Weng and Guan *EURASIP Journal on Image and Video Processing* (2018) 2018:8
DOI 10.1186/s13640-018-0250-5

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Action recognition using length-variable edge trajectory and spatio-temporal motion skeleton descriptor

Zhengkui Weng[1] and Yepeng Guan[1,2*]

Abstract

Representing the features of different types of human action in unconstrained videos is a challenging task due to camera motion, cluttered background, and occlusions. This paper aims to obtain effective and compact action representation with length-variable edge trajectory (LV-ET) and spatio-temporal motion skeleton (STMS). First, in order to better describe the long-term motion information for action representation, a novel edge-based trajectory extracting strategy is introduced by tracking edge points from motion without limiting the length of trajectory; the end of the tracking is depending not only on the optical flow field but also on the optical flow vector position in the next frame. So, we only make use of a compact subset of action-related edge points in one frame to generate length-variable edge trajectories. Second, we observe that different types of action have their specific trajectory. A new trajectory descriptor named spatio-temporal motion skeleton is introduced; first, the LV-ET is encoded using both orientation and magnitude features and then the STMS is computed by motion clustering. Comparative experimental results with three unconstrained human action datasets demonstrate the effectiveness of our method.

Keywords: Human action recognition, Length-variable edge trajectory, Motion clustering, Spatio-temporal motion skeleton

1 Introduction

Human action recognition (HAR) is an active research topic in intelligent video analysis, gained extensive attention in academic and engineering communities [1, 2], and widely used in the fields of human-computer interaction, video surveillance, motion analysis, virtual reality, etc. [3–7]. Usually, the realization of HAR includes two steps: the first is feature extraction based on video information; the second is the classification according to feature vectors. However, due to the presence of background clutter, partial occlusion, varying viewpoints, and camera movement, it is still a challenging task to obtain discriminative action representations from realistic videos.

Recently, the trajectory-based methods were proposed and utilized in various human action recognition approaches owing to the promising results of histogram-based descriptors [8–12]. Different with the method of extracting the local features directly, the trajectory-based method is extracting spatio-temporal trajectory by matching the feature points between adjacent frames to represent human actions [13–15]. Due to the better description of motion changing and long-term target dynamic information, these methods out-performed than the local representations significantly. Therefore, this paper focuses on the trajectory-based method and will especially emphasize the following two aspects, namely, the trajectory extraction and trajectory description.

In general, previous studies ignore the motion feature of the tracking points and the differences between various types of actions. To address this issue, we propose the length-variable edge trajectory extracting method. Moreover, inspired by [16], a model of the human body is composed of different key skeletons (i.e., limbs, trunk, skull, etc.) so is the video. In the video, we regard the various edge trajectories with analogous spatio-temporal and motion features as a set of

* Correspondence: ypguan@shu.edu.cn
[1]School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China
[2]Key Laboratory of Advanced Display and System Application, Ministry of Education, Shanghai 200072, China

skeletons. An overview of the proposed pipeline is shown in Fig. 1. First, we compute the edges and optical flow of each frame of an input video. Second, we generate the length-variable edge trajectories (LV-ET) according to the detected edge points adaptively. The tracking is terminated when a feature point moves to a region with no edge points. Third, we extract spatio-temporal motion skeleton (STMS) descriptor by regarding the generated trajectories as different skeletons, and these skeletons are clustered under a novel motion encoding method. In addition, the descriptors (that is, HOG, HOF, and MBH), which proved discriminative, were also extracted from proposed LV-ET. After the extraction of local descriptors, we adopt fisher vector (FV) [17] to separately encode these descriptors. Before encoding, the dimension of each local descriptor was reduced by principal component analysis (PCA). Finally, the multi-kernel learning-based support vector machine (MKL-SVM) [18] is employed to classify the human actions. As such, the main contribution of this paper is as follows:

- In this paper, we propose to sample edge points in each frame in order to adaptively select the trajectories associated with moving targets. In addition, we do not fix the length of trajectory, when the tracking is terminated depending on whether there is still the edge point in the next frame. So, a compact set of trajectories (LV-ET) can be obtained exactly to better represent the motion information of each action.
- By introducing a spatio-temporal trajectory encoding method, where the videos are represented as a set of distinctive trajectory skeletons with latent motion features, a novel trajectory descriptor named STMS is designed by making use of spectral clustering with motion information, where videos are represented as a set of skeletons.

The remainder of this paper is organized as follows: we give an overview of the related work in Section 2. We describe the details of the LV-ET in Section 3. We explain the process for extracting the STMS descriptor in and discuss the experimental results in Section 4. Finally, the conclusion is drawn in Section 5.

2 Related work

Recently, trajectory-based approaches have gained significant popularity in video interpretation. As illustrated in Fig. 2, Sun et al. [19] tracked spatio-temporal context information between adjacent frames by using SIFT matching, whereas Matikainen et al. [20] extracted trajectory by tracking interesting points with Kanade-Lucas-Tomasi (KLT) feature tracker; interestingly, Bregonzio et al. [21] combined the SIFT matching and KLT interesting points tracker so that the trajectory density was greatly improved. Although the motion changing and long-term target dynamic information can be captured by these methods, it still suffers a problem. These were kinds of sparse tracking method [19–21], which led to sparse trajectory, since the sparse trajectory was not sufficient enough to represent human actions [22]. In consideration of the shortcoming, Wang et al. [23, 24] used dense optical flow to sample dense feature points on a regular grid, and the points located at spatially homogeneous regions were removed to generate a spatio-temporal dense trajectory (DT) and improved dense trajectory (iDT). In order to represent the extracted trajectories, iDT used four descriptors: histogram of oriented gradients (HOG) [25], histogram of optical flow (HOF) [26], motion boundary histograms (MBH) [23], and trajectory shapes (TS) [23], which had

