Egyptian Informatics Journal (2015) 16, 187–198

Cairo University

Egyptian Informatics Journal

www.elsevier.com/locate/eij
www.sciencedirect.com

EGYPTIAN Informatics JOURNAL

ORIGINAL ARTICLE

Human action recognition using trajectory-based representation

Haiam A. Abdul-Azim [a,*], Elsayed E. Hemayed [b]

[a] Physics Department, Faculty of Women for Arts, Science and Education, Ain Shams University, Egypt
[b] Computer Engineering Department, Faculty of Engineering, Cairo University, Egypt

Received 16 January 2015; accepted 27 May 2015

KEYWORDS

Human action recognition;
Spatio-temporal features;
Cuboid detector;
Trajectory-based feature description;
Bag-of-Words

Abstract Recognizing human actions in video sequences has been a challenging problem in the last few years due to its real-world applications. A lot of action representation approaches have been proposed to improve the action recognition performance. Despite the popularity of local features-based approaches together with "Bag-of-Words" model for action representation, it fails to capture adequate spatial or temporal relationships. In an attempt to overcome this problem, a trajectory-based local representation approaches have been proposed to capture the temporal information. This paper introduces an improvement of trajectory-based human action recognition approaches to capture discriminative temporal relationships. In our approach, we extract trajectories by tracking the detected spatio-temporal interest points named "cuboid features" with matching its SIFT descriptors over the consecutive frames. We, also, propose a linking and exploring method to obtain efficient trajectories for motion representation in realistic conditions. Then the volumes around the trajectories' points are described to represent human actions based on the Bag-of-Words (BOW) model. Finally, a support vector machine is used to classify human actions. The effectiveness of the proposed approach was evaluated on three popular datasets (KTH, Weizmann and UCF sports). Experimental results showed that the proposed approach yields considerable performance improvement over the state-of-the-art approaches.

1. Introduction

Over the past years, human action recognition in videos has been a growing field of research in computer vision with many real-world applications, such as video surveillance, video indexing/browsing, recognizing gestures, human–computer interfacing and analysis of sport-events. However, it is still a challenging problem because of cluttered backgrounds, illumination changes, different physiques of humans, variety of

[*] Corresponding author. Tel.: +20 1272588373; fax: +20 224157804.
E-mail addresses: haiamadel@yahoo.com (H.A. Abdul-Azim), hemayed@ieee.org (E.E. Hemayed).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

ELSEVIER Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.eij.2015.05.002
1110-8665 © 2015 Production and hosting by Elsevier B.V. on behalf of Faculty of Computers and Information, Cairo University.

clothing, camera motion, partial occlusions, viewpoint changes, scale variation of video screen, etc.

Generally, human action recognition procedure consists of two main steps: action representation, and action learning and classification. Existing action recognition approaches are classified by Weinland et al. [1] based on action representation into two main approaches: global and local representations. Global representation approaches focus on detecting the whole body of the person by using background subtraction or tracking. Silhouettes, contours or optical flow are usually used for representing the localized person. These representations are more sensitive to viewpoint changes, personal appearance variations and partial occlusions.

In local representation approaches, videos are represented as a collection of small independent patches. These patches involve the regions of high variations in spatial and time domains. Centers of the patches are called spatio-temporal interest points (STIPs). The detected points are described by capturing the appearance and/or motion information from their patches and clustered to form a dictionary of prototypes or visual-words. Each action sequence is then represented by Bag of Words model (BOW) [2]. Recently, these approaches have become very successful approaches for human action recognition. They overcome some limitations of global representation such as sensitivity to noise and partial occlusion and the necessity of accurate localization by background subtraction and tracking.

Several STIPs detectors have been proposed to determine the spatio-temporal interest locations in videos. For example, Laptev [3] extended Harris corner detector for the spatio-temporal case and propose Harris3D detector, Dollar et al. [4] proposed the Cuboid detector by applying 1-D Gabor filters temporally, Willems et al. [5] proposed Hessian detector which measures the saliency with the determinant of the 3D Hessian matrix, and Wang et al. [6] introduced Dense sampling detector that extracts STIPs at regular positions and scales in space and time. Also, various descriptors for STIPs have been proposed such as Gradient descriptor [4], Histogram of Oriented Gradients (HOG) and Histogram of Optical Flow (HOF) descriptors [2], 3D Scale-Invariant Feature Transform (3D SIFT) [7], 3D Gradients descriptor (HOG3D) [8] and the extended Speeded Up Robust Features descriptor (ESURF) [5].

Despite the popularity of local representation approaches, it has some drawbacks. One of the main drawbacks is the ignorance of spatial and temporal relationships between local features. This may be a major problem in human action recognition. The spatial and/or temporal connections between the detected low-level action parts are necessary to introduce intrinsic characteristics of actions. A lot of attempts have been made to weaken this limitation of the local representation approaches based on BOW model. To capture these relationships early work was introduced such as Laptev et al. [2], Liu and Shah [9], Gilbert et al. [10], Zhang et al. [11] and Bregonzio et al. [12].

This paper introduces an enhancement of the recently proposed approaches which called trajectory-based local representation approaches [13–19]. These approaches capture some temporal relationships between the detected interest points by tracking them throughout the video. They differ in the trajectory generation and representation methods used. In this framework, we track the STIPs detected by Cuboid detector using SIFT-matching, then after some refinement we use the tracked points to form the action trajectories and then we describe the volumes around these trajectories' points. These features are represented with a Bag-of-Words (BOW) model. Finally, human actions are classified using a Support Vector Machine. In order to evaluate the proposed approach, we train and recognize action models on three popular datasets, KTH [20], Weizmann [21] and UCF Sports [22].

This paper is organized as follows. Section 2 reviews the previous related work. Section 3 describes the proposed trajectory-based approach for video representation. Section 4 presents the experimental setup, datasets and discusses the obtained results. Finally, Section 5 concludes the paper.

2. Related work

Recent works [13–19] show good results for action recognition in which the local spatio-temporal volumes are determined by using the trajectories of the interest points through video sequences. These trajectory-based approaches leverage the motion information extracted from the spatio-temporal volumes and utilize different methods for representation. Messing et al. [13] allowed a bag-of-features model to capture a significant amount of non-local structure by tracking Harris3D interest points [3] with the Pyramid Lucas–Kanade–Tomasi (KLT) tracker [23]. For action classification, Trajectories are represented by computing quantized velocities over time which called "velocity history". Matikainen et al. [14] introduced a trajectory-based motion features which called "trajectons". Trajectories are produced using a standard KLT tracker. For trajectories representation, clustering trajectories is performed using K-means and then an affine transformation matrix is computed for each cluster center. Sun et al. [15] generated their trajectories by matching SIFT descriptors between consecutive frames based on a unique-match constraint that yields good motion trajectories. Actions are then described in a hierarchical way where three levels of context information are exploited. Sun et al. [16] also extracted long-duration trajectories by combining both KLT tracker and SIFT descriptor matching. Moreover, random points are sampled for tracking within the region of existing trajectories to capture more salient image structures. For action representation, spatio-temporal statistics of the trajectories are used. Raptis and Soatto [17] proposed spatio-temporal feature descriptors that capture the local structure of the image around trajectories. These descriptors are a computation of HOG or HOF descriptors along the trajectories. The final descriptor is applied in action modeling and video analysis. Bregonzio et al. [18] presented an action representation based on fusing the trajectories generated by both KLT tracker and SIFT matching with the extracted spatio-temporal local features. This fusion enhanced the trajectory-based action representation approaches to be able to recognize actions under realistic conditions such as small camera movement, camera zooming, and shadows. Wang et al. [19] introduced dense-trajectories and used the motion boundary histograms (MBH) [24] as a trajectory descriptor. Points are detected by dense sampling detector and checked by Shi and Tomasi criterion [25] then tracked using a dense optical flow field. Trajectories are described with four different descriptors (trajectory shape, HOG, HOF and MBH).

