, it followed egocentric (i.e., simple) and ecocentric approach (i.e., complex, deals with environmental parameters) for recognizing the actions. Moreover, the methods of action recognition were classified as scene-based, full-body-based, body part-based and grammar-based approach. Poppe [28] also discussed about vision-based human action recognition and gave emphasis on action classification and representation module. The representation of an image was classified as global representations (i.e., obtained in a top-down fashion) and local representations (i.e., collection of independent patches). Moreover, Weinland et al. [29] made an extensive survey of more than one-hundred papers in action recognition domain and covered various related methods in almost one decade, from 2001 to 2011. It placed emphasis on spatial and temporal structure of action with variation of camera and viewpoints. The methods of feature point detectors and descriptors were classified into three different categories such as bags of trajectories, feature templates and bags of events. In consequence, Iosifidis et al. [22] made a comprehensive survey about the techniques of multi-view human action recognition. The methods of multi-view human action recognition were classified as 3D multi-view (i.e., multi-camera setup during both training and testing phases) and 2D multi-view methods (i.e., methods that can operate by using an arbitrary number of cameras).

Due to the complexity, the recognition of human activity is very challenging task among the researcher communities. It covers various scientific disciplines such as AI, computer vision, linguistics and neuroscience. Taking this into account, Turaga et al. [30] made a comprehensive survey on machine-based recognition of human activities in real-world environment. The methods of activity recognition were classified as graphical model (e.g., dynamic Bayes nets), syntactic model (e.g., context-free grammar) and knowledge-based model (e.g., constraint satisfaction). Further classification of action recognition methods included parametric (e.g., HMM-hidden Markov models), non-parametric (e.g., template matching) and volumetric (e.g., space–time filtering) approach. In addition with, the article got to grip the impacts of invariance factors. One of the most remarkable challenges in action recognition domain is invariance factor, i.e., unevenness of ascertained features in same action class. Due to the factors of viewpoint, execution rate and anthropometry, invariances of an action analysis system were classified as view-invariance (cause: variation of motion, occlusion, camera effects, etc.), execution rate invariance (cause: temporal variations, time warps, etc.) and anthropometric invariance (cause: shape, gender and size, etc.). Moreover, Aggarwal et al. [6] classified the various techniques of activity analysis using approach-based taxonomy. It considered non-hierarchical approaches (for the recognition of gestures and actions) and hierarchical approaches (analyze high-level interactions between multiple humans with objects) for analysis of the activities. Besides, Ke et al. [31] made up an extensive survey on threesome aspects of human activity recognition system, such as core technology (i.e., human object segmentation), human activity recognition system (i.e., single or multiple people) and applications from low-level to high-level representation. The article put into sets the feature extraction methods such as space–time volume, frequency, local descriptors and body modeling types.

In surveillance and health care system, sensor-based activity recognition technique is very useful for smarter performance. Due to the complicated function of sensing, inference and learning, the sensor-based activity recognition is very challenging task among the researchers. Guan et al. [32] made an extensive survey on video sensor-based (i.e., observes remotely) and physical sensor-based (sensor attach to body) activity recognition systems. The physical sensor-based activity recognition systems were systematized into wearable sensor (sensor stick with human body, e.g., gyroscopes, accelerometers) and object usage-based (sensor stick with object, e.g., radio-frequency identification, binary sensors). Vishwakarma et al. [21] provided a survey work on surveillance-based activity recognition system. The techniques of motion analysis system were grouped into three parts; low- (i.e., human detection), intermediate- (i.e., human tracking) and high-level vision (i.e., behavior understanding). The taxonomy of object detection was grouped into background subtraction, statistical methods, temporal differencing and optical flow. Moreover, the approaches of activity recognition system were classified as non-hierarchical approach (e.g., space–time volume, space–time features, trajectories, state-based) and hierarchical approach (e.g., description-based, syntactic). Despite that, Xu et al. [33] classified human surveillance-based activity recognition systems into template matching and state space-based approach.

In essence, datasets play a crucial role for performance analysis in the action recognition domain. Chaquet et al. [34] discussed about 68 public datasets for video-based activity recognition system. According to the type of actions, it considered a possible taxonomy of actions such as heterogeneous action (i.e., jumping, running, walking, waving, etc.), specific action (i.e., abandoned objects, crowd behavior, detecting falls, gait, post and gesture) and others (i.e., motion capture-MOCAP, infrared and thermal). On top of that, Hassner [35] provided a comprehensive survey on contemporary action recognition systems in Weizmann [55] and Action Similarity LAbeliNg (ASLAN) [61] benchmark datasets.

Recently, some researchers [11,23] have cited the importance of STIP-based methods along with others. In consequence, Akila et al. [23] classified all the methods of human action recognition into spatio-temporal-based, shape-

Table 1 A brief description of related review or survey papers

References	Year	Topic
Cedras et al. [24]	1994	Survey on motion-based recognition and put emphasize on motion trajectory
Aggarwal et al. [4]	1998	Review on human motion analysis and focus on tracking features over image sequences
Gavrila [26]	1998	Survey on visual analysis of human movements with 2D and 3D approaches
Moeslund et al. [25]	2006	Vision-based motion capture and analysis with emphasized on tracking and recognition
Kruger et al. [27]	2007	Survey on action representation, recognition and mapping at different complexity levels
Turaga et al. [30]	2008	Machine-based recognition of human activities with consideration of invariance factors
Poppe [28]	2010	Vision-based human action recognition by labeling image sequences with action labels
Weinland et al. [29]	2011	Survey on vision-based methods for action representation, segmentation and recognition
Aggarwal et al. [6]	2011	A review on human activity analysis using approach-based taxonomy
Guan et al. [32]	2011	Review on video and physical sensor-based activity recognition systems
Vishwakarma et al. [21]	2012	Survey on activity recognition and behavior understanding in surveillance system
Li et al. [11]	2012	Makes comparison on techniques of STIP- based recognition
Hassner [35]	2013	Critically review on action recognition benchmarks : Weizmann [55] and ASLAN [61] set
Ke et al. [31]	2013	Survey on video-based human activity recognition system in different aspects
Chaquet et al. [34]	2013	A taxonomy-based survey of video datasets for human action and activity recognition
Xu et al. [33]	2013	Survey on template matching and state space-based approaches in activity recognition
Iosifidis et al. [22]	2013	Review on 2D and 3D multi-view-based human action recognition techniques
Akila et al. [23]	2014	Comparative analysis of action recognition methods with various phases and techniques

or pose-based, interest point-based and motion- or optical flow-based visual recognition. Along with, they commented on different phases of human action recognition methods such as foreground extraction, tracking, feature extraction and recognition. Li et al. [11] made a survey on various techniques of STIP-based action recognition. It focused on STIP detection, feature classification, action representation and recognition module. The extracted features were categorized into two sets: static feature (based on shape and edge) and dynamic feature (based on optical flow and motion trajectory). The methods of action representation and recognition were grouped into bag-of-words model and state space-based model.

All of the reported authors commented critically on human action, activity and motion analysis techniques and mentioned a lot of future scopes. Table 1 presents a brief description about reported papers. However, most of the previous reviewers have focused on activity and motion analysis system and followed tracking and surveillance-based approach. Hence, this paper presents a comprehensive survey on human action recognition and puts emphasis on spatio-temporal or STIP-based approach.

3 Survey of literature on STIP-based methods

In the general run of things, the STIP-based human action recognition systems have foursome components such as STIP detector, feature descriptor, vocabulary builder and classifier. All of the components are put into two sets. On the basis of feature sets, the STIP detector is grouped into dense and sparse feature-based detector. A dense feature detector densely covers the video content. On the other hand, a sparse (local) feature is a subset of the total feature vector (i.e., large and unbound). The spatio-temporal-based descriptors are classified as local descriptor and global descriptor. A local descriptor captures only local or static information (i.e., color, posture, texture etc.), while a global descriptor captures the global or dynamic information (i.e., scale changes, illumination changes, speed variation, phase variation, etc.) of a video. The shape or edge relevant data of objects are used for determining the local features of a video. However, the global information points towards the description of flow or motion of a video. The vocabulary builder is systematized into bag-of-words (BOW)-based model and state space-based model. “Seeing an article as a collection of many words” is the heart of the matter of BOW-based model. The state space model interprets as each invariable posture as a state and makes mutual relations among the states using probabilities. On the basis of control, the STIP-based classifier is grouped into supervised (i.e., human-guided) and unsupervised (i.e., computes by the software) model-based classifier. Figure 2 summarizes the major functional steps involved in human action recognition. It also categorizes the different techniques used by the researcher in each of these functional steps.