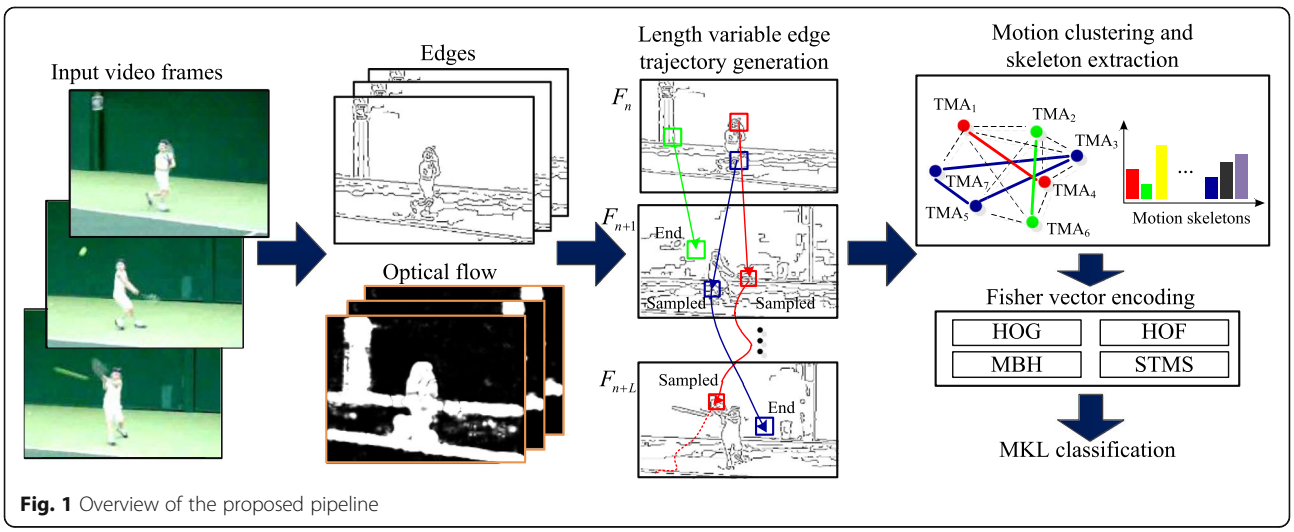

Fig. 1 Overview of the proposed pipeline

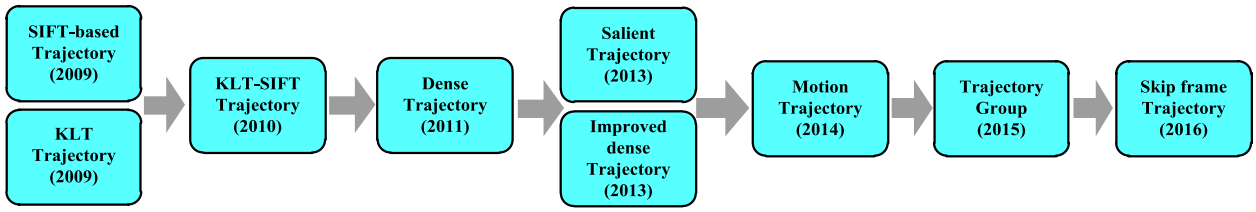

Fig. 2 Illustration of the trajectory-based approaches in chronological order

been proved to be very discriminative in representing actions; Gaidon et al. [27] represented a video as a hierarchy of trajectories. The hierarchy was first computed using a divisive clustering algorithm and then represented a BOF-liked structure where each node was modeled by a bag-of-feature over MBH descriptors. In [28], the video was structured using bag of trajectory groups with latent class variables and employing multiple instance learning (MIL)-based support vector machine (SVM) to classification. More recently, Vig et al. [29] revealed that the HAR performance can be maintained with as little as 30–40% of descriptors picked at random within dense trajectory-based approach [23]. It implies that although plenty of information can be obtained by dense descriptor, most of them are redundant even irrelevant. Therefore, Yi et al. [30] aimed to deal with the problem of HAR with salient trajectories, then the salient trajectories were encoded by a hierarchical representation-based method. Iveel et al. [29] extracted motion trajectories after video segment and build a bag-of-features (BOF) model based on four different types of descriptors (i.e., HOG, HOF, MBH, and TS). Unlike [30, 31] use video saliency to generate trajectory, Seo et al. [16] introduced a trajectory rejection technology to avoid generating redundant trajectories and skip the frames that do not have much movement information, so the complexity of the algorithm is reduced.

Previous work shows that the most informative trajectories were extracted from the region of interest (ROI); motion and shape are two important information sources from human action videos. It is worthy noted that we also cannot guarantee that the points inside ROI are more informative and representative. Besides, we find that different types of action have its own movement rhythm, in other words, the evolution of different actions is unequal (e.g., "run" is a high-speed and continuous action whereas the action of "hug" is relatively slow and discontinuous between two persons). Therefore, the use of fixed length trajectories [30, 23, 24, 31] is not discriminative enough to represent the various types of human actions. In addition, we observe that the spatial-temporal features of trajectories are similar in the same kind of action, whereas those between different actions are dissimilar. So, the HAR performance may be improved if we can better consider the phenomenon aforementioned.

3 Methods

3.1 Proposed length-variable edge trajectory

In this section, we introduce the major components of the proposed LV-ET extraction, including edge point sampling and tracking, trajectory generation, and its pruning strategy.

3.1.1 Edge point sampling

The key to edge trajectory extraction is to select the tracking points exactly. In general, spatio-temporal interest point sampling [9, 10] and dense sampling [23, 24] have been successfully applied to various occasions. However, such approach often involves some irrelevant interest points like the background points with high complexity, which seriously affect the final recognition accuracy and reduce the efficiency of the algorithm. Unlike the sampling approach mentioned above, we utilize Canny detector to detect edge points. Moreover, in order to better obtain human action motion information and edge trajectories, we also leverage optical flow to track edge points across video frames to extract trajectories.

Given a video sequence $F = (F_1,F_2,F_3,\ldots,F_t)$, where F_t is the frame at time t, we first compute the optical flow frame to frame using Farnebäck dense optical flow similar to [32], for each frame F_t, its dense optical flow field $\omega_t = (u_t|_{(x_t,y_t)}, v_t|_{(x_t,y_t)})$ is computed with regarding to the next frame F_{t+1}, where $u_t|_{(x_t,y_t)}$ and $v_t|_{(x_t,y_t)}$ are respectively the horizontal and vertical components of the optical flow of F_t. More specifically, we smooth both components by using a median filtering:

$$\overline{u_t}|_{(x_t,y_t)} = u_t|_{(x_t,y_t)} \times M_{3\times3} \quad (1)$$

$$\overline{v_t}|_{(x_t,y_t)} = v_t|_{(x_t,y_t)} \times M_{3\times3} \quad (2)$$

Where M is a median filtering kernel with 3×3 pixels. We also perform the same mean filtering for too long optical flow displacements in order to further reduce the errors caused by the drift in the tracking process. In

addition, we remove the displacement lower than 0.3 empirically to mitigate the impact of noise. Since background points have less contribution to HAR, whereas a large amount of action information tends to be concentrated in the regions with gradients changing acutely. So, we detect edge points in each frame using Canny detector to distinguish the edge and none edge regions. The result $E = (E_1,E_2,E_3,\ldots,E_t)$ represents the edge points according to each frame, where E_t is a binary matrix, $E_t(x_t,y_t) = 1$ represents the point (x_t,y_t) in F_t as an edge point. As illustrated in Fig. 3, where in Fig. 3a, b, the points are sparsely and densely sampled as the same as in [20, 23], we can observe that no matter the points are in foreground or background, they are all sampled by the tracker, so the background noises may be brought into tracking process. Through the edge-based sampling strategy as shown in Fig. 3c, the noises from background areas are greatly reduced. But we can also find that the sampled points are too sparse, which leads to tracking failure. In order to make the tracking process more effective and informative, after edge point detection, the edge images per frame E_t are dilated by a 3×3 square-shaped structuring element. So, the succeeding candidate points may have more chance to be found as valid. Therefore, the tracking loss is alleviated and the completeness and continuity of the edge trajectory are guaranteed.

3.1.2 Tracking and edge trajectory generation

In the following, the LV-ET in the video is extracted through a tracking process. To begin with, for each frame F_t, the edge points E_t are sampled as the start point of an edge trajectory, given an edge point (x_t,y_t), the trajectory's succeeding tracking point (x_{t+1},y_{t+1}) in the net frame F_{t+1} is simply decided by the horizontal and vertical components of the optical flow $(\overline{u_t}|_{(x_t,y_t)}, \overline{v_t}|_{(x_t,y_t)})$respectively:

$$x_{t+1} = x_t + \overline{u_t}|_{(x_t,y_t)} \tag{3}$$

$$y_{t+1} = y_t + \overline{v_t}|_{(x_t,y_t)} \tag{4}$$

By observing the computational process, it is not difficult to find that if we compute the trajectory and only use (3) and (4), the succeeding tracking points may have a high-risk drifting to none informative regions, no matter whether the initial edge points are stipulated or not. In consideration of this, we utilize a novel tracking strategy that when computing the succeeding trajectory point, we prejudge whether the succeeding candidate point is an edge point. In other words, we compute every frame's edge information E_t and use it to determine whether the round position (x_{t+1},y_{t+1}) is an edge point in F_{t+1}. If so, we regard it as the valid sampled point in F_{t+1}. Otherwise, we consider the succeeding sampled point (x_{t+1},y_{t+1}) is not a valid trajectory point and terminated the tracking process at (x_t,y_t). In the next step, for each edge point in F_{t+1}, if it is not a succeeding sampled point of F_t, we use it as the initial point of a new trajectory. This frame-by-frame tracking process is iterated until the last sampling points are found in the last frame.