In order to generate reliable and robust trajectories, it is necessary to extract good keypoints and track them accurately. In [14–17], trajectories are extracted by tracking spatially salient points, such as SIFTs or corners. However, in low-resolution videos with cluttered background and fast motion, the number of resulting keypoints is far from sufficient which increases redundancy and noise level. Due to its spares representation base, tracking Harris3D STIPs as in [13] enhances the generated trajectories but using KLT tracker does not guarantee the corresponding point in the next frame is a feature point. Also, the temporal scalability of 3D interest points is limited to capture only the short (simple) movements. Finally, tracking the densely sampled STIPs achieved a great success [19], but still not discriminative enough because of the large number of extracted trajectories and its computationally consumption. The effect of using optical flow in tracking is decreased in dense trajectories due to the large number of feature points.

In this paper, we go further on the trajectory-based approaches for a local representation of human actions. Our trajectories are characterized by its generation. We select the cuboid detector as a point detector to generate denser trajectories and matching the spatial information of the detected STIPs over consecutive framers to extract their actual movements. Furthermore, to overcome the short temporal scalability of the detected STIPs, the flow vectors computation is used as a linker and/or explorer for short trajectories.

3. Proposed approach

The goal of the proposed approach is to represent actions by extracting local spatio-temporal features with rich motion information from videos. STIPs are detected, as a first step, using Cuboid detector [4]. Then, around every point, 2D SIFT descriptor [26] is computed to build trajectories using SIFT-matching between STIPs in consecutive frames. The extracted trajectories are enhanced and described. Finally, each video sequence is represented using the Bag of features models. The system architecture is illustrated in Fig. 1.

3.1. Feature extraction

In this approach, we initially extract STIPs in each video sequence. Among different detectors that have been proposed in the last few years, we have selected Cuboid detector proposed by Dollar et al. [4]. It overcomes the drawback of the sparsity STIP detectors (e.g. Harris3D and 3D Hessian) which detect small number of stable interest points detected [1]. Furthermore, the sampling of spatial–temporal volume of Cuboid detector is denser than other dense sampling detectors [6]. Dollar et al. [4] detected the locally periodic motion by applying a set of separable linear filters (Gaussian filter applied on the spatial domain and 1-D Gabor filter applied temporally). The response function for a stack of images denoted by $I(x,y,t)$ is given by

$$R = (I * g * h_{even})^2 + (I * g * h_{odd})^2 \tag{1}$$

where $g(x,y;\sigma)$ is the 2D spatial Gaussian smoothing kernel, and h_{even} and h_{odd} are a quadrature pair of 1D Gabor filters which are applied temporally. They are defined as

$$h_{even}(t;\tau,\omega) = -\cos(2\pi t\omega)e^{-t^2/\tau^2}, \text{and}$$

$$h_{odd}(t;\tau,\omega) = -\sin(2\pi t\omega)e^{-t^2/\tau^2} \tag{2}$$

with $\omega = 4/\tau$. The two parameters σ and τ of the response function R correspond roughly to the spatial and temporal scales of the detector. The local maxima of the response function are the interest points. These interest points correspond to the local regions where the complex motion patterns are occurred. In the presented experiments, we detect the interest points at multiple scales of $\sigma = \{2,\ 4\}$ as spatial scales and $\tau = \{2,\ 4\}$ as temporal scales for each video sequence to capture the essence of human action.

3.2. Trajectory-based feature description

The second stage in the presented approach consists of four steps to describe the detected interest points by their trajectories throughout the video frames. These steps are described as follows:

3.2.1. Trajectories generation

Trajectories are extracted from each video by describing the spatial information of the detected STIPs using SIFT descriptor [26] and then searching for their matches between the consecutive frames. Due to its robustness to the variations of viewing conditions, SIFT descriptors were used for STIP matching and tracking in consecutive frames. To mitigate the effect of incorrect matches we follow the windowing approach for matching proposed by Sun et al. [15]. Such approach considers that any point p in frame i can be matched with one candidate point p_0 in frame $i + 1$ and must be located within a spatial window $M \times M$ around point p. This consideration discards the matches that are too far apart based on the observation of most realistic motions which cannot be very fast. The generated trajectories by this windowing approach may automatically end when reaching the shot boundaries or with considerable occlusions. Experimental results showed that the spatial matching window of size 32×32 gives good results over the tested dataset.

3.2.2. Trajectories enhancement (linking and exploring)

After tracking STIPs over all frames, motion trajectories of varying lengths are generated. Examples for generated trajectories are shown in Fig. 2. It can be observed that there are very short and long trajectories generated for motion representation. Due to the matching errors, occlusions of interest points, camera motion, viewpoint changes, and scales variation of video screen, short trajectories can be treated as uncompleted trajectories. Therefore, the future of these uncompleted trajectories must be explored.

The exploration process is carried out by estimating the optical flow vectors for the next frame by the KLT method [23] within a window of size $W \times W$. Then, describing the estimated location by SIFT descriptor and again searching for its match in the next frames. Finding a match means that a linking between two trajectories has been performed. The exploration by KLT method continues till finding SIFT match or reaching to a limited length of frames. Fig. 3 illustrates the proposed linking and exploration approach. Experimental

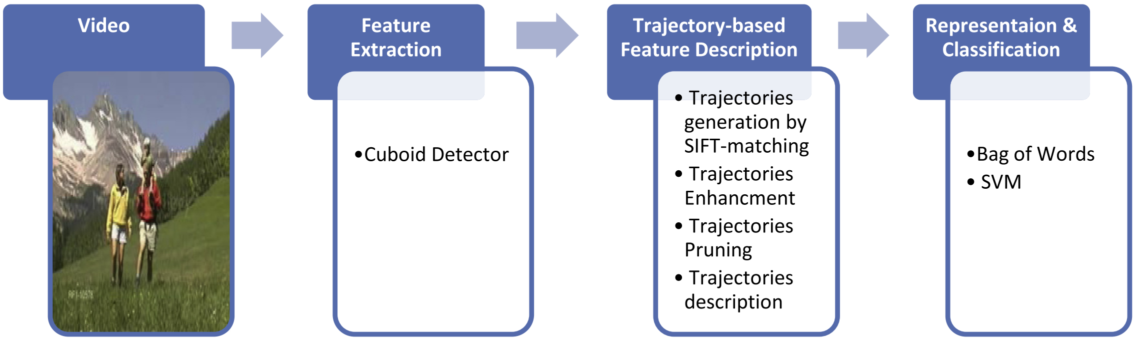
Video
Feature Extraction
•Cuboid Detector
Trajectory-based Feature Description
• Trajectories generation by SIFT-matching
• Trajectories Enhancment
• Trajectories Pruning
• Trajectories description
Representaion & Classification
•Bag of Words
• SVM

Figure 1 Diagram of the proposed method for human action recognition.