3.1 Action recognition in controlled environment

This section summarizes the major works done by several researchers on STIP-based methods tested on benchmark

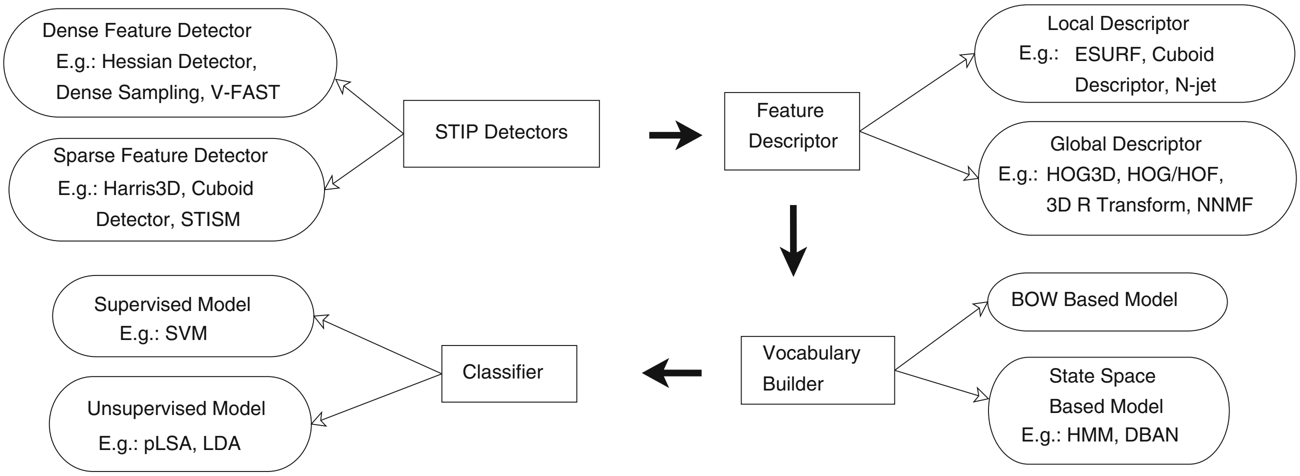

Fig. 2 Components of STIP-based human action recognition technique

datasets. Most of the datasets are bounded with some controlled environment. Here controlled environment does not necessarily mean the ideal laboratory condition. What is meant is that, data are acquired in a condition with no or very less influence of illumination changes, cluttered background, viewpoint changes, occlusions, etc.

The spatio-temporal model was proposed by Baumberg et al. [1]. The concept of physically based (i.e., vibration mode) spatio-temporal model was evolved from a set of training samples for describing the motion patterns of moving deformable objects. This model provides a low dimension shape descriptor and generates training information in a better way. In addition with, the point distribution model (PDM) was used for analysis of image sequences in robust tracking environment. The PDM is also applicable in gait analysis system. However, there is no temporal aspect of PDM and problem may arise on video-based motion analysis system.

In consequence, Laptev et al. [36] drew attention to detect interest points in spatio-temporal domain by means of space–time-based event detection with scale selection mechanism. The event-based interest point detection technique is very useful for analysis of repetitive motion pattern. A scale-invariant interest point detector is obtained from Harris interest point detector by incorporating the nominalized Laplace operator. This model detects interest points by following multi-scale approach and handles the limitation of two traditional approaches i.e., “optic flow and feature tracking”. Nevertheless, this approach provides an abstract representation of video data and the system is not robust against occlusions and cluttered backgrounds. Indeed, Laptev et al. [2] brought into focus the compact representation of video data by incorporating the idea of Harris and Forstner operator for detecting the interest points in space–time domain along with Lindeberg’s proposal for local scale selection in the spatial domain. In classification module, k-means clustering of Gaussian derivatives for each interest point in normalized form was utilized for developing local neighborhood-based approach to classify the events on the basis of similar nature and noise. However, the system is not invariant against Galilean transformation and motion-based recognition. In Laptev et al. [37] considered Galilean motion with transformation of input images for smoothing purpose. The histogram-based statistical framework was used for smoothing the image sequences. In addition to this, the velocity adaptation technique was evolved for interpreting the events against relative motion of cameras. The stabilization of non-adaptive filtered data was used for considering the velocity factor. The velocity adaptation Laplacian operator in each spatio-temporal scale is very useful for adapting the local velocity. However, the process may fail in case of non-static backgrounds, complex scene or multiple events of interests. Hence, improvement is needed in the section of scale selection module for appropriate scale selection in statistical framework. Figure 3 represents a visual comparison on spatial interest point (SIP) with STIP.

In the context of feature description of interest points, Laptev et al. [38] introduced a local space–time-based

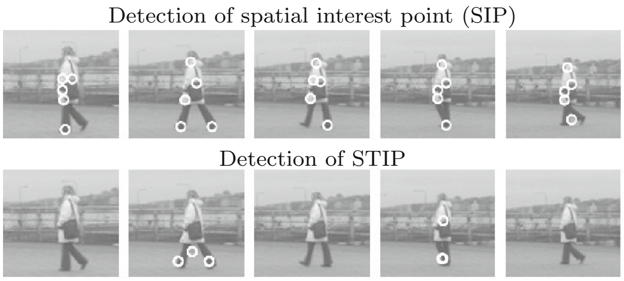

Fig. 3 Visually comparison on spatial interest point (SIP) with STIP. Courtesy: [36]

descriptor and event detector for representing the video data by utilizing the concept of various image descriptors such as histogram-based descriptor, principal component analysis (PCA), and N-jet. Moreover, a matching algorithm concept opened out for matching between STIPs with the image descriptor by employing the local greedy method (i.e., k-nearest neighbor classification technique). However, the local descriptor-based approach is fully dependent on relative motion between camera and objects and it varies in large amount for certain change of motion pattern. Hence, the improvement was needed in case of matching algorithm. In addition, for making a better matching algorithm, consider the Mahalanobis distances for matching with Support Vector Machine (SVM) classifier. To overcome the limitations of [38], Laptev et al. [39] proposed a velocity adaptation descriptor (i.e., adapt velocity in automatically) by incorporating the idea of μ-descriptor for neutralizing the effects of Galilean transformation. In essence, the Galilean transformation is applicable in temporal domain where relative motion tends to null. The velocity adaptation technique is very useful for increasing the stability of the descriptors, where motion pattern is unknown or background is movable. Laptev et al. [19] focused on event-based local motion representation (i.e., combination of event-based representation [36] with velocity adaptation technique [39]) for recognizing action in complex scene with dynamic backgrounds. Besides, the system made use of local motion events [2] for detecting neighbors of each event in space–time domain. The use of nearest neighbor (NN) and SVM classifier for event-based action recognition has become very popular among the researchers. However, the event-based local motion representation of a video rigidly depends on motion and appearance of an object.

Laptev et al. performed an admirable research on spatio-temporal domain for action recognition. Still, 3D counterpart of 2D methods is inadequate for action recognition in spatio-temporal domain. Dollar et al. [15] introduced sparse features for behavior recognition (for human as well as animal) in spatio-temporal domain. The concept of behavior recognition is very similar to object recognition. Moreover, it shows adeptness for thinking about different parameters such as posture, size, appearance, illumination and image clutter. On the whole, the sparse features are robust in anticipation of noise and pose variation. The method made use of spatio-temporal features by incorporating the idea of SIFT (i.e., Laplacian of Gaussian) detector in space–time domain together with 3D-Harris corner method (i.e., gradient vectors change in all directions: x, y and t) for detecting the features in a short video. By utilizing the concept of PCA-SIFT descriptor for same type of cuboid of each interest point, it considered behavior descriptor (i.e., histogram of cuboid of same type) using spatio-temporal order of cuboids. Apart from activity recognition, this is also applicable in recognition of facial expression, mouse behavior and distance measurement. The 3D-Harris corner detector identified some abrupt points where motion pattern is changing suddenly such as starting and stopping time.

The time and space complexity of any action recognition algorithm can be minimized, if the total number of interest point is kept low. On the contrary, the reliability of an algorithm may increase, if the total number of interest points is kept up. In essence, the more stable and unique STIPs (i.e., selective STIPs) are used for getting distinct interest points. In due course, Chakraborty et al. [13,40] proposed a novel action recognition algorithm using selective STIPs. The procedure made use of 2D-Harris corner detector with multiple spatial scales in each frame, along with found set of spatio interest points at different scales. The process is made up by removing the unwanted interest points in the background texture by calculating gradient weighting factor. Instead of foreground extraction, the process was involved for suppressing the background by means of non-maxima suppression technique. Thereafter, the system was obtained to selective STIPs with the help of temporal constraint (i.e., removes static interest points) together with matching algorithm (i.e., removes common points). The method of working made use of N-jet descriptor for feature extraction and BOW model for building vocabulary. Finally, SVM was used for action classification and recognition. In addition, Chakraborty et al. were the first to report exhaustive cross-data evaluation. However, the system comes behind greedy approach making the complexity of the system high sometimes.

The concept of scale-invariant interest point implied that the detected interest points are robust against scale-changing operation. In essence, the scale-invariant interest points are deducible from Gaussian derivatives of each interest point at different scales by selecting the local extrema over multiple scales. Willems et al. [12] extended the idea of scale-invariant interest point detector into spatio-temporal domain using Hessian-based STIP (Hes-STIP) detector with γ-normalization. The operating procedure was made use of 2D scale-invariant Harris–Laplace corner detector [2] into spatio-temporal domain together with SURF descriptor for describing dense feature points. Moreover, the system can handle a motion of moving camera. However, it brings out a very large number of feature sets for using dense sampling for STIP-based feature detection.

The global or dynamic information of a video consists of information regarding scaling, rotation, illumination changing, speed variation, phase variation, etc. (i.e., all flow or motion-based information). On this account, Wong et al. [41] contributed an introduction to novel feature extraction method for identifying the moving parts of an object by utilizing the global information. The dynamic texture and non-negative matrix factorization (NNMF) were used for extracting and representing the global information of video data. In addition, Difference of Gaussian (DoG) detector was used for faster execution. Apart from action recognition domain, this technique [41] is also applicable in the field of activity recognition, gesture recognition, facial expression identification, interest point counting, etc. However, it ignores a valuable global information regarding spatio-temporal distribution of STIPs. Moreover, some BOW models [16,17] also ignore the local informations of STIPs.