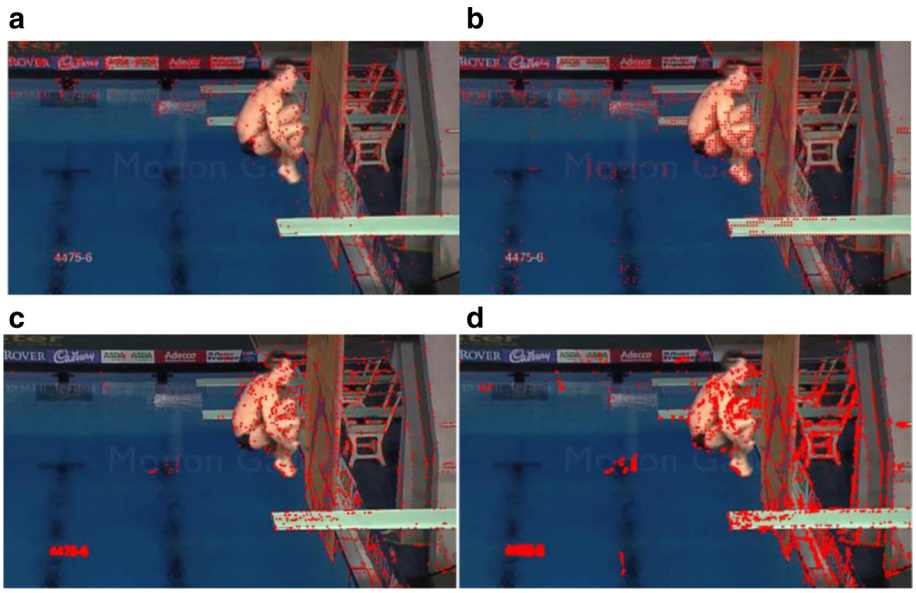

Fig. 3 Comparison between different sampling strategies: **a** interest points sampling, **b** dense sampling, **c** edge-based sampling, and **d** edge-based sampling after dilating

Based on the process aforementioned, we can obtain all the trajectories frame-by-frame from begin to end. Namely, the subsequent sampling point positions are concatenated orderly to form the edge trajectory: $Tra_t = (P_t, P_{t+1}, P_{t+2}, \ldots)$. We compute the trajectory length according to the number of the frames it goes through. It is worth noting that we do not stipulate the length of one trajectory, whether the tracking is terminated depending on the round position of succeeding sampling point that is an edge point or not. We think there are several reasons to do so. First, different types of action have different motion information, or we may call it motion speed. Some actions are changing at a constant velocity whereas some involve drastic changes. For instance, "run" is a high-speed, continuous, and uniform action whereas the action of "hug" is a relatively slow and discontinuous action between two persons. Figure 4 illustrates the differences of trajectory length histogram between "run" and "hug," where the length of trajectories are normalized by dividing every video clips' frame numbers. In Fig. 4a, long trajectories make up a large part of all trajectories in the histogram of "run," but the histogram of trajectories length in Fig. 4b does not appear this tendency; the trajectories' length in action "hug" is relatively shorter than that in "run." This phenomenon exactly illustrates the discrimination of proposed LV-ET.

3.1.3 Trajectory pruning strategy

To sum up, the fixed length trajectory cannot fully represent the diversity of different actions and may also bring down the discrimination of the trajectory-based feature descriptors. Moreover, considering the continuity and completeness of a trajectory is crucial to represent some long-term non-periodic actions. Although the optical flow vector has been smoothed to eliminate the vector displacement drifting and the edge-based sampling strategy reduces the number of the trajectories while retaining the discrimination information, we have to admit that some problems still exist during this process. First, edge-based sampling strategy may lead to tapping in local minima and also may lead to a very long trajectory with point drifting. Meanwhile, the static edge points may overlap and generate trajectories which are too short and contain little useful information. Considering the displacement of these static point trajectories is slight, we remove these trajectories by introducing coefficient of variation (var) as follows:

$$\mathrm{var}(Tra_t) = \frac{1}{2L}\left[\sum_{i=t}^{L}(x_t - \bar{x})^2 + (y_t - \bar{y})^2\right] \tag{5}$$

where L is the length of a trajectory, $\bar{x}$ and $\bar{y}$, respectively, represent the mean value of horizontal and vertical coordinates of a trajectory. For a trajectory

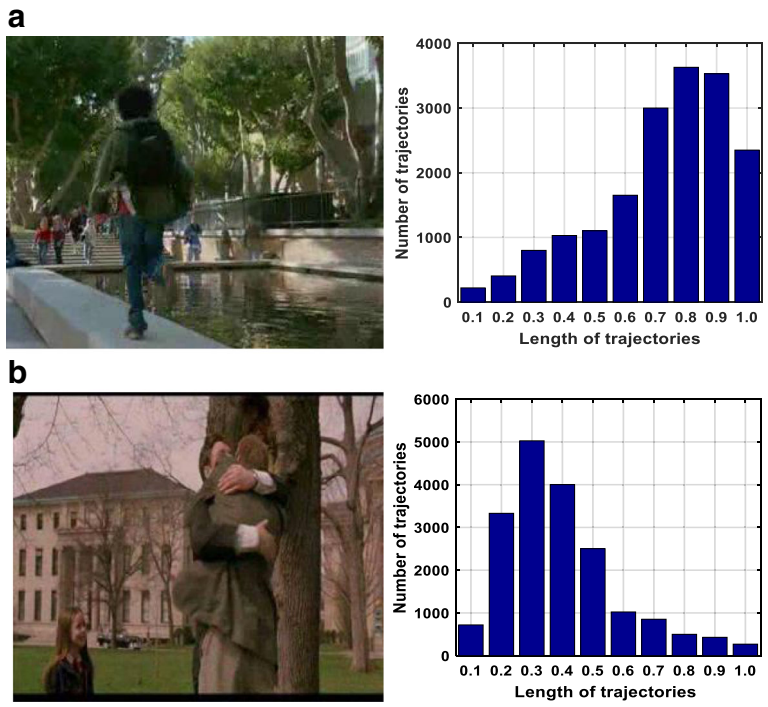

Fig. 4 Comparison of LV-ET length histogram between two sample actions after normalization: **a** run and **b** hug

Tra_t, the coordinate series with var smaller than 10 is pruned. Second, camera motion is a very common phenomenon in unconstrained datasets. Both action-related points and background points are moving together in such videos, like dive and football match. It is hard to distinguish motion trajectories and background trajectories. In this paper, an effective technology is utilized to solve the problem. We observe that the trajectories generated by camera motion are always uniform than those of motion trajectories, in other words, there is no obvious changing of direction for the trajectories caused by camera motion. So, we empirically prune the trajectories with mean curvature (mc) smaller than 30°. The mc for every trajectory is computed according to formula (6).

$$\mathrm{mc}(Tra_t) = \frac{1}{L-1}\sum_{t=1}^{L-1}\left|\arctan\left(\frac{y_t - y_{t+1}}{x_t - x_{t+1}}\right)\right| \times (x_t \neq x_{t+1}) \tag{6}$$

Considering the rigor of the process, when the adjacent horizontal coordinate x_t is equal to x_{t+1}, we deem the curvature of this part is 0°. In addition, we remove the $L < 5$ trajectories because most of them are caused by outliers and uncertain noises. In this section, we report three sample actions (brush hair, riding horse, and diving) from the datasets. The actions of brush hair and riding horse are relatively static actions with slightly background movement. While the action of diving involves drastic camera motion, many points are moving together in each frame. Figure 5 illustrates the edge-based trajectory and the pruning of irrelevant trajectories. We can find that trajectories caused by camera motion are removed, and the trajectories produced by action-related edge are well kept.

3.2 Proposed spatio-temporal motion skeleton

In this section, we introduce the trajectory descriptors and the generation of spatio-temporal motion skeleton, including trajectory encoding-based similarity measurement and motion clustering.