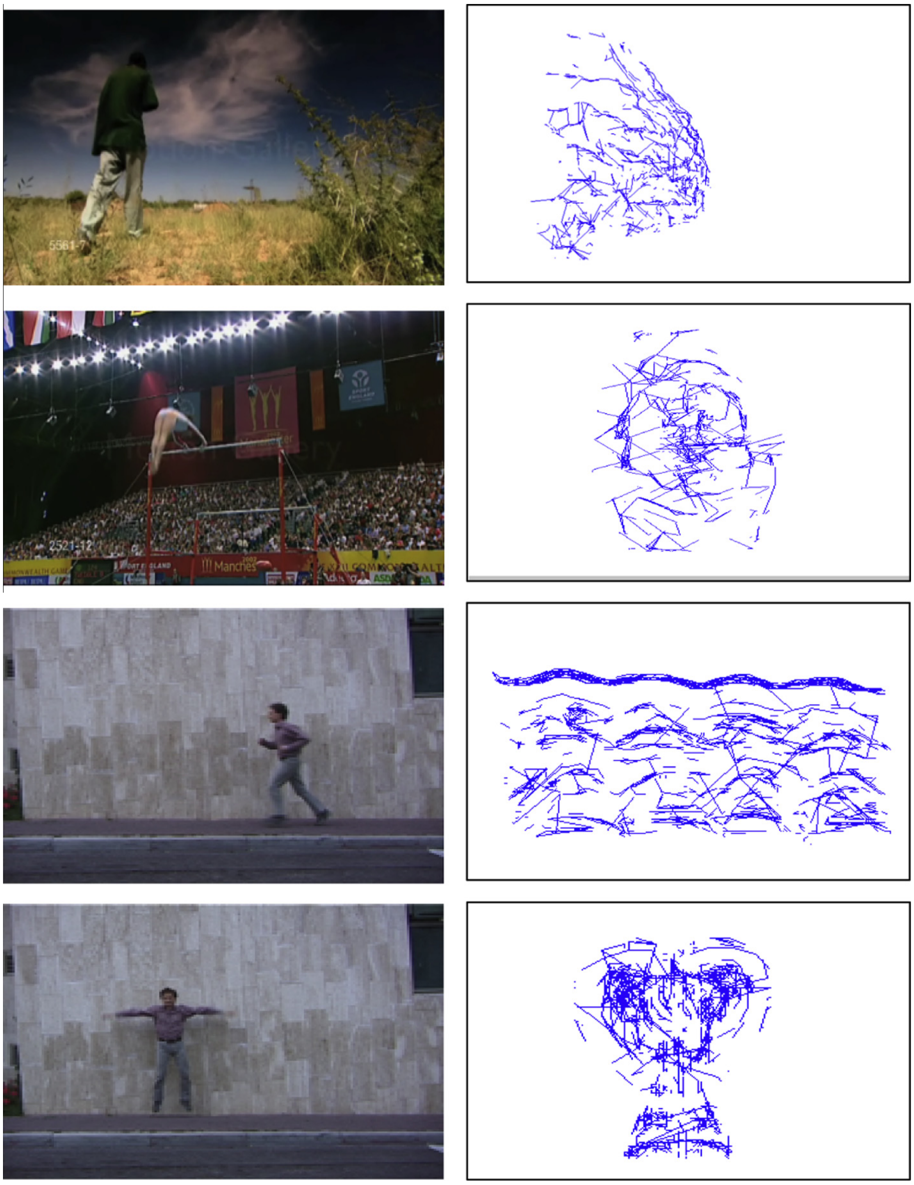

Figure 2 Examples of generated trajectories for four actions. The first row is for walking action, the second row is for swinging bar sport, the third row is for running action and the fourth is for jacking action.

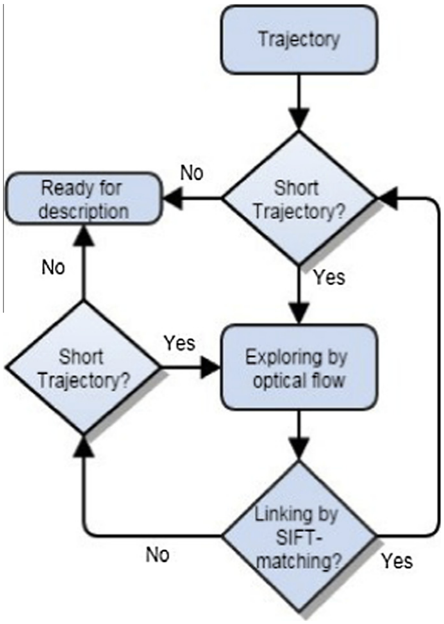

Figure 3 Illustration of linking and exploring process.

results showed that $W = 10$ gives good results over the tested dataset.

In all presented experiments, trajectories less than 10 frames length are considered as uncompleted trajectories and thus need to be linked or explored. The exploration process is terminated after the length of trajectory reaches 10 frames. Fig. 4 shows the generated trajectories before and after linking and exploring. From Fig. 4, it can be observed that the proposed linking and exploring method performs well under different video conditions such as varying video scales as shown in the first row and complex background with small camera motion as shown in the second row. Furthermore, it generates reliable trajectories under controlled video environments as shown in the last two rows in Fig. 4.

3.2.3. Trajectories pruning

After linking and exploring the trajectories, short trajectories still exist. These trajectories can be considered as noisy tracks and can be pruned. Moreover, static trajectories that do not have motion information are removed. Also, trajectories with sudden large displacements are pruned. We follow Wang et al. approach [19] to detect such trajectories by measuring the displacement vector between two consecutive trajectory points. The trajectory is removed if any displacement is larger than 70% of the overall displacement. The effect of pruning stage is illustrated in Fig. 5. Finally, trajectories longer than a pre-defined minim length (10 frames) are automatically segmented into smaller fixed length trajectories ($L = 10$ frames). Again any short trajectory segments are removed. Now our trajectories are ready for description.

3.2.4. Trajectories description

In this work we adopted the approach of Wang et al. [19] to describe our trajectories. For each motion trajectory, four descriptors were computed: Trajectory shape descriptor [19], HOG (histograms of oriented gradients) [2], HOF (histograms of optical flow) [2] and MBH (motion boundary histograms) [19]. Each descriptor captures some specific characteristics of the video content.

In our experiments, as we fix the trajectory length to 10 frames, the final dimension of the trajectory shape descriptor is 18. It captures the shape information of a trajectory by computing its displacement vectors. Moreover, a volume along the extracted trajectories of size $32 \times 32 \times 10$ and a grid size $3 \times 3 \times 2$ are used for HOG, HOF and MBH descriptors. To describe the appearance information inside the volumes, 4-bins histogram of gradient orientations (HOG) is computed for each cell which yields a feature vector of length 72. Also, 5-bins histogram of optical flow (HOF) is used to quantize the grid cells to capture the motion information. The HOF descriptor vector length is 90. MBH is another motion descriptor for trajectories which describes the motion in x and y directions respectively. For each direction, 5-bins histogram of optical flow is used to quantize the grid cells of trajectory volume. The final feature vector dimension is 180.