Yuan et al. [3] proposed a 3D $\Re$ transform (extension of 3D discrete Random transform) for capturing global distributions (in geometrically) of STIPs in an efficient manner (i.e., easy to compute). The method of working made use of $\Re$ feature (i.e., applicable for global information) and BOW model (i.e., applicable for local cuboid features) for representing the detected STIPs. In essence, BOW model operates on HOG/HOF features and $\Re$ descriptor carries off position-dependent feature (in both space and time). Taking advantage of context-aware kernel of selected contexts, the operating procedure opened out a context-aware fusion method for making a better pairwise relationship of these two feature vectors. At the bottom, the context-free kernels are very sensitive to outliers of video data and noise. The context of each video is derived from k-nearest neighbor classification approach. By analyzing the experimental result of that system, $\Re$+BOW combination furnishes with optimum result among other combinations like BOW+$\Re$, BOW+BOW, $\Re$+$\Re$, $\Re$, BOW, etc.

Apart from the STIP detection and feature extraction module, the vocabulary building section is also significant for recognizing the action classes. In essence, BOW model is used for generating the codebook of features. This is a histogram representation of particular visual pattern (according to number of occurrences). The significant benefits of this representation are relatively sparse, robust against occlusion and viewpoint changing operation, detect locally and compute efficiently. One of the key features of this technology is to make better classification strategy of multiple actions in a single video clip. In consequence, Niebles et al. [16] proposed an unsupervised learning method using spatio-temporal words (or features). The algorithm of unsupervised learning method made use of probabilistic Latent Semantic Analysis (pLSA) model and Latent Dirichlet Allocation (LDA) for learning the parameters of spatio-temporal words in an unsupervised way. The system is robust against noisy feature point, dynamic backgrounds and moving camera. On the contrary, the BOW model [16] is too local and may fail for capturing the adequate relationship of objects in spatial or temporal domain. Also, this may fail for capturing the information regarding relative layout of objects and actions of motion trajectories.

Kovashka et al. [17] proposed a hierarchy-based learning proposal for making a richer vocabulary of BOW-based action recognition system by means of features of discriminative spatio-temporal neighborhoods. A discriminative spatio-temporal neighborhood (size of the neighborhood may be fluctuated) feature provides a class-specific vocabulary and activity-specific information. The method of working made use of level-0 vocabulary using HOG3D descriptor (i.e., applicable for dense interest points) and HOG/ HOF descriptor (i.e., applicable for sparse points). Taking advantage of normalized Euclidean distance of N closest interest points, it formulated spatio-temporal neighborhood for generating features using compound descriptor (e.g., HOG, HOF, HOG3D) on next level of hierarchy (say level-1). Moreover, the multiple kernel learning (MKL) model was introduced for deducing the discriminative neighborhood of each interest point using distance matrices of multiple interest points. However, the BOW model [17] ignores spatio-temporal relationships among local descriptors. This model is not so enough to discriminate the multiple actions which are very closed to each other.

On that account, Yan et al. [42] proposed histogram of interest point location (HIPL) algorithm as supplement of bag-of-interest point (BIP) descriptor for capturing information regarding spatial distribution of STIPs. HIPL is a weaker descriptor than BIP, but also has more power to handle large amount of feature vectors. Moreover, the method of working made use of Adaptive Boosting (AdaBoost) and sparse representation (SR) with combination of weighted-output classifier (WOC) for better classification of feature sets. The AdaBoost is a learning algorithm, depicts probability of various classes and makes ensemble of some weak learners (e.g., stumps). The SR can handle large intra-class variation of action against corruption and noise distribution. The WOC model optimizes weight among combinations of multiple classifiers using their output strength, pros and cons etc. However, this HIPL model does not provide any temporal information of video data such as velocity. Subsequently, the system [42] is unable to properly discriminate the action classes which are very close to each other; as for example running and jogging.

3.2 Action recognition in real-world scenario

In most of the cases, it is seen that, the study of human action recognition is bounded with some controlled environment, as for example clean background, simple scene, no occlusion, etc. Even so, feature detection is very difficult task in case of cluttered videos. In this subsection, it summarizes the methodologies dealing with cluttered background, real-time processing , activity prediction, etc. As a consequence, Cao et al. [44] proposed an idea for dealing with occlusion and cluttered background by considering the multiple STIP features. In fact, the problems may arise for identifying the actors or distinguishing the actions or tracking of motion field in complex scene (i.e., presence of occlusion with cluttered background), where some body parts are occluded by other objects. The associated feature set for action recognition was classified as motion-based feature (e.g., motion history, optical flow) and appearance-based feature (e.g., edge, color). On top of that, it made use of heterogeneous features such as Hierarchical Filtered Motion Field (HFMF, dealing with crowded scene), sparse feature of histograms of oriented gradient with optic flow (HOG/HOF, histogram descriptors for the space–time volume) and adaptive action detection concept for combining the multiple features by means of Gaussian mixture model (GMM). Having said that, due to the computational complexity and low response time, this approach is not suitable for recognizing the action in real-time processing.

Subsequently, Matikainen et al. [45] considered pair-wise spatio-temporal relationship among features in the BOW-based framework using discriminative formulation and Naive-Bayes formulation. Moreover, it made use of appearance-based (i.e., STIP-HOG) and motion-based feature (i.e., quantized trajectories) for generating base feature codeword (i.e., corresponding to its spatio- temporal position).

Zhang et al. [18] gave thought to 4D-local spatio-temporal features (i.e., combination of both intensity and depth information) for handling camera motion in complex background. It used hyper 4D cuboid for each feature point (i.e., x, y, z, t: 3D spatial dimension and 1D in temporal dimension) and Latent Dirichlet Allocation (LDA) for categorization of human activities. However, the system is not robust against scale-changing operation; improvement is needed in the vocabulary building section.

The real-time processing of human action recognition is very popular in HCI, surveillance, video indexing, etc. Even so lacking of efficient algorithm, quick repose time and structural information (i.e., information regarding relationship among local descriptors), real-time processing is very difficult task among the researchers. Yu et al. [20] provided a real-time solution by considering the spatio-temporal semantic and structural forest for recognizing the actions. It introduced pyramidal spatio-temporal relationship match (PSRM) technique for capturing structural information, those are connected with descriptors. Also, the operating procedure made use of Video FAST (i.e., V-FAST) interest point detector (i.e., 3D counterpart of FAST corner detector) for collecting accurate dense interest point in short time sequence. The V-FAST interest point detector provides dense interest point, which has more power to classify the spatio-temporal semantic texton forests (i.e., STF). The spatio-temporal STF generates hierarchical information and codewords by imposing on spatio-temporal patches. Furthermore, it used kernel k-means forest classifier (i.e., PSRM + k-mean forest algorithm) for efficient classification of vocabularies. Despite that, it used semantic textons (BOST) for analyzing the texture perception and interaction among local space–time elements.

With the aid of activity prediction, human action recognition technique is proven to be very useful in video-based surveillance system, for example stopping of some criminal activity at its initial stage for avoiding unfortunate outcomes. Due to the incomplete observation, identification of action from short video clips is very challenging task among the researchers. Subsequently, Yu et al. [46] developed Spatial–Temporal Implicit Shape Model (i.e., STISM: 3D counterpart of Implicit Shape Model) for capturing the space–time structure of local sparse features. Due to additive nature of STISM, it can predict multiple actions simultaneously with incomplete observation from segmented video clips. The course of action made use of Multi-class Balanced Random Forest (MBRF) for efficient (i.e., save memory and computational cost) and discriminative random matching from training set to testing set. Instead of matching all the interest points from training to testing set, the MBRF model brings into focus the interest point pairs; those are falling in same leaf. Despite that, the system is not suitable for unsegmented video clips. Table 2 presents an overview idea of the reported papers.

3.3 Discussions

This section presents a comprehensive summary of the reported techniques on STIP-based detector. Table 3 represents a component-wise analysis and short remark of the reported STIP papers. Most of the researchers worked on sparse feature-based STIP detector with local descriptor-based approach. Also, dense feature-based STIP detectors generate good result for handling large number of feature sets. In the context of vocabulary building, BOW-based model is very suitable for STIP-based action recognition technique. Due to the human-guided classification strategy, supervised model-based classifiers are very compatible for both dense and sparse feature-based detector.

Laptev et al. played a vital role in the development of STIP-based approach. Laptev et al. [2] introduced an idea of Harris detector in spatio-temporal domain by utilizing the concept of Harris and Forstner interest point operators. It detected sparse feature set of each separate feature candidates in iterative fashion and handled very low number of feature sets for keeping the computation time under control. Dollar et al. [15] introduced 1D Gabor-filters and convolution of Gaussian filter for selection of interest points along with cuboid-based detector and descriptor. As specified in [15], due to the unavailability of true corner detector, direct 3D model of 2D interest point detectors was insufficient for the detection of spatio-temporal feature points.