3.2.1 Trajectory descriptors

In order to well describe the motion information of each trajectory, four descriptors are calculated in our method, namely, HOF, MBH, HOG, and the proposed STMS. The first two descriptors capture the motion information from optical flow, the HOG descriptor capture the local appearance information, and the STMS represents the relationship of trajectories between different types of action.

Similar to other trajectory-based methods, the descriptors HOG, HOF, and MBH of the LV-ET are also

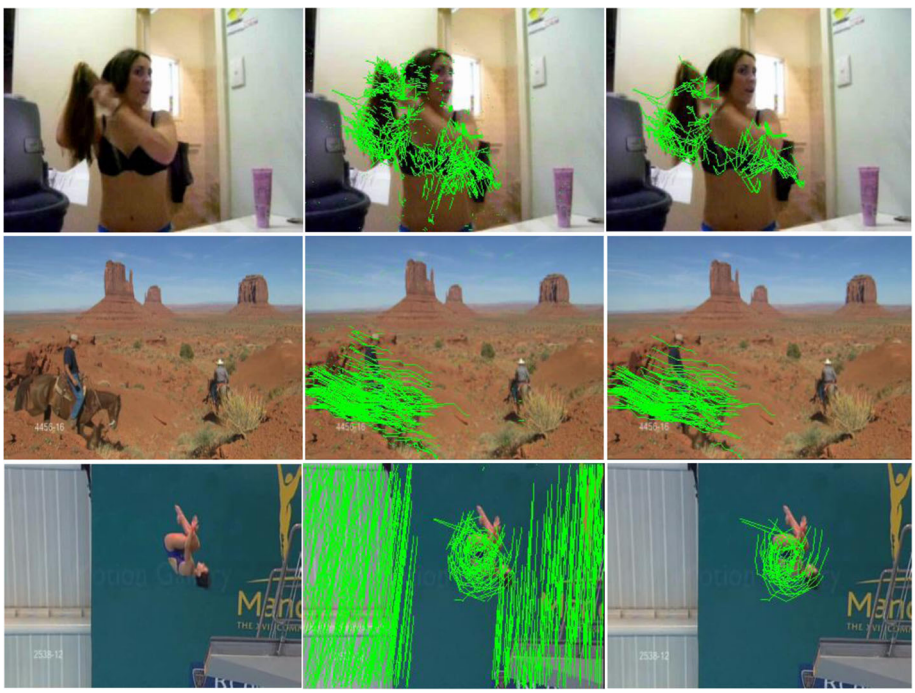

Fig. 5 Visualization of proposed LV-ET. From top to bottom: brush hair, riding horse, and diving. The above two rows correspond the case of relatively static action with slight camera motion. The bottom row corresponds to the drastic camera motion case, lots of points moving together in one frame. From left to right: a frame of an action video, the trajectories unpruned and the trajectories pruned

acquired from a spatio-temporal volume centered along the trajectory, which is usually divided into spatio-temporal cells to embed structure information. In our work, we use a volume of $24 \times 24 \times L$ pixels for each trajectory, where L is the length of LV-ET, and divide the volume into $4 \times 4 \times 1$ cells. Then, the three above descriptors are calculated with the same parameters which are used in [24]. The detailed information of proposed STMS descriptor is introduced in the next section.

3.2.2 Trajectory encoding

For a certain type of action, there are always some specific trajectories existing to describe the specified action. How to represent these specific trajectories in a proper way is a crucial problem in HAR process. In this paper, our goal is to learn discriminative spatio-temporal trajectory clusters from a video that are most relevant to a specific type of action and to encode the trajectory as motion skeleton for each action class. Our descriptor, STMS, is extracted by using spectral clustering algorithm under a novel trajectory encoding method. The illustration of extraction STMS is presented in Fig. 6. First, LV-ET can be considered as time series of 2D coordinates with different lengths $Tra_t = [(x_t,y_t),(x_{t+1},y_{t+1}),\ldots,(x_{t+L},y_{t+L})]$. We first encode every trajectory into proposed trajectory motion histogram (TMH) according to the motion information. Then, the TMH is clustered by spectral clustering algorithm, the number of the clustering centers k is discussed in detail in Section 4.3.2. Then, we regard each cluster center as the skeleton of this action and obtain the STMS by counting the numbers of trajectories belonging to each type of skeleton to gain video level representation.

For a LV-ET, there are two aspects essential to compute the similarity, namely, displacement magnitude and orientation. In this work, each LV-ET is considered as a spatio-temporal series composed of above two aspects, and we expect to extract such a compact representation to depict the motion information of each trajectory. Process to trajectory encoding and motion similarity computing is illustrated in Fig. 7. For two points P and P' between two adjacent frames along the same LV-ET, we denote the displacement between them as a vector $PP' = (\Delta x,\Delta y)$. To perform a reasonable quantization on PP', we take both magnitude and orientation into consideration. There are two quantization uniform levels for magnitude, where $||PP'||$ is normalized by its max value in each trajectory. This normalization guarantees that this quantization method is scale invariant. For orientation, we divide the 360° into eight intervals, each of which is 45°. The representation of magnitude and orientation quantization results in 16 bins in polar coordinates. Moreover, we set an additional bin for horizontal displacement of the trajectory. That is to say, the LV-ET is encoded into a TMH and the dimension of each TMH is 1×17 equally. Finally, the motion similarity is computed by Euclidean distance between two TMH.

3.2.3 Motion clustering and skeleton extraction

From the above analysis, we can obtain the motion similarity of LV-ET, which is the foundation for the motion clustering and motion skeleton extraction. In recent years, spectral clustering has become one of the most popular clustering algorithms. In this paper, we adopt this clustering similar to [33] to get the cluster center as motion skeleton. Spectral clustering algorithm is based on graph theory, we regard every TMH as a node in an undirected graph $V = (v_1,v_2,\ldots,v_n)$, the proposed motion similarity between trajectories is quantified as the edge weight between nodes. Therefore, we can transform the motion clustering problem into the sub-graph partitioning problem.

Given a set of trajectories $T = \{Tra_1,Tra_2,\ldots,Tra_n\}$ in R^l, each with its own $TMH = \{TMH_1,TMH_2,\ldots,TMH_n\}$ and that we want to cluster them into k motion skeletons:

1. Form the affinity matrix $W \in R^{n \times n}$ defined by the proposed motion similarity between two trajectories, where if $i \neq j$, $W_{ij} = \Delta Ms_{ij}$, and the elements on the diagonal is 0.

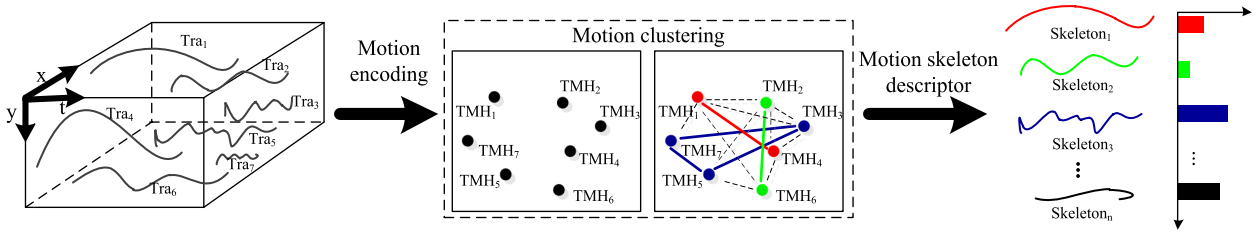

Fig. 6 Illustration of the proposed STMS descriptor based on motion similarity

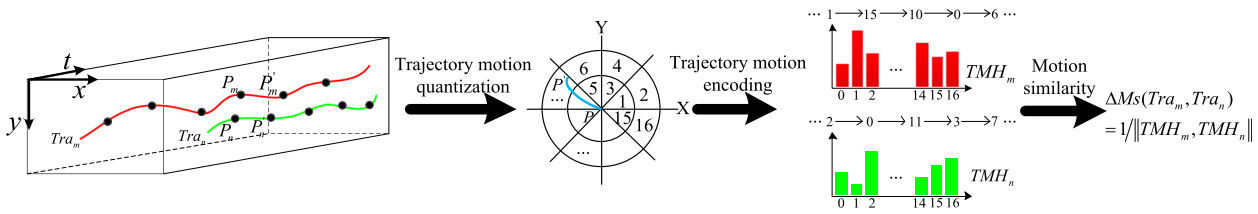

Fig. 7 Illustration of the process to trajectory encoding and motion similarity computing