3.3. Representation & classification

We follow the general framework of the local-based action representation approaches and use the standard bag of words model for video representation. The visual words (or codebook) are built from training data by using k-means clustering algorithm with the Euclidean distance. Then, each video is represented as the frequency histogram of the codebook elements. In our experiments, the codebook is built on a subset of 60,000 randomly selected training features to limit the complexity. Due to the random initialization of k-means clustering algorithm, we report the best result over 10 runs. When using different descriptors for action description, the codebook for each one is built separately.

For classification, we use a non-linear support vector machine (SVM) with the χ^2 kernel [2];

$$K(H_i, H_j) = exp\left(-\frac{1}{2A}\sum_{n=1}^{V}\frac{(h_{in} - h_{jn})^2}{h_{in} + h_{jn}}\right)$$

where $H_i = \{h_{in}\}$ and $H_j = \{h_{jn}\}$ are the histograms of word occurrences, V is the codebook size and A is the mean value of distances between all training samples. The average accuracy over all classes is reported for the performance measurement. The C parameter trades off the misclassification of training examples against simplicity of the decision surface. Low C makes the decision surface smooth, while high C aims to classify all training examples correctly. The best classification parameters C and A are selected by a grid search on all values of 2^x where x is in the range -5 to 16 and 2^y where y is in the range 3 to -15, respectively.

4. Experimental setup and results

The proposed approach is firstly evaluated based on the classification accuracy with four descriptors: TD, HOG, HOF, MBH and their combination at different codebook sizes. Secondly, more tests have been done to evaluate the influence of the trajectories parameters: matching window size, exploring window size, trajectory length, neighborhood size and cell grid structure size. The final evaluation has been carried out by

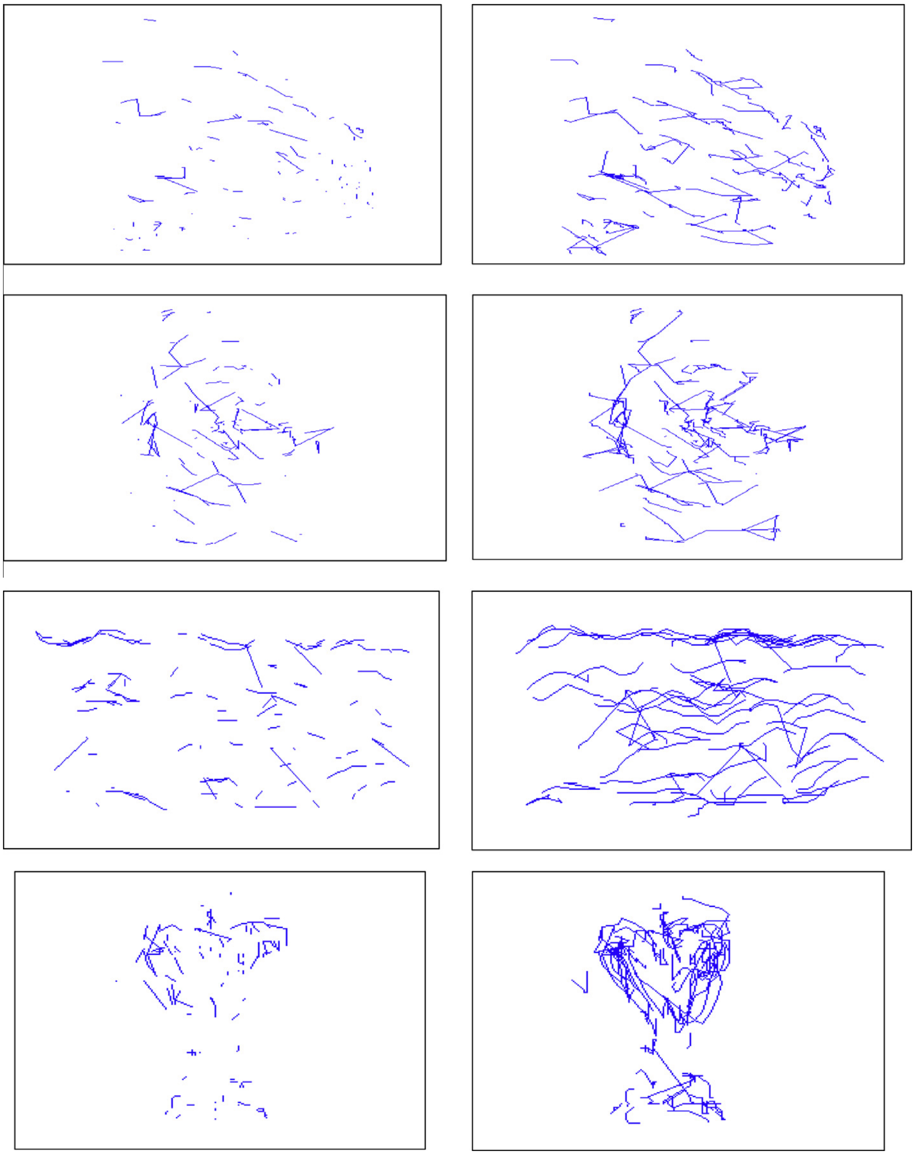

Figure 4 Examples of the generated trajectories before and after linking and exploring for four actions. First column is before linking and second column is after. The first row is for walking action, the second row is for swinging bar sport, the third row is for running action and the fourth is for jacking action.

comparing our results with the state-of-the-art results. The evaluations have been performed on three popular datasets: Weizmann, KTH and UCF sports. These datasets are chosen to test the performance of our proposed approach on the constrained and non-constrained environments with different conditions.

4.1. Datasets

KTH Dataset is the most commonly used dataset for the evaluations of action recognition approaches [20].[1] It consists of 600 videos performed by 25 different actors for six basic actions: walking, jogging, running, boxing, hand waving, and hand clapping (see Fig. 6, top). The actions were captured with a static camera under constrained environment in four different scenarios: indoors, outdoors, outdoors with scale variation, and outdoors with different clothes. Each video is sampled by 25 fps with the resolution 160 × 120 pixel. Following the original setup of the authors, a multi-class SVM classifier has been trained on sequences of 16 actors and evaluated by the sequences of the remaining 9 actors according to 5-fold cross-validation.

Weizmann Dataset [21][2] contains 93 video sequences showing nine different actors, each performing ten actions: bend, skip, jumping-jack, jump-forward-on-two-legs, jump-in-place-on-two-legs, gallop sideways, wave-one-hand, wave-two-hands, run and walk (see Fig. 6, middle). They have a spatial resolution 180 × 144 pixel and captured with a fixed camera

[1] http://www.nada.kth.se/cvap/actions/.

[2] http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html.