For the first time, Willems et al. [12] detected scale-invariant (i.e., both spatially and temporally) spatio-temporal interest points using Hessian detector. It also performs real-time action recognition for low-resolution videos such as

Table 2 Related spatio-temporal papers

References	Year	Topic
Baumberg et al. [1]	1996	Generates spatio-temporal models and emphasized on tracking for deforming objects
Laptev et al. [36]	2003	Detects interest point and select scale in space–time
Laptev et al. [2]	2003	Detects interest points by extending the notion in spatio-temporal domain
Laptev et al. [37]	2003	Introduces velocity adaptation techniques in space–time for activity recognition
Laptev et al. [38]	2004	Introduces local descriptors to represent and recognize motion patterns
Laptev et al. [39]	2004	Introduces automatic velocity adaptation techniques for video representation
Dollar et al. [15]	2005	Recognized behavior using sparse features in spatio-temporal case
Laptev [14]	2005	Implementation of STIPs for compact representation of video data
Wong et al. [41]	2007	Extracts interest points using global features to identify moving parts
Laptev et al. [19]	2007	Adapt velocity in locally for event-based motion recognition in space–time domain
Niebles et al. [16]	2007	Unsupervised learning concept for action recognition using spatio-temporal words
Willems et al. [12]	2008	Introduces dense and scale-invariant STIP detector
Wang et al. [43]	2009	Evaluates local features for action recognition by comparing existing methods
Cao et al. [44]	2010	Action detection using multiple STIP features and focused on cluttered video
Kovashka et al. [17]	2010	Introduces discriminate space–time neighborhood features for action recognition
Yu et al. [20]	2010	Recognized action with temporal semantic and structural forests in real time
Matikainen et al. [45]	2010	Evaluates pairwise spatial and temporal relations for action recognition
Zhang et al. [18]	2011	Introduces 4-D local STIP features, combination of dense and intensity information
Yu et al. [46]	2012	Predicts human activities via STIPs detector by introducing forest structures
Yan et al. [42]	2012	Recognized human action using descriptor-based weighted-output classifier for STIPs
Chakraborty et al. [40]	2012	Introduces selective STIPs concept using local descriptor-based approach
Yuan et al. [3]	2013	Recognized actions using global feature-based STIPs with 3D R transform

Table 3 Systematic analysis of the reported papers on the basis of Fig. 2

References	STIPs detector	Feature descriptor	Vocabulary builder	Classifier	Remark
Laptev [14]	Sparse	Local	–	Semi-supervised	High computation time
Dollar et al. [15]	Sparse	Local	–	–	Features are not scale-invariant
Wong et al. [41]	Sparse	Global	–	Both	Unable to capture spatio-temporal distribution
Niebles et al. [16]	Sparse	Local	BOW	Unsupervised	The prescribed BOW model is too local
Willems et al. [12]	Dense	Local	BOW	Supervised	Difficult to handle for huge amount feature set
Cao et al. [44]	Both	Both	State space	Supervised	High computation time with low response rate
Kovashka et al. [17]	Both	Global	Extended BOW	Supervised	Unable to discriminate very similar action type
Yu et al. [20]	Dense	Local	State space	Supervised	Contemporary good
Matikainen et al. [45]	Sparse	Both	BOW	Supervised	Simple and computationally efficient
Zhang et al. [18]	Both	Local	BOW	Unsupervised	Vocabulary building section is too poor
Yu et al. [46]	Sparse	Local	BOW	Supervised	Nonsuitable for unsegmented video clips
Yan et al. [42]	Sparse	Global	BOW	Supervised	Unable to capture temporal information
Chakraborty et al. [40]	Sparse	Local	BOW	Supervised	Follow greedy approach
Yuan et al. [3]	Dense	Global	Mixing BOW	Supervised	Considerable performance on similar action type

the KTH human action dataset [56]. However, the time complexity depends on total number of available features. Chakraborty et al. [40] proposed a novel approach for detecting selective interest points. The system is robust against occlusion and jumbled background. Tables 4, 5, and 6 provide a brief overview of some popular features, detectors and descriptors.

Apart from STIP-based action recognition techniques, some researchers presented a lots of methodologies with reasonable performances. Scovanner et al. [47] proposed 3D

Table 4 Concise overview of some popular features

Features	Heart of the matter
HIPL	Histogram of interest point locations; generates polar coordinates; captures location information
BIP	Bag-of-interest points; Generates codebook for histogram of cuboids, cluster into several groups
PHOG	Pyramid histogram of oriented gradients, HOG computes in salient region, describes appearance and motion
PHOF	Pyramid histogram of optical flows, HOF computes in salient region, describes appearance and motion

Table 5 Concise overview of some popular detector

Detectors	Heart of the matter
Harris3D	Extended version of Harris detector; detects corner, edge and flat region using eigenvalue analysis
Cuboid detector	Based on Gaussian kernel (2D) and Gabor Filter (1D); detects local maxima of response function
Hessian detector	Extension of Hessian saliency measure; measures saliency with determinant of 3D Hessian matrix
Dense sampling	Extracts video blocks by locations and scales at five dimension, 3D in spatially and 2D in temporally

Table 6 Concise overview of some popular descriptors

Descriptors	Heart of the matter
Cuboid descriptor	Computes gradient of every single pixel in patch and minimize dimension by using PCA
HOG/HOF	Histograms of spatial gradient and optic flow; similar to SIFT descriptor
HOG3D	Based on histogram of 3D gradient orientation,3D SIFT descriptor; captures motion and shape data
ESURF	Extended SURF; detects robust local features in spatio-temporally

Table 7 Performance comparison of an STIP-based method with others methods by using recognition result

References	Weizmann dataset [55] (%)	KTH dataset [56] (%)
Scovanner et al. [47]	82.80	–
Gorelick et al. [55]	97.83	–
Laptev et al. [48]	–	91.8
Weinland et al. [49]	100	92.4
Mukherjee et al. [7]	–	92.8
Wang et al. [50]	–	94.4
Chakraborty et al. [13]	100	96.35

SIFT descriptor for action recognition by extending the idea of SIFT descriptor for 2D images to 3D video data. Due to the fast computation time and better classification strategy, 3D SIFT descriptor is better compared to 2D SIFT and 3D gradient magnitude. Gorelick et al. [55] introduced action as space–time shapes and evolved space–time feature using Poisson-based descriptor to show robust against partial occlusion and significant scale-changing operation. The space–time shape of an action is represented as a sequence of silhouettes of frames in a row at certain time period. Laptev et al. [48] introduced an interesting method for annotating automatically human action using movie scripts and subtitles by following 3D-Harris corner detector at multiple scales with HOG and HOF feature descriptor. This is an erroneous method and most of the errors come from the temporal alignment of the scripts. Weinland et al. [49] introduced a robust technique against occlusions and viewpoint changing operations using dense HOG3D descriptor for representing the video data. For action recognition in complex scene with dynamic background, Wang et al. [50] used feature trajectories (for action detection) and spatio-temporal tube of maximum mutual information (for action modeling) for keeping both time efficiency and localization accuracy. However, it is unable to discriminate the actions which have common motion pattern, as for example dialing a phone and answering a phone. Using pose information and motion pattern, Mukherjee et al. [7] evolved key (i.e., important or meaningful) pose-based action recognition method for repetitive motion pattern; however, it takes longer time to train and shorter time to test.

Table 7 brings to light the performance comparison of a famous STIP-based human action recognition technique [13] with the other detectors. The potential of Weinland et al. [49] and Chakraborty et al. [13] is almost identical on Weizmann [55] dataset; though it is the fact that Weinland et al. [49] and Chakraborty et al. [13] used global descriptor (i.e., HOG3D) and local descriptor (i.e., N-jet), respectively. In KTH dataset [56], Weinland et al. [49] and Mukherjee et al. [7] show almost same accuracy; however Mukherjee et al. [7] is more time efficient than Weinland et al. [49] dataset.

Moreover, Singh et al. [51] performed an effective research work regarding tracking and action recognition by utilizing the concepts of Dynamic Bayesian Action Network (DBAN) for 2D body model. Jiang et al. [52] introduced a hierarchical model for recognizing action with real-time processing. The video-based human action recognition system has exert influence on occlusions, view changing, fluctuating execution rate, anthropometry, etc. Ramanathan et al. [53] did a survey work on contemporary action recognition methods to take into account some of these potential challenges. Wu et al. [54] utilized audiovisual feature selection and fusion in realistic scenario for human action recognition.

As found in various state-of-art papers, the researchers recommended a lots of detectors, descriptors, and classifiers for recognizing the human action. However, it is seen that some datasets or application-specific detectors, descriptors or classifiers performed better among rest. As for example, cuboid detector [15] is performed better in Hollywood2 [59] and UCF dataset [57]; 3D-Harris corner detector is performed better in KTH dataset [56], etc. Moreover, the combination of HOG/HOF descriptor together with dense sampling has carried out always better result in Hollywood2 dataset [59] and dense sample together with 3D-HOG descriptor has operated better in UCF dataset [57]. Despite that among other descriptors, the gradient-based optic flow descriptor is always reliable for recognizing the action. In Wang et al. [43] discussed about experimental results on various state-of-art technology using KTH action dataset [56], UCF Sports action dataset [57] and Hollywood2 dataset [59]. In their experimental setup, the evolution method made use of four types of detectors (i.e., Harris3D, Cuboid, Hessian and Dense), five types of descriptors (i.e., HOG3D, HOG/HOF, HOF, cuboid, extended SURF or ESURF) and one classifier (i.e., SVM with x^2-kernel). It evaluated various results regarding dataset-specific parameters such as find best combination of detector and descriptor of a particular dataset, average accuracy of dense sampling of a particular dataset, average speed (frames/second) of feature detection of various descriptors, average number of features (features/frame), etc. As specified in [40], SVM with x^2-kernel gives almost cent percent average recognition accuracy in Weizmann dataset [55].

Table 8 represents a comparison analysis of major headway papers. The STIP detection ratio is evaluated on the basis of total number of detected STIPs on the actors with respect to the total number of STIPs in the background texture. It becomes more clear that Chakraborty et al. [40] evaluated more accurate and distinct interest points compared to others. Tables 9 and 10 confer performance comparison of variety of STIP-based techniques, and exhibit the ability of various reported techniques. These tables confer the idea that spatio-temporal based approaches are very suitable for repetitive action analysis. Table 11 represents average accuracy of the two famous feature detectors (i.e., Harris3D, Hessian) with some descriptors in KTH dataset [56]. In KTH dataset [56], HOF descriptor performs similar to HOG/HOF descriptor in both dense and sparse type of STIP detector. However, the performance of HOG is not up to the mark. Still, there are some loopholes for STIP-based techniques such as:

Local approach Generally, the STIP-based methods capture information from video sequences locally. There is also a gap for capturing global information more beneficially. In due course, it may fail for recognizing actions in complex scene.