2. Define D to be the diagonal matrix which diagonal element is the sum of W's i-th row, and construct the Laplacian matrix $L = D^{-1/2} AD^{-1/2}$.
3. Calculate the first k eigenvalues $x_1, x_2,\ldots,x_k$ of L and form the matrix $X = [x_1x_2\ldots x_k] \in R^{n \times k}$ by stacking the eigenvectors in columns.
4. Form the matrix Y from X by normalizing every X's row to a unit length, where $Y_{ij} = X_{ij}/(\sum_j X_{ij}^2)^{1/2}$.
5. Regarding each row of Y as a point in R^k, cluster them into k cluster $C_1, C_2,\ldots,C_k$ via K-means algorithm.
6. The clustering centers are regarding as the motion skeleton of this action. Then, each trajectory type is assigned to its nearest cluster centroid using Euclidean distance. The STMS descriptor with a dimension of k is constructed for each type of trajectories to represent the video. In general, once we have extracted the LV-ET, we can obtain the STMS by applying the following algorithm:

4 Experimental results and discussion

In this section, we evaluate the performance of the proposed LV-ET and STMS descriptor on three challenging unconstrained HAR datasets including UCF Sports [34], YouTube [35], and HMDB51 [36]; Fig. 8 shows some examples from these datasets.

4.1 Datasets

UCF Sports [34] dataset contains ten human actions: diving, golf, swinging, kicking, weight lifting, horseback riding, running, skating, swinging bench, swinging side, and walking. For most action classes, there is considerable variation inaction performance, human appearance, camera movement, viewpoint, illumination, and background. Besides, due to its high resolution, we resize each video in it to half its original spatial to reduce the time consumption. Similar to [23], we add a horizontally flipped version of each sequence to the dataset to

Algorithm 1 Motion clustering and skeleton extraction

Input: Length variable trajectories $T=\{Tra_1, Tra_2, \ldots, Tra_n\}$ and the skeleton number k.

Output: STMS descriptor $C_1, C_2, \ldots C_k$.

For $Tra_t \in T$ **do**

TMH_t = trajectory encoding(Tra_t)

end for

For $TMH_t \in TMH$ **do**

motion similarity = $||TMH_t, TMH_{t'}||(t \neq t')$

clustering into k clusters $C_1, C_2, \ldots C_k$. using spectral clustering algorithm

end for

for each cluster C_k **do**

if $\|TMH_t, C_k\|_{min}$ **do**

$C_k = C_{k+1}$

end if

end for

Fig. 8 Sample frames from the three unconstrained HAR datasets used in our experiments. From top to bottom: UCF Sports, YouTube, and HMDB51

increase data samples and use the leave-one-out setup, i.e., testing on each original sequence while training on all other sequences except the flipped version of the tested sequence. We report average accuracy over all classes.

YouTube [35] dataset includes 11 action categories: basketball shooting, biking, diving, golf swinging, horse-back riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog. This dataset is very challenging due to large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, etc. For each category, the videos are grouped into 25 groups. In total, 1168 videos are used in the experiment. Leave-one-out setup is utilized, and average accuracy over all classes is reported as the performance measure.

HMDB51 [36] dataset is collected from a variety of sources. There are a total of 6766 videos distributed in 51 action categories. The action categories can be grouped in five types: general facial actions, facial actions with object manipulation, general body movements, body movements with object interaction, and body movements for human interaction. For evaluation, there are three distinct training and testing splits. We follow the original protocol using three train-test splits and report average accuracy over these splits.

4.2 Experimental setup

The experiment is divided into two parts: the first part is feature extraction to represent the videos, and the second part is video classification. In the first part, Farnebäck's optical flow algorithm [32] is employed to estimate optical flow field. During LV-ET extraction, we adopt Canny operator to detect frame edges and a square-shaped structuring element with 3×3 pixels is used to compute the dilated edge image. The LV-ET can be obtained through the above steps, which is also reused later to compute motion-related descriptors like HOF and MBH. Moreover, in order to extract STMS descriptor, we introduce a novel trajectory encoding method and based on the proposed encoding method, the motion similarity is computed under Euclidean distance. Then, the spectral clustering algorithm is employed to construct the motion skeleton within 256 dimensional, so the video can be represented by a 256-dimension STMS descriptor. We also extract three baseline descriptors (i.e., HOG, HOF, and MBH) from the trajectories. Empirically, we use a volume of $24 \times 24 \times L$ pixels and divide it into $4 \times 4 \times 1$ cells for each trajectory when extracting the three baseline descriptors. In order to fairly compare with the baseline trajectory DT [23] and iDT [18], the same parameters are utilized to extract the descriptors as used in [23]. Note that iDT uses human detection to detect the targets from the background. In contrast, the proposed method does not make use of human detection.

After the trajectory descriptors are computed, the principle component analysis (PCA) is individually applied to reduce the dimensionality of each descriptor (i.e., HOG, HOF, MBH, and STMS) by a factor of two as

suggested in [24] so as to better mitigating the impact of noise. Then, the fisher vector (FV) model [18] is adopted in this paper. For each type of descriptor extracted from trajectory, these PCA-reduced vectors are separately encoded into a signature vector by the FV model. Similar to [24], the projection matrix of PCA is learned using 256,000 descriptors randomly sampled from the training set. Then, we use a Gaussian mixture models (GMM) with 256 components to encode the projected descriptors as the same in [24] and apply ℓ_2 normalization [18] to each type of descriptor to obtain the video-level representation. For each video, four types of video-level representations are computed.

In the second part, after gaining the high-level video representations, we use multi-kernel learning-based support vector machine to predict action class [37], where four linear kernels are used, and each corresponds to one type of representations [38]. The one-against-all approach is adopted, and the predicted class is selected with highest score.

4.3 Evaluation of parameters

In this subsection, we compare and analyze different edge detectors and the numbers of motion skeletons for HAR.

4.3.1 Evaluation of edge detectors

As the default setting is introduced in Section 4.2, we use Canny operator as the default edge detector when extracting LV-ET. In order to evaluate the impact of edge detectors on the performance of HAR, the edge information is also detected by Sobel and Log edge detectors. Then, we compare the HAR performances between these three detectors on UCF Sports and YouTube datasets. Figure 9 illustrates the comparison result on the two datasets and some example frames from different edge detectors. It is obvious that on UCF Sports, all three edge detectors achieved appreciable results, while on YouTube, Canny detector performs better than the other two. The reasonable explanation for this phenomenon is that edge detection is the foundation of all subsequent work. So, the quality of edge LV-ET is highly depending on the integrity of edge information. For most action clips in UCF Sports, the edge of targets is relatively clear and easy to be detected by the three edge detectors. Whereas for some action clips in YouTube, the targets are relatively weak which lead to false detection or weak edge losing, thus increases the difficulty of edge detection. Among the three edge