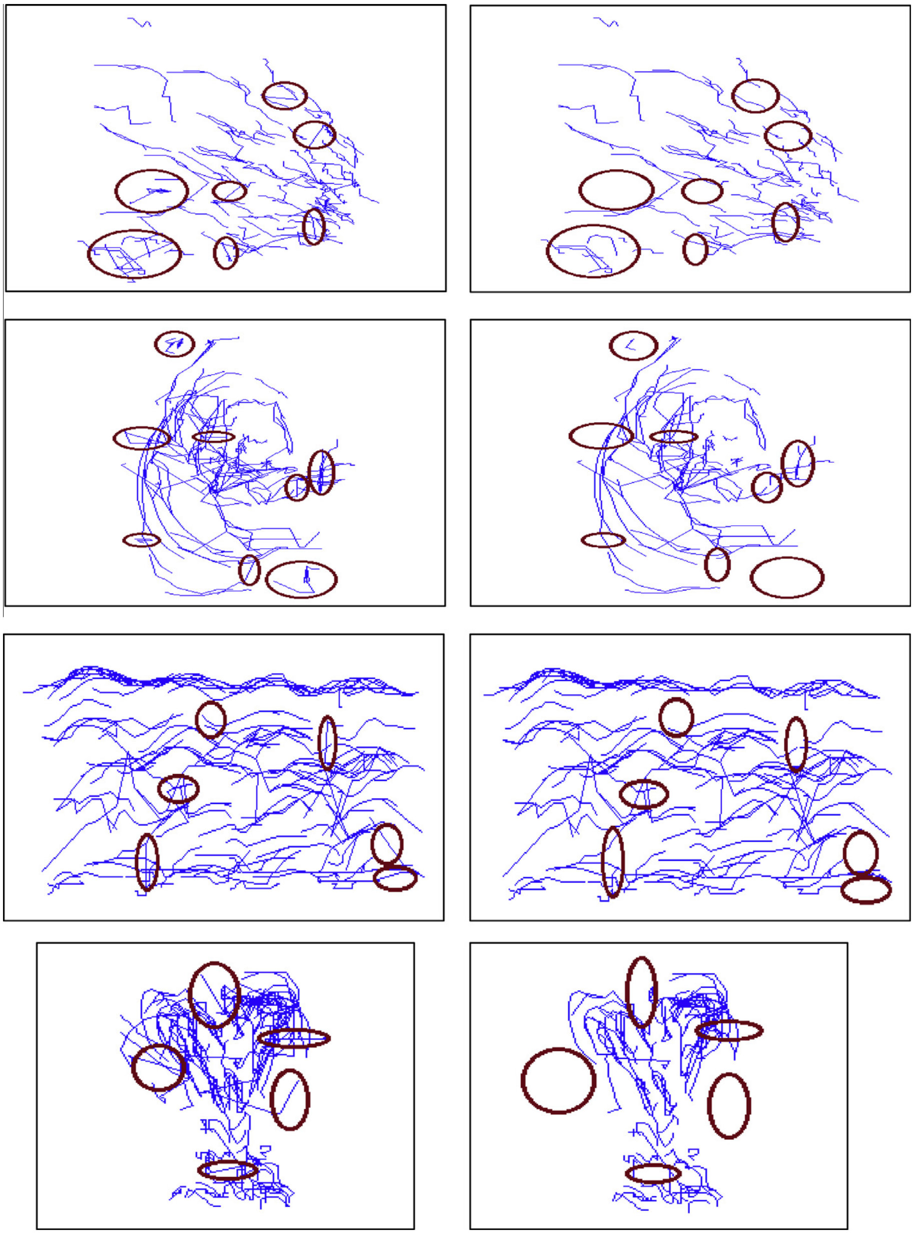

Figure 5 Examples of the generated trajectories before and after pruning for four actions. First column is before pruning and second column is after. The first row is for walking action, the second row is for swinging bar sport, the third row is for running action and the fourth is for jacking action.

25 fps under the same lighting condition. Leave-One-Out Cross-Validation (LOOCV) is used to train SVM classifier.

UCF sports dataset [22][3] contains 150 video samples for ten different types of human actions in sport broadcasting videos: diving, golf swinging, kicking, weight-lifting, horse-riding, running, skateboarding, swinging 1 (on the pommel horse and floor), swinging 2 (at the high bars) and walking (see examples in Fig. 6, bottom). The videos have different frame rate and high image resolution. We follow the setup of Wang et al. [6], the dataset is extended by adding a horizontally flipped version of each sequence to the dataset and each sequence is subsampled to its half spatial resolution to decrease the high memory requirements. For evaluation, Leave-One-Out Cross-Validation (LOOCV) is used to train SVM.

4.2. Classification accuracy

Table 1 presents the best performance of the proposed approach for different descriptors on three datasets (KTH, Weizmann and UCF sports). Also, the classification accuracies at different codebook sizes for the three datasets are shown in Fig. 8.

For *KTH dataset*: The best result of our approach 95.36% is obtained with the HOF descriptor at codebook size 3000 as shown in Fig. 7(a). In the second place, MBH and the combination report 94.9% at codebook size 3000. Finally, TD and

[3] http://www.crcv.ucf.edu/data/UCF_Sports_Action.php.

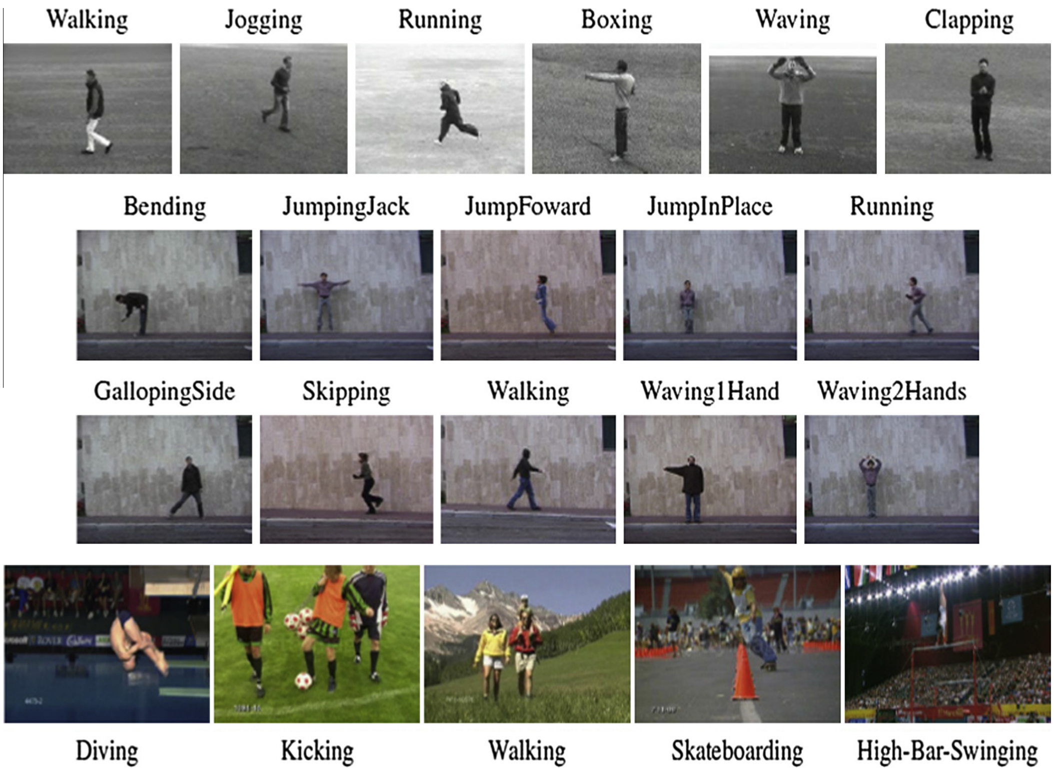

Figure 6 Sample frames from KTH (top), Weizmann (middle), and UCF Sports (bottom) human action datasets.

Table 1 The action recognition accuracy of the proposed approach using different descriptors and applied to KTH, Weizmann and UCF Sports datasets.