Table 8 Comparison on the basis of STIP detection ratio in MSR 1 Dataset [60]

Method	Feature set	Detection (%)
Laptev et al. [2]	Sparse	18.73
Dallar et al. [15]	Sparse	21.36
Willems et al. [12]	Dense	24.02
Chakraborty et al. [40]	Sparse	76.21

Table 9 Performance comparison on the basis of action recognition ratio (%) in KTH dataset

Method	KTH dataset [56] (%)
Laptev et al. [39]	91.80
Cao et al. [44]	95.02
Kovashka et al. [17]	94.53
Wong et al. [41]	86.62
Niebles et al. [16]	81.50
Yu et al. (PSRM+BOST) [20]	95.67
Wang et al. [43]	92.1
Yan et al. (AdaBoost) [42]	93.98
Chakraborty et al. [40]	96.35
Yuan et al. [3]	95.49

Table 10 Performance comparison on the basis of action recognition ratio (%) in UCF sports action dataset

Method	UCF sports action [57] (%)
Wang et al. [43]	85.6
Kovashka et al. [17]	87.57
Yan et al. (AdaBoost) [42]	90.67
Yuan et al. [3]	87.33

Table 11 Average accuracy of two most famous feature detectors with some descriptors in KTH dataset [56]

Descriptor	Harris3D (sparse) (%)	Hessian (dense) (%)
HOG (local)	80.9	77.7
HOF (global)	92.1	88.6
HOG3D (global)	89.0	84.6
HOG/HOF (global)	91.8	88.7

Stability It is a fact that a maximum number of STIP-based methods are unstable due to the disparity of local characteristic of the detectors. Sometimes, the STIP-based methods have failed for detecting operation with moving cameras.

Redundancy Due to the property of local descriptors of surrounding image region, redundancy can occur.

However, the recognition of human action in real-world scenario is a very challenging task. By considering a lots of complex issues, the algorithmic complexity may increase. Still, application-based recognition techniques require some extra features apart from accurate recognition. Most of the reported methods were tested in some special dataset, which are captured in some controlled environment. Hence, it is very difficult for guessing their performances in real-world scenario. Some related complex issues are: (1) If occlusion occurs in some part of the body, then tracking of moving objects in different viewing direction are very difficult to perform. (2) A silhouette image where interior portion is featureless. The translation, rotation and scaling operations are not suitable for the silhouette images. (3) The complexity of an action depends on number of body parts involved in the action. (4) The speed or motion of different body parts is not in synchronized fashion. (5) Speed variation of actors. For example, it is very difficult to identify walking and slow running. (6) Phase and scale variation of actors. (7) For variation of light in image sequences, background subtraction is a very challenging task. (8) Unusual or excess clothing of actors. (9) Video captures in moving camera with variable speed or multi-view objects. (10) Physically challenged actor. For example, an actor having single leg moving with a crutch.

4 Related datasets for human action recognition

This section presents a synopsis about various related datasets for human action recognition. All the datasets are publicly available for research. These datasets are useful for the comparison of various techniques.

4.1 Weizmann Human action dataset [55]

The Weizmann human action dataset was introduced by Weizmann Institute of Science in 2001 and 2005. The Weizmann Institute of Science provided two datasets: Weizmann event-based analysis (ground truth: temporal annotation, in 2001) and Weizmann actions as space–time shapes (ground truth: silhouettes, in 2005). The key features of these two datasets are: low resolution and static background. The Weizmann event-based analysis consists of around six thousand frames and performing four activities: running in place, waving, running, and walking. Weizmann actions as space–time sh apes are a collection of ninety low-resolution video sequences and contain ten types actions; i.e., run, walk, skip, jumping-jack or shortly jack, jump-forward-on-two-legs or jump, jump-in-place-on-two-legs or pjump, gallopsideways or side, wave-two-hands or wave2, wave-one-hand or wave1, and bend. As specified in [16], contradiction may arise in the action classes skip, jump and run. Chakraborty et al. [40] obtained almost cent percent recognition accuracy with and without cross-data evaluation on this dataset. Due to the static background feature, the unsupervised learning method (i.e., pLSA, LDA) is also applicable for this dataset. Figure 4 provides a snapshot of this dataset.

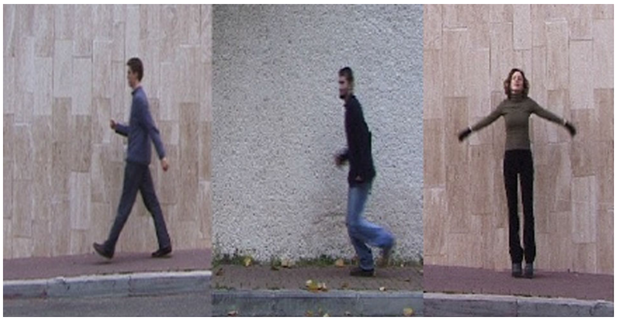

Fig. 4 Example of of Weizmann dataset (Walking, Running, and Jack)

4.2 KTH Human action dataset [56]

The Royal Institute of Technology in Sweden created KTH (in Swedish: Kungliga Tekniska hgskolan) dataset. It consists of 25 × 6 × 4 = 600 video sequences where 25 individuals performed 6 actions with 4 scenarios. The KTH video database contains six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping). All actions were performed several times by 25 subjects in four different scenarios: outdoors s1, outdoors with scale variation s2, outdoors with different clothes s3 and indoors s4. The ground truth of this dataset is simple action annotation. This dataset contains video with static camera and homogeneous background. Due to its features, NNMF has proven to give very good result on this dataset [41]. This dataset contains some actions which are very close. It is proven by the several researchers [12,16,20,41,42], because the similarity contradiction may arise on running and jogging action class. In addition to this, extended version of KTH dataset (i.e., Multi-KTH) was released in 2008. The key features of Multi-KTH dataset are multiple actors. Sparse feature-based detector (i.e., Harris3D) and local descriptor (i.e., N-jet)-based approach are proven to be very suitable for Multi-KTH dataset [40]. Figure 5 provides a snapshot of this dataset.

4.3 UCF Sports action dataset [57]

The Department of Computer Science and Electrical Engineering of University of Central Florida (UCF) created the UCF sports action dataset in 2008. This is real sport broadcasting videos. They provides various sports such as diving, golf swinging, kicking lifting, horseback riding, running,