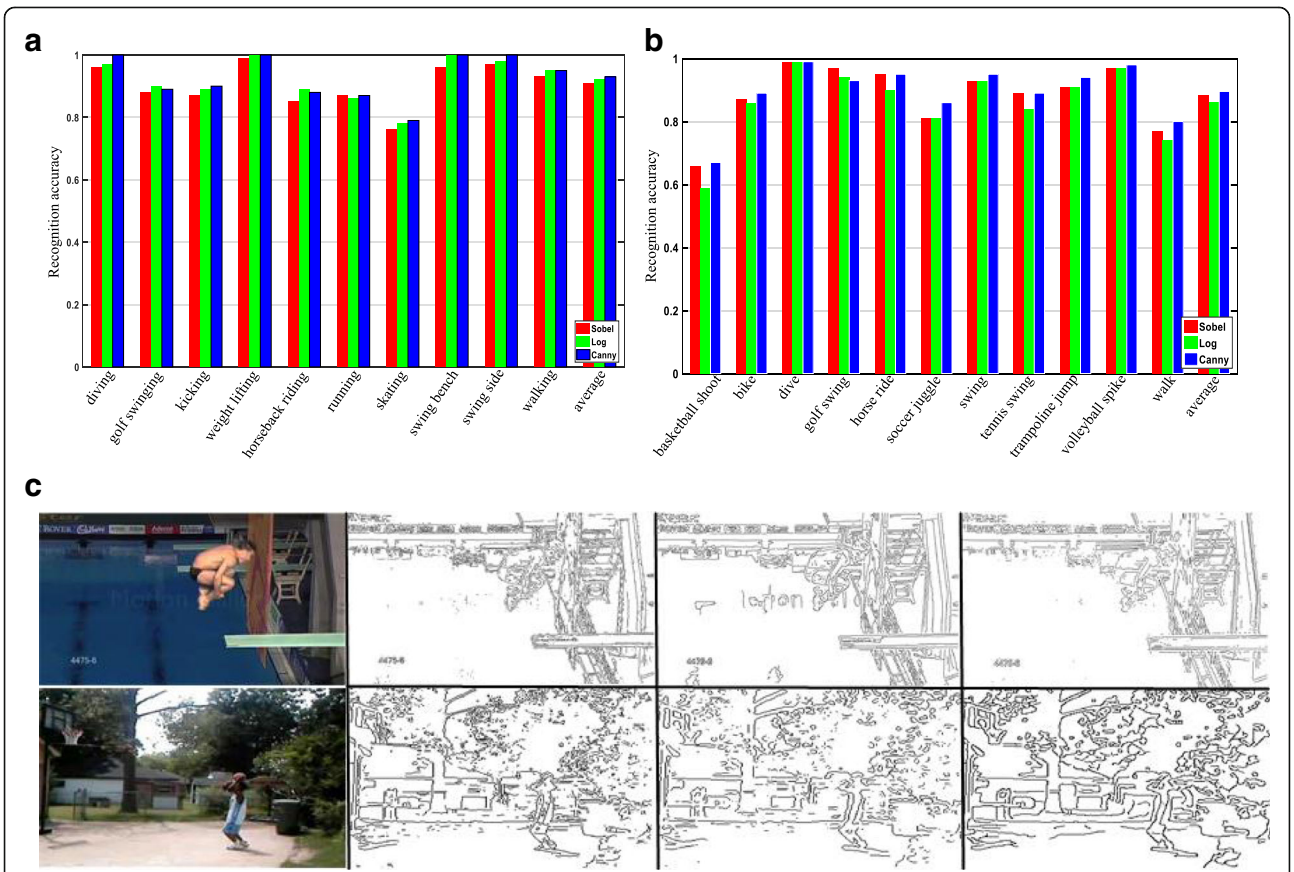

Fig. 9 Comparison of edge detectors: **a** recognition accuracy on UCF Sports, **b** recognition accuracy on YouTube, and **c** Example frames of different edge detectors. From left to right: Original, Sobel, Log and Canny

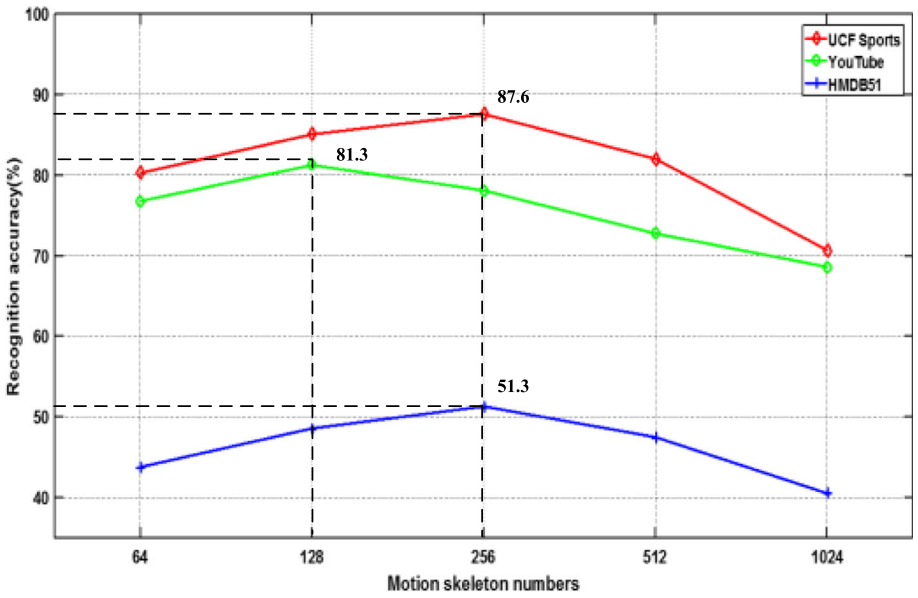

Fig. 10 Evaluation of motion skeleton number on three datasets

detectors, Canny is more likely to detect true weak edges, which provides the necessary guarantee for the follow-up works.

4.3.2 Evaluation of motion skeleton numbers

In Section 4.2, we introduced the pre-defined parameters motion skeleton number. In order to analyze the impact of this parameter, we evaluated the HAR performance of the parameter among three datasets aforementioned and the comparison result is illustrated in Fig. 10. It is obvious that in each dataset, when the motion skeleton number is too small, increasing it will improve the performance. But when it reaches a certain value, performance begins to decrease. To be specific, the performance on UCF Sports, YouTube, and HMDB51 reaches the highest respectively at 256, 128, and 256. It is partly because the resolution of action clips in YouTube is lower than that those of two.

Table 1 Comparison of STMS with baseline descriptors

Descriptors	UCF Sports (%)	YouTube (%)	HMDB51 (%)
HOG	84.3	74.3	38.4
HOF	79.2	70.9	42.5
MBH	84.5	*82.5*	48.1
STMS	*87.6*	81.3	*51.3*
Combine 1	90.4	87.0	56.6
Combine 2	*92.8*	*89.6*	*58.2*

The italicized values in the above four rows indicate the best result of single descriptor and those in the below two rows shows the best result between "Combine 1" and "Combine 2"

4.4 Comparison with baseline descriptors

A large number of experimental studies [23, 25, 26] have proved the discriminative and representative of HOG, HOF, and MBH, so we set these three descriptors as baseline descriptors. In order to have a fair comparison, both STMS and baseline descriptors are extracted from the proposed LV-ET and employ the same parameters as presented in Section 4.2. The comparison result between proposed STMS and the baseline descriptors is given in Table 1, where "Combine 1" is the method of combining three baseline descriptors, "Combine 2" is the combination of the four descriptors (i.e., HOG, HOF, MBH, and STMS). We compare the recognition results of baseline descriptors with the combination of the descriptors. It is obvious that the combination with STMS outperforms that without STMS. This indicates the discriminative and complementary of proposed STMS descriptor.

4.5 Comparison with baseline trajectories

Due to the excellent performance obtained by DT and iDT, we select both of them as baseline trajectories. The default parameters of the baseline trajectories are set as the same as in [24, 25]. In order to get the best

Table 2 Comparison of LV-ET with baseline trajectories

Trajectories	UCF Sports (%)	YouTube (%)	HMDB51 (%)
DT	88.2	84.1	52.1
iDT	89.3	86.6	57.2
LV-ET	*92.8*	*89.6*	*58.2*

The italicized values indicate the best performance among three types of trajectories

performance of iDT, a human detector is employed. In addition, the parameters of both baseline trajectories and LV-ET are configured as the same as presented in Section 4.2. The combination of the four descriptors (i.e., "Combine 2" in Table 1) is utilized to evaluate the performance, and the results are reported in Table 2, where we report the average accuracy over all three action datasets.