	KTH	Weizmann	UCF sports
TD	88.42	94.44	75
HOG	88.42	**97.77**	84.26
HOF	**95.36**	96.66	88.152
MBH	94.9	95.55	86.77
Combination	94.9	96.66	**89.97**

HOG descriptors achieve accuracy 88.42% at codebook size 4000. The confusion matrix of the best classification result on the KTH dataset is shown in Table 2. It is interesting to note that the leg-related actions (“jog”, “run”, “walk”) are more confused with each other. There is a small confusion between “wave” and “clap” by a ratio 2.77%.

Weizmann dataset: According to the results in Table 1 (second column) and Fig. 7(b), HOG descriptor achieves the best result of 97.77% among the tested descriptors at codebook size 2000. Comes in the second place is HOF descriptor and the combination by 96.66% at codebook sizes 1000 and 3000, respectively. MBH descriptor obtains 95.55% at codebook size 1000. In the last place, TD descriptor achieved 94.44% at codebook size 3000. It can be observed that Weizmann dataset needs small visual words for video representation than KTH dataset. It may be explained by the little variations in this dataset.

The confusion matrix of recognition results for HOG descriptor on the Weizmann dataset is presented in Table 3. The confusion occurs between the actions “jump” and “skip” by a ratio 11.11% when using the proposed trajectory-based approach with HOG descriptor.

Our approach is also tested for more complex actions and background video variations on *UCF Sports dataset*. According to the presented results in Table 1 (third column) and Fig. 7(c), the combination of the computed descriptors is superior descriptor for the UCF dataset which obtains 89.97% accuracy at codebook size 4000. HOF descriptor obtains 88.15% and 86.77% for MBH descriptor at codebook size 3000. TD and HOG descriptors achieve accuracy 74.82% and 84.26% at codebook sizes 1000 and 4000, respectively. The high performs of TD descriptor at small codebook size 1000 can be explained by that it described the actions with short dimension vectors (18 dimension) which yields a lot of similarity between the different actions representation.

The confusion matrix of the best classification result on the UCF sports dataset is shown in Table 4. It can be observed that the recognition accuracies of actions “dive,” “kick,” “lift,” “walk,” “swing1” and “swing2” reach 100%. Another observation is that high confusion is occurred between the actions “ride,” “run” and “skate”. It should be noted that the videos of these actions have a cluttered backgrounds and some of them were captured with a moving camera. Therefore, this confusion is reduced by using the combination.

4.3. Evaluation of trajectory parameters

In this section, the influence of the different parameter settings for the proposed trajectory is evaluated. KTH dataset is selected to report the results because of its middle level of

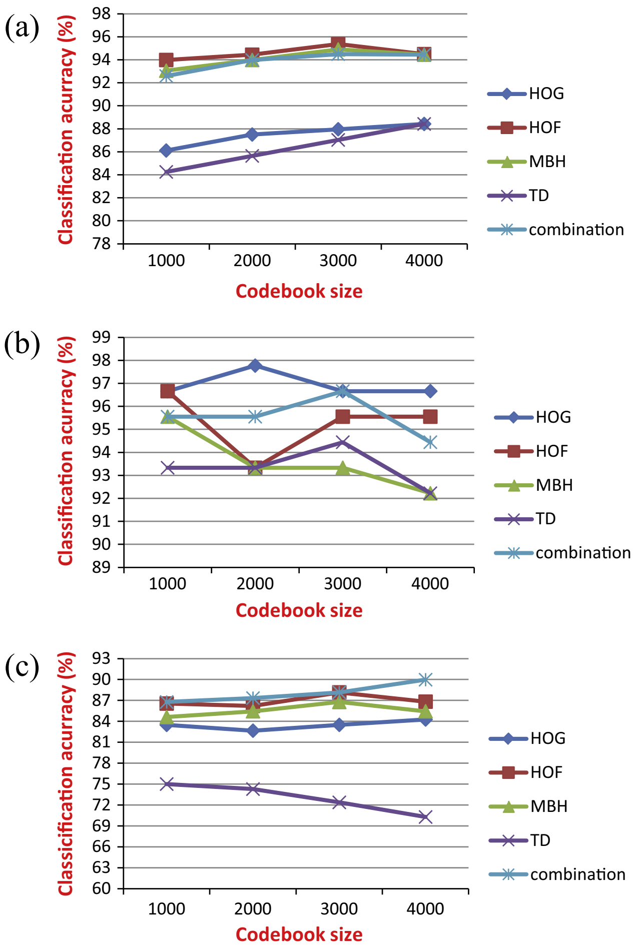

Figure 7 Classification accuracies for different descriptors at different codebook sizes for (a) KTH dataset, (b) Weizmann dataset, and (c) UCF sports dataset.

Table 2 Confusion matrix for the best classification result of HOF descriptor on KTH dataset.

(%)	Box	Clap	Wave	Jog	Run	Walk
Box	**100**	0	0	0	0	0
Clap	0	**100**	0	0	0	0
Wave	0	2.77	**97.22**	0	0	0
Jog	0	0	0	**94.44**	5.55	0
Run	0	0	0	16.69	**83.33**	0
Walk	0	0	0	2.77	0	**97.22**

difficulty. We study the impact of matching window size, exploring window size, trajectory length, neighborhood size, cell grid structure. All evaluations are carried out for one parameter at a time and the other parameters are fixed to the default values, i.e., matching window size $M = 32$, trajectory length $L = 10$, exploring window size $W = 10$, neighborhood size $N = 32$ and cell grid structure $n_{x,y} = 3$, $n_t = 2$. Based on the experimental results presented in Fig. 8, the codebook size is fixed to 3000 and HOF descriptor is used for trajectory description. The trajectory parameters are evaluated based on the classification accuracy.

The evaluation results of the matching window size (M) in Fig. 8(a) show that the classification accuracy is increased by increasing the window size till reaching the size 32 × 32. A decrement in classification accuracy is observed after this size. This occurs because using large window size causes a high ratio of matching errors and generates a large number of bad trajectories. The same reasons can explain the evaluation results of the exploring window size (W) that presented in Fig. 8(b).

The evaluation results of the trajectory length (L) presented in Fig. 8(c) show that the classification accuracy is decreased by increasing the length beyond $L = 10$. It should be noted that most of the STIPs which were extracted with different detectors represent regions of short range motion and varied rapidly with time. Therefore, depending on the used detector and trajectory generation approach, the ratio of wrong paths is increased beyond the predefined trajectory length. For the same reasons, the suitable neighborhood size used for trajectory description is 32 × 32 with grid cell structure 3 × 3 × 2 as shown in Fig. 8(d) and (e).

4.4. State-of-the-art comparison

KTH dataset: Among the results of the previously published trajectory-based local representation approaches [13,16,17,19], our method achieves the highest recognition accuracy of 95.36% as shown in Table 5 (first column). Messing et al. [13] reported 74% using the velocity histories description for their trajectories. Sun et al. [16] obtained 86.8% using the trajectories statistics description. When we compare with the TD description of our trajectories, an improvement of 14.42% is obtained above Messing et al. [13] and 1.62% above

Table 3 Confusion matrix for the best classification result of HOG descriptor on Weizmann dataset.