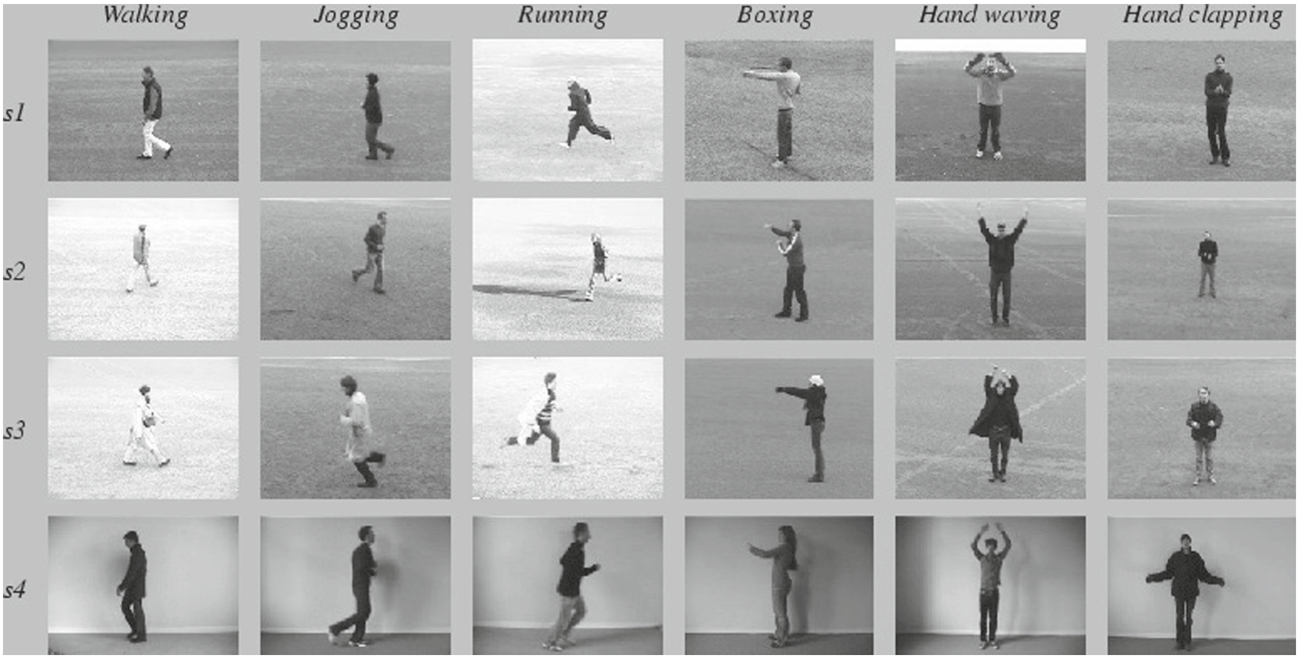

Fig. 5 Screenshots of KTH dataset

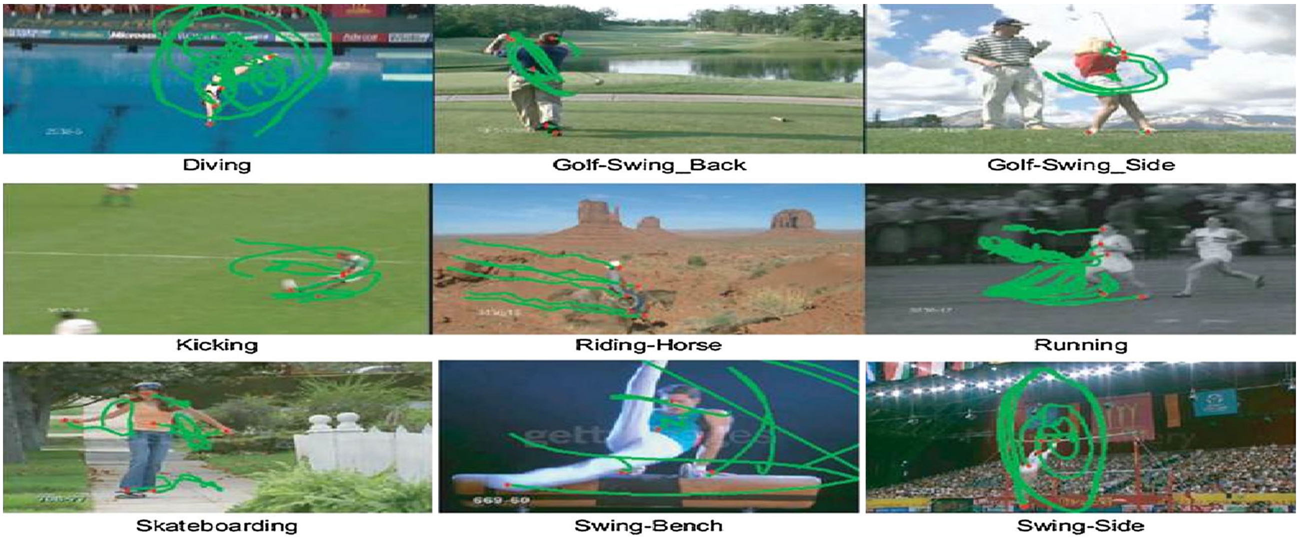

Fig. 6 Frame example of UCF sports action dataset

skating, swinging, and walking. The key features are: simple background, wide range of scenes and view points, simple action annotation. The sparse feature-based STIP detector is not suitable for this dataset due to the wide range of scenes and view points. Both local (i.e., HOG) and global descriptor (i.e., HOG3D, HOG /HOF) are compatible with this dataset. As specified by [43], the dense sampling with minimal spatial size gives good result for this dataset. Dense sampling gives a more complete description of motions and rich context information of the dataset. Figure 6 provides a snapshot of this dataset.

4.4 Hollywood and Hollywood2 human action database [58,59]

Hollywood (in 2008) and Hollywood2 (in 2009) datasets were created by IRISA institute in France. These datasets consist of short movie sequences. Hollywood action dataset contains eight actions (i.e., answer phone, get out of car, handshake, hug, kiss, sit down, sit up and stand up) from movies' extraction. Hollywood 2 dataset is a second version of the Hollywood dataset with the addition of dynamic background features. Hollywood2 dataset provides four additional actions (i.e., drive car, eat, fight, run). The key feature of these datasets are: large intra-class variability, multiple persons, challenging camera motion, rapid scene changes, unconstrained and cluttered background, label ambiguity, high quality. These dataset contains some special features such as expression, posture and clothing. Due to its complexity, this dataset is not so popular in action recognition field. However, these same features make the researchers to come face to face with real-life challenges. Both sparse (i.e., Harris3D, Cuboid) and dense feature detector (i.e., Hessian detector, dense sampling) are suitable for these datasets [43]. Because the local descriptor (like N-jet) did not provide good result in these datasets as found [40], however, due to the large intra-class variability and challenging camera motion, the global descriptor (i.e., HOG3D, HOG/HOF) is more suitable for these datasets. Figures 7 and 8 provide snapshots of these datasets.

Fig. 7 Video samples of Hollywood action dataset

Fig. 8 Video samples of Hollywood2 action dataset

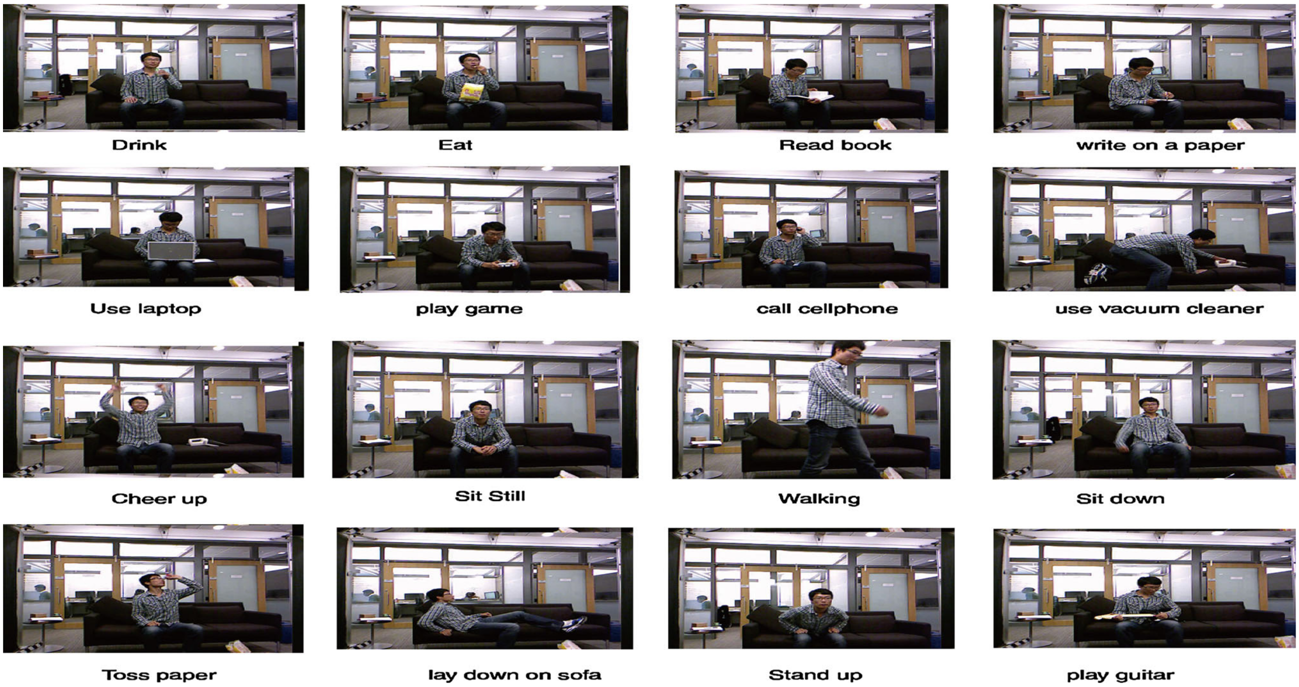

Fig. 9 One frame example of some actions from Microsoft Research action (MSR) action dataset

4.5 Microsoft Research (MSR) action dataset [60]

The Microsoft research team created MSR action dataset in 2009 for analyzing the behavior of human being. This dataset deals with realistic action with complex and non-static backgrounds. It provides both indoor and outdoor scene. The dataset contains 14 hand clapping, 24 handwaving and 25 boxing actions with multiple types of clutter and moving backgrounds. It is found that sparse feature detector (i.e., Harris 3D) together with local descriptor (i.e., N-jet) performed well on this dataset [40]. Figure 9 provides a snapshot of this dataset.

4.6 Action similarity LAbeliNg (ASLAN) dataset [61]

The ASLAN dataset was introduced by Kliper-Gross et al. [61]. This dataset contains 3631 unique action samples and 432 action classes. It provides both color and grayscale videos in AVI and mp4 format along with different types of resolution and aspect ratio. This dataset was captured in uncontrolled environment with pair-matching benchmark. It provides unified testing protocol for measuring the performances on various techniques. It is proven that this dataset is compatible with various STIP-based detector and local descriptor (i.e., HOG, HOF and combina-