As given in Table 2, the HAR performance on three datasets of LV-ET outperform DT by 4.6, 5.5, and 7.1%.

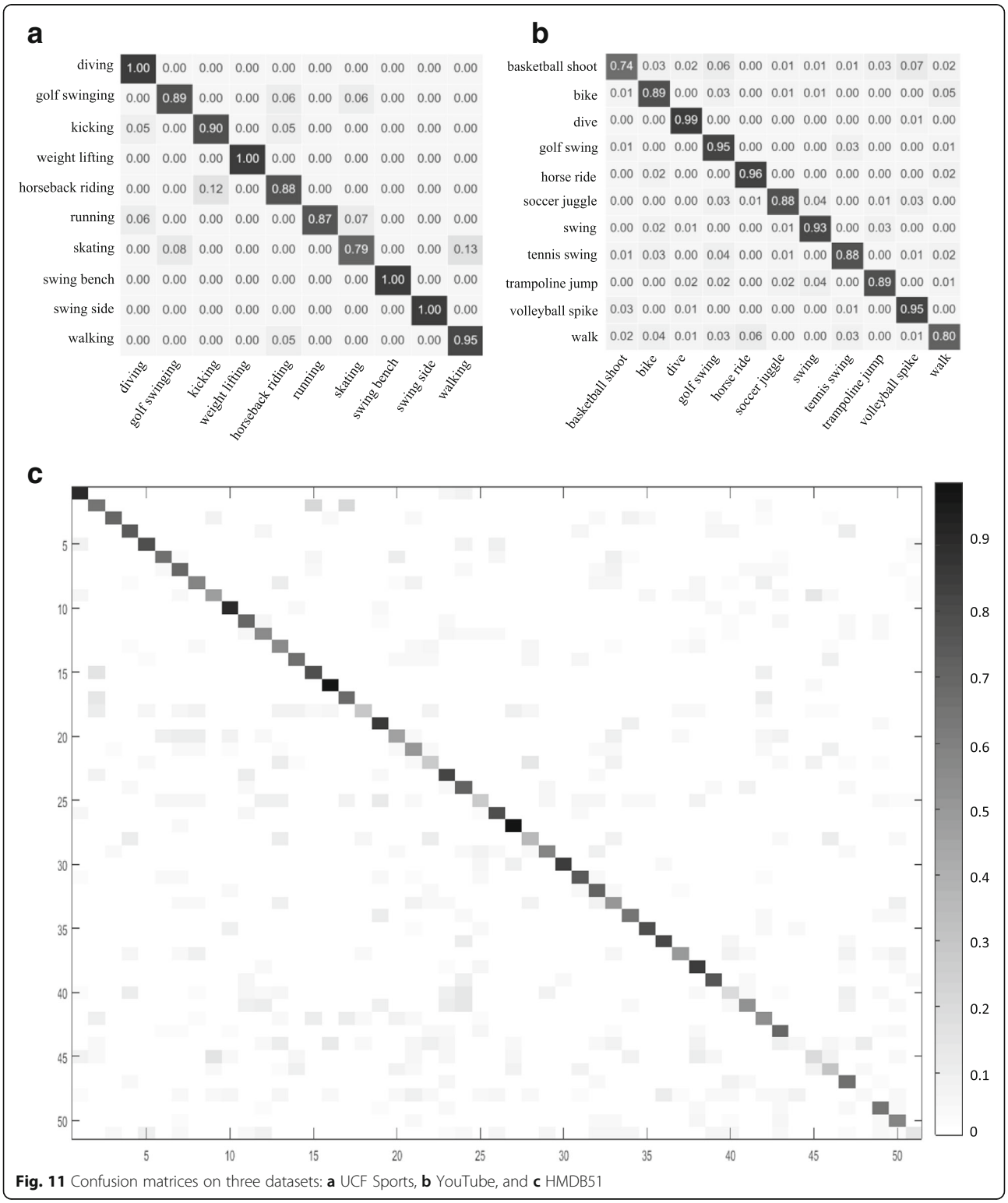

Fig. 11 Confusion matrices on three datasets: **a** UCF Sports, **b** YouTube, and **c** HMDB51

Table 3 Per class average accuracy for HMDB51

brushhair:0.89	eat:0.53	kiss:0.80	shakehands:0.62	swordexercise:0.19
cartwheel:0.63	fallfloor:0.53	laugh:0.69	football:0.76	swordfight:0.29
catch:0.67	fencing:0.64	pick:0.27	shootbow:0.78	talk:0.64
chew:0.71	backhandflip:0.76	pour:0.77	shotgun:0.48	throw:0.04
claphands:0.77	golf:0.97	pullup:0.97	sit:0.82	turn:0.62
climb:0.64	handstand:0.66	punch:0.33	situp:0.73	walk:0.57
climbstairs:0.68	hit:0.28	push:0.57	smile:0.19	wave:0.14
dive:0.57	hug:0.83	pushup:0.82	smoke:0.50	
drawsword:0.46	jump:0.44	ridebike:0.72	somersault:0.52	
dribble:0.88	kickball:0.49	ridehorse:0.69	standup:0.67	
drink:0.67	kick:0.27	run:0.49	swingbaseball:0.06	

As compared with iDT, the performances are increasing 3.5, 3.3, and 2.0% respectively. It is obvious that the proposed LV-ET obtains the best HAR performance among three types of trajectories. This is because LV-ET describes the evolution features between different types of actions and uses the edge information of the target to reduce background interference and camera motion.

4.6 Evaluation of the overall recognition performance

We report the overall recognition performance over three datasets mentioned above. Figure 11 shows the confusion matrix for UCF Sports, YouTube, and HMDB51 respectively. As seen from Fig. 11a, it can be observed that the actions of diving, weight lifting, swing bench, and swing side can all be identified and the average recognition accuracy is 92.80% which is a sufficiently high accuracy on UCF Sports. In Fig. 11b, it can be observed that the average recognition accuracy is 89.64% which is comparable to the state-of-the-art. The recognition accuracy among these actions is over 85% except basketball shoot and walk, which are confused with each other. This is one of the problems that we need to consider in the future.

Figure 11c shows the confusion matrix of HMDB51, and we use a heatmap to represent the confusion matrix. The color legend is drawn at the right and the detailed per-class accuracy is given in Table 3. We can see that the per class accuracy of smile, swing baseball, sword exercise, throw, and wave are all less than 20%. Smile is confused with chew, laugh, and talk which are all in the group of general facial actions. Sword exercise is most confused with draw sword, and throw is confused with kick ball. These errors mainly occur between classes which are visually similar. Table 3 also shows that the proposed method gains a relatively low performance in the type of body movements with object interaction like pick, sword exercise, swing baseball and throw. In the future, we will consider how to better distinguish the action in one type of group especially the group of body movements with object interaction.

Table 4 Comparison of the overall performance of our method and the trajectory-based methods

Methods	UCF Sports (%)	YouTube (%)	HMDB51 (%)
Wang et al. [23]	88.0	84.1	46.6
Seo et al. [16]	–	–	57.8
Yi et al. [30]	90.0	–	–
Wang et al. [39]	92.0	88.9	54.3
Peng et al. [40]	–	87.6	51.8
Cho et al. [41]	89.7	86.1	–
Ours	*92.8*	*89.6*	*58.2*

The italicized values indicate the best HAR performance among the three datasets

4.7 Comparison with state of the arts

To further verify the effectiveness of the proposed method, we compare our method with the recent trajectory-based methods on all three datasets and the results are reported in Table 4. For the sake of fairness, we only report the results for the case in which we combined all descriptors. For the three datasets evaluated, we can observe that the proposed method achieves a comparable HAR performance.