(%)	Bend	Jack	Jump	P-Jump	Run	Side	Skip	Walk	Wave1	Wave2
Bend	**100**	0	0	0	0	0	0	0	0	0
Jack	0	**100**	0	0	0	0	0	0	0	0
Jump	0	0	**100**	0	0	0	0	0	0	0
P-Jump	0	0	0	**100**	0	0	0	0	0	0
Run	0	0	0	0	**88.88**	0	11.11	0	0	0
Side	0	0	0	0	0	**100**	0	0	0	0
Skip	0	0	11.11	0	0	0	**88.88**	0	0	0
Walk	0	0	0	0	0	0	0	**100**	0	0
Wave1	0	0	0	0	0	0	0	0	**100**	0
Wave2	0	0	0	0	0	0	0	0	0	**100**

Table 4 Confusion matrix for the best classification result of the descriptors combination on UCF sports dataset.

	DIVE	GOLF	KICK	LIFT	RIDE	RUN	SKATE	SWING1	SWING2	WALK
DIVE	**100**	0	0	0	0	0	0	0	0	0
GOLF	0	**88.88**	0	0	0	0	0	0	0	11.11
KICK	0	0	**100**	0	0	0	0	0	0	0
LIFT	0	0	0	**100**	0	0	0	0	0	0
RIDE	0	0	0	0	**75**	8.33	16.66	0	0	0
RUN	0	0	0	0	15.4	**69.23**	0	0	0	7.7
SKATE	0	0	0	0	0	16.66	**66.66**	0	0	16.66
SWING1	0	0	0	0	0	0	0	**100**	0	0
SWING2	0	0	0	0	0	0	0	0	**100**	0
WALK	0	0	0	0	0	0	0	0	0	**100**

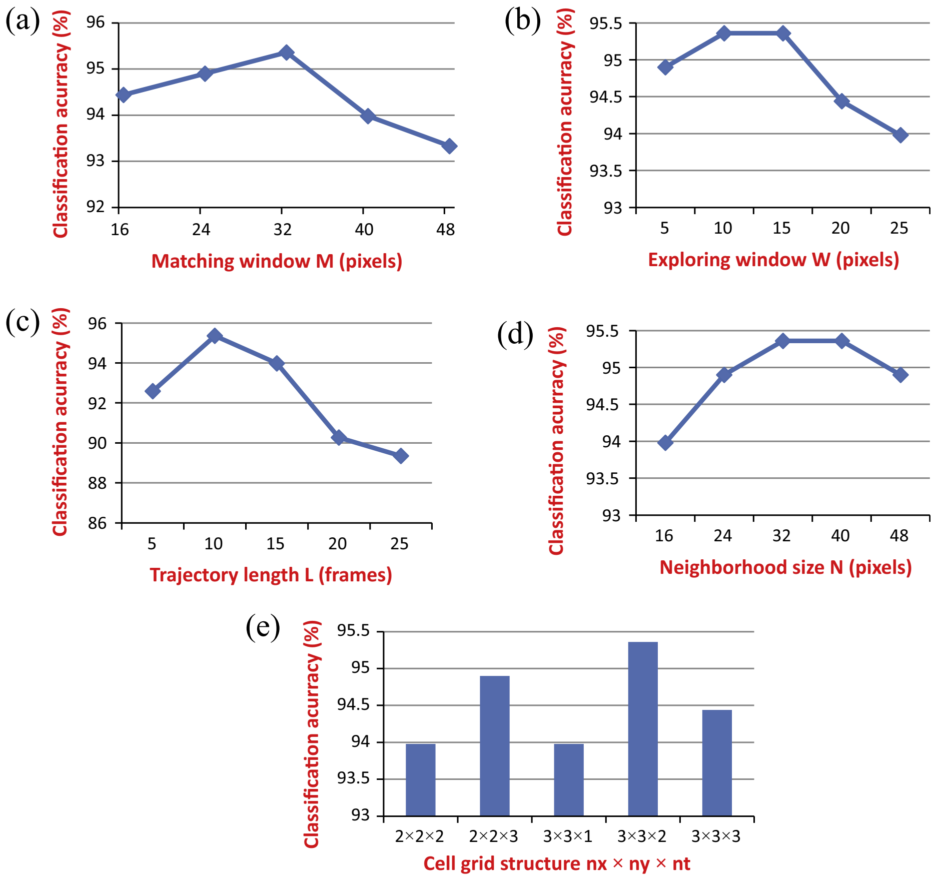

Figure 8 Evaluation of trajectory parameters (a) matching window size, (b) exploring window size, (c) trajectory length, (d) neighborhood size, (e) grid structure.

Sun et al. [16]. Moreover, Raptis and Soatto [17] reported 94.8% using the HOG/HOF tracklets description.

Comparing to dense trajectories proposed by Wang et al. [19] (see Table 6, second column), our trajectories achieve 0.1% lower performance in case of MBH descriptor. An improvement of 1.4% is achieved by HOG descriptor and 2% by HOF descriptor. TD descriptor of the proposed trajectories reports 1.4% lower than the accuracy obtained by dense trajectories. The descriptors combination of our trajectories results in 0.7% improvement over the dense trajectories. We have to note that the length of dense trajectories is fixed to 15 frames in [19] and the codebook size = 4000. It is longer than our trajectories and that explains the lower results of TD descriptor. Also, due to the use of Cuboid detector (denser than dense sampling) and SIFT matching, the number of generated trajectories is lower than dense trajectories and more curved.

The proposed method obtains large improvement over the traditional KLT and SIFT-trajectories computed in Wang et al. [19]. Additional comparison of the proposed method with the state-of-the-art in the literature is presented in Table 7. On KTH dataset, we obtain 95.63% with the HOF descriptor

Table 5 Comparison with previous trajectory-based approaches of local action representation for KTH and UCF sports datasets.

KTH		UCF sports	
Messing et al. [13]	74.00%		
Sun et al. [16]	86.80%		
Raptis and Soatto [17]	94.80%	Bregonzio et al. [18]	86.90%
Wang et al. [19]	95.00%	Wang et al. [19]	88.00%
Our method	**95.36%**	Our method	**89.97%**

which is comparable to the state of the art, i.e., 95.7% [27]. Note that the several works of human action recognition in the literature were evaluated using different methods and with different conditions.

Comparing our approach on *Weizmann dataset* with the state-of-the-art results shows that we obtain a comparable result of 97.77% using HOG descriptor with [28] as shown in Table 7 (second column).

For UCF sports dataset, as shown in Table 5 (second column) Bregonzio et al. [18] reported recognition accuracy 86.90% by combining the KLT and SIFT tracks with the detected spatio-temporal points. When comparing the proposed approach performance with that of dense trajectories [19] (see Table 6, third column), we observed that a large improvement is achieved over dense trajectories. The largest improvement is 11.35% gained for HOF descriptor. Around 3% is improved for MBH descriptor and 2% for the descriptors combination. The HOG descriptor of the proposed trajectories gives similar results as dense trajectories. Again TD descriptor reports 0.4% lower than dense trajectories. The achieved improvement of the motion descriptors indicates that our trajectories capture good motion information from videos under uncontrolled environments. As shown in Table 7 (third column), the proposed trajectory-based approach obtained better results than the current state-of-the-art approaches.