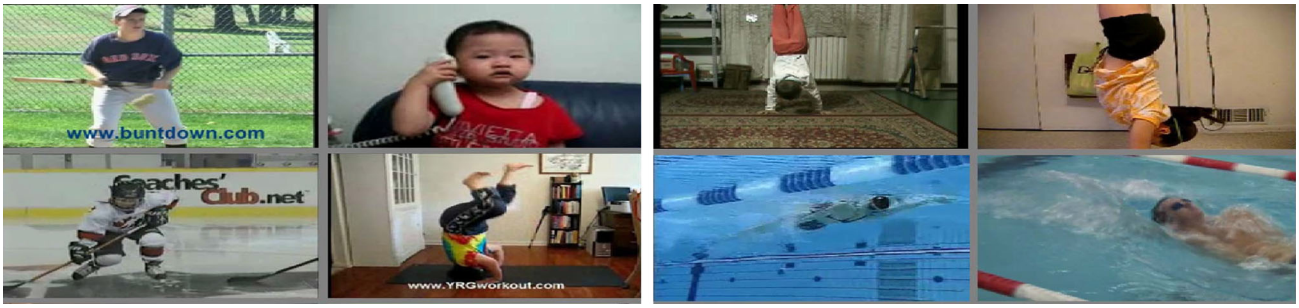

Fig. 10 Sample image frame from the ASLAN video dataset

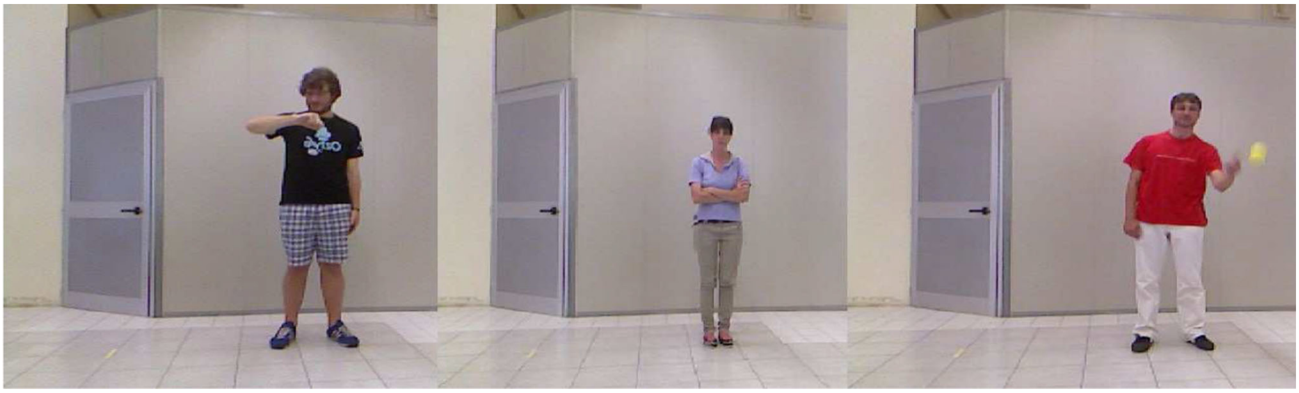

Fig. 11 Example video sequences of IAS-Lab action dataset (check watch, cross arms, and throw from bottom up)

Table 12 Datasets: brief overview

Name of the datasets	Year	Number of actors	Scenes	Camera movement	View type
Weizmann dataset [55]	2001	9	Outdoor	Static	Mono-view
KTH dataset [56]	2004	25	In/outdoor	Static	Mono-view
UCF action dataset [57]	2008	–	In/outdoor	Several	Mono-view
Hollywood human action [58]	2008	–	In/outdoor	Several	Mono-view
Multi-KTH dataset [56]	2008	5	outdoor	Moving	Multi-view
Hollywood2 dataset [59]	2009	–	In/outdoor	Several	Mono-view
MSR action dataset [60]	2009	10	In/outdoor	Static	Mono-view
ASLAN dataset [61]	2012	–	In/outdoor	Static	Multi-view
IAS-Lab action [62]	2013	12	Indoor	Static	Mono-view

tion of two) [61]. Figure 10 provides a snapshot of this dataset.

4.7 IAS-Lab action dataset [62]

Munaro et al. introduced IAS-Lab action dataset in 2013. The dataset provides 15 actions, performing 12 people with 540 samples video (RGB-Depth and gray scale); in addition with, it supplies skeleton pose for every frame. It contains 15 human actions such as CheckWatch, Cross arms, Get up, Kick, Pick up, Point, Punch, Scratch head, Sit down, Standing, Throw from bottom up, Throw over head, Turn around, Walk, and Wave. Munaro et al. made use of Microsoft Kinect sensor with tracking system (i.e., NITEs skeletal tracker) for detecting and tracking people in the scene. This dataset will be helpful for analyzing four-dimensional spatio-temporal features [18]. In essence, the four-dimensional STIP feature analysis deals with intensity and depth information. The contradiction may arise on the action classes such as Standing, Sit down and Get up [62]. Figure 11 provides a snapshot of this dataset.

Over the last decade, primitive datasets Weizmann [55] and KTH [56] are saturated to write about performance measurement on action recognition domain. The realistic datasets MSR action [60], Hollywood [58,59] and UCF sports action dataset [57] are the ideal for STIP-based action analysis. While these datasets are focused on atomic actions, the recent work ASLAN dataset [61] and IAS-Lab action [62] dataset are focused on complicated actions. Apart from these, some datasets are really important for this domain such as IXMAS dataset [63], i3DPost Multi-view [64], MuHAVi human action [65], VIRAT dataset [66]. Consequently, most of the datasets provide actions classes of mono-view type. However, the datasets [63–65] provide multi-view analysis of actions. By putting background clutter and diversity in the video data, VIRAT dataset [66] becomes challenging for the researcher community. Table 12 provides a concise overview about related datasets.

5 Conclusions

This paper presents a comprehensive review work on human action recognition based on spatio-temporal approaches. The gradual developments in the field of STIP-based detectors have been identified. Also the limitations of various methodologies proposed by different researchers have been summarized. Related benchmark databases are also reviewed in this context. STIP-based human action recognition is a promising field of research with several interesting applications like HCI, surveillance, health care systems, etc.

There remain various dynamic factors and complex issues in STIP-based approaches in real-life scenarios. Action recognition becomes very difficult for multiple moving objects in the presence of shadow, illumination changes in the scene. Segmentation of foreground objects from the background and correct localization of the objects in the video frame are also challenging tasks. Action recognition in multi-view moving objects is very promising in handling the issues involved in real-life scenarios. Another interesting extension of this approach is on human skeleton images produced by different motion sensor-based devices (e.g., KINECT, LEAP). It will be easier to detect STIPs in skeleton images with better accuracy. Therefore, it increases the rate of correct classification of different human actions.

References

1. Baumberg, A., Hogg, D.: Generating spatiotemporal models from examples. In: 6th British Machine Vision Conference, vol. 14, pp. 525–532 (1996)
2. Laptev, I., Lindeberg, T.: Space–time interest points. In: Proceedings ICCV'03, pp. 432–439. France (2003)
3. Yuan, C., Li, X., Hu, W., Ling, H., Maybank, S.: 3D R transform on spatio-temporal interest points for action recognition. In: CVPR'13 (2013)
4. Aggarwal, J.K., Cai, Q.: Human motion analysis: a review. In: IEEE Proceedings of Nonrigid and Articulated Motion Workshop, vol. 73, pp. 428–440. San Juan (1997)
5. Wu, J., Hu, D., Chen, F.: Action recognition by hidden temporal models. Vis. Comput. **30**(12), 1395–1404 (2013)
6. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: a review. ACM Comput. Surv. (CSUR). **43**(3), 16:1–16:43 (2011)
7. Mukherjee, S., Biswas, S.K., Mukherjee, D.P.: Recognizing human action at a distance in video by key poses. IEEE Trans. Circuits Syst. Video Technol. **21**(9), 1228–1241 (2011)
8. Moravec, H.: Obstacle avoidance and navigation in the real world by a seeing robot rover. In:tech. report CMU-RI-TR-80-03, Robotics Institute, Carnegie Mellon University & doctoral dissertation, Stanford University (1980)
9. Harris, C., Stephens, M.: A combined corner and edge detector. In: Proceedings of Fourth Alvey Vision Conference, pp. 147–151 (1988)
10. Fstner, M.A., Glch, E.: A fast operator for detection and precise location of distinct Points, corners and centers of circular features. In: ISPRS Intercommission Workshop (1987)
11. Li, Y., Kuai, Y.: Action recognition based on spatio-temporal interest points. In: Proceedings on 5th International Conference on Bio-Medical Engineering and Informatics (BMEI) (2012)
12. Willems, G., Tuytelaars, T., Van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. Comput. Vis. ECCV **5303**, 650–663 (2008)
13. Chakraborty, B., Holte, M.B., Moeslund, T.B., Gonzalez, J., Xavier Roca, F.: A selective spatio-temporal interest point detector for human action recognition in complex scenes. In: IEEE International Conference on Computer Vision (ICCV), pp. 1776–1783 (2011)
14. Laptev, I.: On space-time interest points. Int. J. Comput. Vis. **64**, 107–123 (2005)
15. 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)

16. Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial–temporal words. Int. J. Comput. Vis. **79**, 299–318 (2007)
17. Kovashka, A., Grauman, K.: Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: Proceedings on IEEE Conference, Computer Vision and Pattern Recognition (CVPR), pp. 2046–2053. San Francisco (2010)
18. Zhang, H., Parker, L.E.: 4-Dimensional local spatio-temporal features for human activity recognition. In: Proceedings of IEEE International Conference on Intelligent Robots and Systems, pp. 2044–2049. San Francisco (2011)
19. Laptev, I., Caputo, B., Schuldt, C., Lindeberg, T.: Local velocity-adapted motion events for spatio-temporal recognition. Comput. Vis. Image Underst. **108**, 207–229 (2007)
20. Yu, T.-H., Kim, T.-K., Cipolla, R.: Real-time action recognition by spatiotemporal semantic and structural forests. In: Proceedings of the British Machine Vision Conference, pp. 52.1–52.12. BMVA Press (2010)
21. Vishwakarma, S., Agrawal, A.: A survey on activity recognition and behavior understanding in video surveillance. Vis. Comput. **29**, 983–1009 (2012)
22. Iosifidis, A., Tefas, A., Pitas, I.: Multi-view human action recognition: a survey In: Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp. 522–525. Beijing (2013)
23. Akila, K., Chitrakala, S.: A comparative analysis of various representations of human action recognition in a video. Int. J. Innov. Res. Comput. Commun. Eng. **2**(1), 2829–2837 (2014)
24. Claudette, C., Shah, M.: Motion-based recognition: a survey. Image Vis. Comput. **13**(2), 129–155 (1995)
25. Moeslund, T.B., Hilton, A., Kruger, V.: A survey of advances in vision-based human motion capture and analysis. Special Issue on Modeling People: Vision-based understanding of a person's shape, appearance, movement and behavior, vol. 104, pp. 90–126 (2006)
26. Gavrila, D.M.: The visual analysis of human movement: a survey. Comput. Vis. Image Underst. **73**, 82–98 (1999)
27. Krger, V., Kragic, D., Geib, C.: The meaning of action: a review on action recognition and mapping. Adv. Robot. **21**(13), 1473–1501 (2007)
28. Poppe, R.: A survey on vision-based human action recognition. Image Vis. Comput. **28**, 976–990 (2010)
29. Weinlanda, D., Ronfardb, R., Boyer, E.: A survey of vision-based methods for action representation, segmentation and recognition. Comput. Vis. Image Underst. **115**, 224–241 (2011)
30. Turaga, P., Chellappa, R., Subrahmanian, V.S., Udrea, O.: Machine recognition of human activities: a survey. IEEE Trans. Circuits Syst. Video Technol. **18**(11), 1473–1488 (2008)
31. Ke, S.-R., Thuc, H.L.U., Lee, Y.-J., Hwang, J.-N., Yoo, J.-H., Choi, K.-H.: A review on video-based human activity recognition. Act. Detect. Nov. Sens. Technol. **2**, 88–131 (2013)
32. Guan, D., Ma, T., Yuan, W., Lee, Y.-K., Jehad Sarkar, A.M.: Review of sensor-based activity recognition systems. IETE Tech. Rev. (Medknow Publications and Media Pvt. Ltd.) **28**, 418 (2011)
33. Xu, X., Tang, J., Zhang, X., Liu, X., Zhang, H., Qiu, Y.: Exploring techniques for vision based human activity recognition: methods, systems, and evaluation. Sensors **13**, 1635–1650 (2013)
34. Chaquet, J.M., Carmona, E.J., Fernndez-Caballero, A.: A survey of video datasets for human action and activity recognition. Comput. Vis. Image Underst. **117**, 633–659 (2013)
35. Hassner, T.: A critical review of action recognition benchmarks. In: IEEE Conference in Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 245–250. Portland (2013)
36. Laptev, I., Lindeberg, T.: Interest point detection and scale selection in space-time. In: Proceedings of 4th International Conference, pp. 372–387, UK (2003)
37. Laptev, I., Lindeberg, T.: Velocity adaptation of spatio-temporal receptive fields for direct recognition of activities: an experimental study. In: ECCV'02 workshop on Statistical Methods in Video Processing, pp. 61–66. Copenhagen (2003)
38. Laptev, I., Lindeberg, T.: Local descriptors for spatio-temporal recognition. In: First International Workshop, SCVMA, pp. 91–103. Prague (2004)
39. Laptev, I., Lindeberg, T.: Velocity adaptation of space-time interest points. In: Proceedings of the 17th International Conference on Pattern Recognition, ICPR, pp. 52–56 (2004)
40. Chakraborty, B., Holte, M.B., Moeslund, T.B., Gonzalez, J.: Selective spatio-temporal interest points. In: Special issue on Semantic Understanding of Human Behaviors in Image Sequences, vol. 116(3), pp. 396–410 (2012)
41. Wong, S.-F., Cipolla, R.: Extracting spatiotemporal interest points using global information. In: Proceedings on 11th IEEE International Conference of Computer Vision, ICCV. Rio de Janeiro, pp. 1–8 (2007)
42. Yan, X., Luo, Y.: Recognizing human actions using a new descriptor based on spatialtemporal interest points and weighted-output classifier. J. Neurocomput. **87**, 51–61 (2012)
43. Wang, H., Ullah, M.M., Klser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: Proceedings British Machine Vision Conference, pp. 1–18 (2009)
44. Liangliang, C., Tian, Y.L., Liu, Z., Yao, B., Zhang, Z., Huang, T.S.: Action detection using multiple spatial–temporal interest point features. In: International Conference on Multimedia and Expo, pp. 340–345. IEEE (2010)
45. Matikainen, P., Hebert, M., Sukthankar, R.: Representing pairwise spatial and temporal relations for action recognition. In: Proceedings on 11th European conference of the Computer vision: Part I, pp. 508–521. ECCV (2010)
46. Yu, G., Yuan, J., Liu, Z.: Predicting human activities using spatio-temporal structure of interest points. In: Proceedings of the 20th ACM international conference on Multimedia, pp. 1049–1052. New York (2012)
47. Scovanner, P., Ali, S., Shah, M.: A 3-Dimensional SIFT descriptor and its application to action recognition. In: Proceedings of the 15th international conference on Multimedia, pp. 357–360 (2007)
48. Laptev, I., Marszaek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2008)
49. Weinland, D., Ozuysal, M., Fua, P.: Making action recognition robust to occlusions and viewpoint changes. In: 11th European Conference on Computer Vision, pp. 635–648 (2010)
50. Wang, T., Wang, S., Ding, X.: Detecting human action as the spatio-temporal tube of maximum mutual information. IEEE Trans. Circuits Syst. Video Technol. **24**(2), 277–290 (2013)
51. Singh, V.K., Nevatia, R.: Simultaneous tracking and action recognition for single actor human actions. Vis. Comput. **27**(12), 1115–1123 (2011)
52. Jiang, X., Zhong, F., Peng, Q., Qin, X.: Online robust action recognition based on a hierarchical model. Vis. Comput. **30**(9), 1021–1033 (2014)
53. Ramanathan, M., Yau, W.-Y., Teoh, E.K.: Human action recognition with video data: research and evaluation challenges. IEEE Trans. Hum. Mach. Syst. **44**(5), 650–663 (2014)
54. Qiuxia, W., Wang, Z., Deng, F., Chi, Z., Feng, D.D.: Realistic human action recognition with multimodal feature selection and fusion. IEEE Trans. Syst. Man Cybern. Syst. **43**(4), 875–885 (2013)
55. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space–time shapes. Trans. Pattern Anal. Mach. Intell. **29**(12), 2247–2253 (2007)
56. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In:17th International Conference on ICPR 3, pp. 32–36 (2004)

57. Rodriguez, M.D., Ahmed, J., Shah, M.: Action MACH: a spatio-temporal maximum average correlation height filter for action recognition. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (2008)
58. Laptev, I., Marszaek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
59. Marszałek, M., Laptev, I., Schmid, C.: Actions in context. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2929–2936 (2009)
60. Yuan, J., Liu, Z., Wu, Y.: Discriminative subvolume search for efficient action detection. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)
61. Kliper-Gross, O., Hassner, T., Wolf, L.: The action similarity labeling challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 34(3) (2012)
62. Munaro, M., Ballin, G., Michieletto, S., Menegatti, E.: 3D flow estimation for human action recognition from colored point clouds. In: Biologically Inspired Cognitive Architectures (BICA) vol. 5, pp. 42–51 (2013)
63. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. Comput. Vis. Image Underst. **104**(2–3), 249–257 (2006)
64. Gkalelis, N., Kim, H., Hilton, A., Nikolaidis, N., Pitas, I.: The i3DPost multi-view and 3D human action/interaction. In: Conference for Visual Media Production, pp. 159–168 (2009)
65. Singh, S., Velastin, S.A., Ragheb, H.: MuHAVi: A multicamera human action video dataset for the evaluation of action recognition methods. In: 2nd Workshop on Activity monitoring by multi-camera surveillance systems(AMMCSS), pp. 48–55 (2010)
66. Oh, S., Hoogs, A., Perera, A., Cuntoor, N., et al.: A large-scale benchmark dataset for event recognition in surveillance video. In: Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR) (2011)

Debapratim Das Dawn completed his school education from Ramakrishna Mission Vidyapith, Purulia, West Bengal, India. He received his B.Tech. degree in computer science and engineering from the West Bengal University of Technology, West Bengal, India, and M.Tech. degree in computer science and application from the University of Calcutta, West Bengal, India. His research interests include computer vision and image processing.

Dr. Soharab Hossain Shaikh is an assistant professor at the Department of Computer Science and Engineering, NIIT University, Neemrana. Before joining NU, Dr. Shaikh has worked as a faculty member at A. K. Choudhury School of Information Technology, University of Calcutta, for almost 6 years. He has also taught at the UG-level B.Tech. for 2 years in engineering/technology institutions in West Bengal. After receiving his B.Sc. degree with honours in computer science in 2001, Soharab completed M.Sc. in computer and information science in 2003, followed by M.Tech. in computer science and engineering in 2005 from the University of Calcutta. He received his Ph.D. in computer science and engineering from the Department of Computer Science and Engineering, University of Calcutta, in 2014. Earlier, Dr. Shaikh had also received a fellowship from the Italian Ministry of Education and Research (MIUR) for pursuing research work at Ca' Foscari, University of Venice, Italy, in 2006-2007. His research interests include image segmentation, biometrics hand gesture recognition for HCI, emotion recognition, etc. He works in active collaboration with AGH University of Science and Technology, Bialystok Technical University (Poland), University of Calcutta, (India), University of Warwick (UK), etc. Soharab jointly holds a US patent. He has co-authored two books published by Springer-Verlag. He has served as the guest editor/reviewer/committee member in a number of international conferences/symposiums and journals. Dr. Shaikh has published a number of research papers in peer-reviewed journals/conferences. He is also a member of IEEE Computer Society and ACM Kolkata Chapter.