5. Conclusions

In this paper, a trajectory-based approach for local action representation has been proposed. The objective is to explore the temporal relationships between the spatio-temporal interest locations in video sequences which characterized human actions. Our method differs from previous trajectory-based approaches in the extracted key points and tracking methodology. The spatio-temporal interest points are selected to reduce the redundancy and noise level. On the other hand, the tracking methodology is designed to extract reliable and robust trajectories which are able to describe the motion information under occlusions, camera motion, viewpoint changes, and scales variations. The experiments are carried out on three popular datasets (KTH, Weizmann and UCF sports) under the framework of the Bag-of-Words model and non-linear Support Vector Machine.

The experimental results demonstrated that an improvement of 2% is achieved by the proposed approach when evaluated on the UCF sports dataset over the latest trajectory-based approaches (dense trajectories) proposed by Wang et al. [19]. Also, an improvement of 0.4% is achieved when evaluated on the KTH dataset over the dense trajectories approach.

Comparing with the traditional local-based action representation approaches, the proposed approach reports a comparable results for KTH and Weizmann datasets and improves the current state-of-the-art results for UCF sports dataset.

Table 6 Comparison with the dense trajectories approach proposed by Wang et al. [19] for KTH and UCF sports datasets.

Description method	KTH		UCF sports	
	Dense trajectories (%)	Our trajectories (%)	Dense trajectories (%)	Our trajectories (%)
TD	**89.8**	88.4	**75.4**	75
HOG	87.0	**88.4**	84.3	84.26
HOF	93.3	**95.36**	76.8	**88.152**
MBH	95.0	94.9	84.2	**86.77**
Combination	94.2	**94.9**	88.0	**89.97**

Table 7 Comparison of the proposed approach with the state-of-the-art for three datasets.

KTH	Weizmann	UCF sports			
Dollar et al. [4]	81.20%	Bregonzio et al. [12]	96.66%		
Laptev et al. [2]	91.80%	Seo and Nilanfar [28]	97.50%		
Gilbert et al. [10]	94.50%	Wang et al. [30]	96.70%		
Kovashka and Grauman [29]	94.53%			Wang et al. [6]	85.60%
Gilbert et al. [27]	**95.70%**			Kovashka and Grauman [29]	87.27%
Our method	95.36%	Our method	**97.77%**	Our method	**89.97%**

References

[1] Weinland D, Ronfard R, Boyer E. A survey of vision-based methods for action representation, segmentation and recognition. Comput. Vis. Image Underst. 2011;115(2):224–41.

[2] Laptev I, Marszalek M, Schmid C, Rozenfeld B. Learning realistic human actions from movies. IEEE Conf Comput Vis Pattern Recognit; 2008.

[3] Laptev I. On space-time interest points. Int. J. Comput. Vision 2005;64(2–3):107–23.

[4] Dollar P, Rabaud V, Cottrell G, Belongie S. Behavior recognition via sparse spatio-temporal features. IEEE Int Work Vis Surveill Perform Eval Track Surveill 2005.

[5] Willems G, Tuytelaars T, Van Gool L. An efficient dense and scale-invariant spatio-temporal interest point detector. Comput Vision – ECCV 2008.

[6] Wang H, Ullah MM, Klaser A, Laptev I, Schmid C. Evaluation of local spatio-temporal features for action recognition. In: Proceedings br mach vis conf; 2009. p. 11.

[7] Scovanner P, Ali S, Shah M. A 3-dimensional sift descriptor and its application to action recognition. In: 15th international conference on Multimedia; 2007. pp. 357–60.

[8] Kl A, Schmid C, Grenoble I. A spatio-temporal descriptor based on 3D-gradients. In: British Machine Vision Conference; 2008.

[9] Liu J, Shah M. Learning human actions via information maximization. In: 26th IEEE conference on computer vision and pattern recognition. CVPR; 2008.

[10] Gilbert A, Illingworth J, Bowden R. Fast realistic multi-action recognition using mined dense spatio-temporal features. In: Proceedings of the IEEE international conference on computer vision; 2009. pp. 925–31.

[11] Zhang Z, Hu Y, Chan S, Chia L-T. Motion context: a new representation for human action recognition. In: Computer vision – ECCV 2008, vol. 5305. Springer, Berlin/Heidelberg; 2008. p. 817–29.

[12] Bregonzio M, Gong S, Xiang T. Recognising action as clouds of space-time interest points. In: IEEE computer society conference on computer vision and pattern recognition workshops. CVPR Workshops; 2009. p. 1948–55.

[13] Messing R, Pal C, Kautz H. Activity recognition using the velocity histories of tracked keypoints. In: Proceedings of the IEEE international conference on computer vision; 2009. pp. 104–11.

[14] Matikainen P, Hebert M, Sukthankar R. Trajectons: Action recognition through the motion analysis of tracked features. In: IEEE 12th international conference on computer vision workshops. ICCV Workshops; 2009. pp. 514–21.

[15] Sun J, Wu X, Yan S, Cheong LF, Chua TS, Li J. Hierarchical spatio-temporal context modeling for action recognition. In: IEEE Computer society conference on computer vision and pattern recognition workshops. CVPR Workshops; 2009. pp. 2004–11.

[16] Sun J, Mu Y, Yan S, Cheong LF. Activity recognition using dense long-duration trajectories. In: Multimed. Expo (ICME), IEEE int. conf.; 2010.

[17] Raptis M, Soatto S. Tracklet descriptors for action modeling and video analysis. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), vol. 6311 LNCS(PART 1); 2010. pp. 577–90.

[18] Bregonzio M, Li J, Gong S, Xiang T. Discriminative topics modelling for action feature selection and recognition. In: Proceedings br mach vis conf; 2010. pp. 8.1–8.11.

[19] Wang H, Kläser A, Schmid C, Liu CL. Dense trajectories and motion boundary descriptors for action recognition. Int J Comput Vis 2013;103(1):60–79.

[20] Schuldt C, Laptev I, Caputo B. Recognizing human actions: a local SVM approach. In: Proc 17th Int Conf Pattern Recognition, vol. 3; 2004.

[21] Gorelick L, Blank M, Shechtman E, Irani M, Basri R. Actions as space-time shapes. IEEE Trans Pattern Anal Mach Intell 2007;29:2247–53.

[22] Rodriguez MD, Ahmed J, Shah M. Action MACH: A spatio-temporal maximum average correlation height filter for action recognition. In: 26th IEEE conference on computer vision and pattern recognition. CVPR; 2008.

[23] Lucas BD, Kanade T. An iterative image registration technique with an application to stereo vision. Imaging, vol. 130(x); 1981. pp. 674–9.

[24] Dalal N, Triggs B, Schmid C. Human detection using oriented histograms of flow and appearance. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), vol. 3952 LNCS; 2006. pp. 428–41.

[25] Shi JSJ, Tomasi C. Good features to track. In: Comput. vis. pattern recognition; 1994.

[26] Lowe DG. Object recognition from local scale-invariant features. Proc Seventh IEEE Int Conf Comput Vis 1999;2.

[27] Gilbert A, Illingworth J, Bowden R. Action recognition using mined hierarchical compound features. IEEE Trans Pattern Anal Mach Intell 2011;33(5):883–97.

[28] Seo HJ, Milanfar P. Action recognition from one example. IEEE Trans Pattern Anal Mach Intell 2011;33(5):867–82.

[29] Kovashka A, Grauman K. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition; 2010. pp. 2046–53.

[30] Wang H, Yuan C, Hu W, Sun C. Supervised class-specific dictionary learning for sparse modeling in action recognition. Pattern Recognit 2012;45(11):3902–11.