As given in Table 4, the proposed method achieves comparable result on all three datasets. We note that the UCF Sports dataset, the proposed method obtains at least 0.8% improvement and obtains 92.8% accuracy; for the HMDB51 dataset, there is at least 1.5% improvement compared with other methods and obtains 58.2% accuracy; for the YouTube dataset, our method outperforms 0.7% than the others and obtains 89.6% accuracy.

Table 5 Comparison of the average time consumption per video for the trajectory

	Trajectory numbers	Trajectory extraction (s)	Descriptor computation (s)	Descriptor encoding (s)	Total (s)
DT	26,749	7.04	70.48	10.17	87.69
LV-ET	2354	19.15	6.39	2.86	28.40

4.8 Evaluation of computational complexity

To evaluate the computational complexity of the proposed LV-ET, we compute the average time consumption per video for both trajectory extraction and descriptor representation on UCF sports. The experiment compared the proposed LV-ET with DT in Matlab with Intel I7 (3.6GHz CPU), only a single CPU core is used, and both of the codes are not optimized. The time consumption for the above two trajectories are compared in Table 5. The extraction of LV-ET is somewhat more computationally expensive than DT. It is because LV-ET use edge-based sampling strategy and needs to traverse every edge points in the next frame which is time consuming. But the number of LV-ET is much smaller than DT, so the descriptor computation and encoding process for the former is much more time-saving. In addition, we do not count the time consumption on optical flow, PCA, and GMM training because the process between these two types of trajectory is similar.

5 Conclusions

In this paper, a new trajectory generation strategy LV-ET is proposed and a novel descriptor STMS is designed for human action recognition. The LV-ET, extracted by tracking edge points across video frames based on optical flow with the aim of better descript the evolution of different type of actions, which proves representative and informative. In the process of extracting STMS, a novel encoding method for trajectory clustering is proposed. The motion similarity is adequately considered during TMH clustering. STMS is designed for extracting the most representative trajectories in one action, so we call it motion skeleton.

Through experimentation with three publicly unconstrained datasets, we demonstrated that the proposed LV-ET outperforms the baseline approach (e.g., DT and iDT). Regarding STMS, it is also comparable to the current trajectory-based descriptors and proved to be discriminative and complementary to existing descriptors. Note that we do not leverage background subtraction and thus can be well applied to unconstrained realistic action recognition.

Acknowledgements

Not applicable.

Funding

This work was supported by the National Natural Science Foundation of China under Grants 11176016 and 60872117.

Availability of data and materials

All data are fully available without restriction.

Authors' contributions

ZKW implemented the core algorithm and drafted the manuscript. YPG reviewed and edited the manuscript. All authors discussed the results and implications, commented on the manuscript at all stages, and approved the final version.

Authors' information

Zhengkui Weng was born in Jiaxing, Zhejiang Province, China. He is now a Ph.D. candidate of School of Communication and Information Engineering in Shanghai University, China. His major research interests include computer vision and pattern recognition.

Yepeng Guan was born in Xiaogan, Hubei Province, China, in 1967. He received the B.S. and M.S. degrees in physical geography from the Central South University, Changsha, China, in 1990 and 1996, respectively, and the Ph.D. degree in geodetection and information technology from the Central South University, Changsha, China, in 2000. Since 2007, he has been a professor with School of Communication and Information Engineering, Shanghai University.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Received: 14 August 2017 Accepted: 22 January 2018
Published online: 01 February 2018

References

1. R Poppe, A survey on vision-based human action recognition. Image Vis. Comput. **28**(6), 976–990 (2010)
2. Y Yi, M Lin, Human action recognition with graph-based multiple-instance learning. Pattern Recogn. **53**, 148–162 (2016)
3. C Yan et al., Supervised hash coding with deep neural network for environment perception of intelligent vehicles. IEEE Trans. Intell. Transp. Syst. PP(**99**), 1–12 (2017)
4. C Yan et al., Effective Uyghur language text detection in complex background images for traffic prompt identification. IEEE Trans. Intell. Transp. Syst. PP(**99**), 1–10 (2017)
5. GI Parisi, C Weber, S Wermter, Self-organizing neural integration of pose-motion features for human action recognition. Front. Neurorobotics **9**, 3 (2015)
6. C Yan et al., A highly parallel framework for HEVC coding unit partitioning tree decision on many-core processors. IEEE Signal Process. Letters **21**(5), 573–576 (2014)
7. H Zheng, Z Li, Y Fu, Efficient human action recognition by luminance field trajectory and geometry information. Transplant. Proc. **42**(3), 987–989 (2009)
8. C Yan et al., Efficient parallel framework for HEVC motion estimation on many-Core processors. IEEE Trans. Circuits Syst. Video Technol. **24**(12), 2077–2089 (2014)
9. J Dou, J Li, Robust human action recognition based on spatio-temporal descriptors and motion temporal templates. Optik Int. J. Light Electron Opt. **125**(7), 1891–1896 (2014)
10. I Laptev, On space-time interest points. IJCV. Int. J. Comput. Vis. **64**(2), 107–123 (2005)
11. O Kliper-Gross et al., Motion interchange patterns for action recognition in unconstrained videos. Eur. Conf. Comput. Vision, 256–269 (2012)
12. A Oikonomopoulos et al., Trajectory-based representation of human actions. ICMI 2006 IJCAI 2007 Int. Conf. Artifical Intell. Hum. Comput., 133–154 (2007)
13. A Ciptadi, MS Goodwin, JM Rehg, Movement pattern histogram for action recognition and retrieval. Eur. Conf. Comput. Vision, 695–710 (2014)
14. B Ni et al., Motion part regularization: Improving action recognition via trajectory group selection. IEEE Conf. Comput. Vision Pattern Recog., 3698–3706 (2015)

15. Y Yi, H Wang, Motion keypoint trajectory and covariance descriptor for human action recognition. Vis. Comput., 1–13 (2017). https://doi.org/10.1007/s00371-016-1345-6
16. JJ Seo et al., Effective and efficient human action recognition using dynamic frame skipping and trajectory rejection. Image Vision Comput. **58**, 76–85 (2016)
17. PF Felzenszwalb et al., Object detection with discriminatively trained part-based models. IEEE Trans. Patt. Anal. Mach. Intell. **32**(9), 1627 (2010)
18. J Sanchez et al., Image classification with the fisher vector: Theory and practice. Int. J. Comput. Vis. **105**(3), 222–245 (2013)
19. J Sun et al., Hierarchical spatio-temporal context modeling for action recognition. Comput. Vision Pattern Recog., 2004–2011 (2009)
20. P Matikainen, M Hebert, R Sukthankar, Trajectons: Action recognition through the motion analysis of tracked features. IEEE Intern. Conf. Comput. Vision Workshops, 514–521 (2009)
21. M Bregonzio et al., Discriminative topics Modelling for action feature selection and recognition. Br. Mach. Vision Conf., 1–11 (2010)
22. N Sundaram, T Brox, K Keutzer, Dense point trajectories by GPU-accelerated large displacement optical flow. Eur. Conf. Comput. Vision, 438–451 (2010)
23. H Wang et al., Dense trajectories and motion boundary descriptors for action recognition. Int. J. Comput. Vis. **103**(1), 60–79 (2013)
24. H Wang, C Schmid, Action recognition with improved trajectories. IEEE Int. Conf. Comput. Vision, 3551–3558 (2014)
25. N Dalal, B Triggs, Histograms of oriented gradients for human detection. Comput. Vision Patt. Recog. **1**(12), 886–893 (2005)
26. I Laptev et al., Learning realistic human actions from movies. Comput. Vision Pattern Recogn., 1–8 (2008)
27. A Gaidon, Recognizing activities with cluster-trees of tracklets. Br. Mach. Vision Conf. 3201–3208 (2012)
28. I Atmosukarto, N Ahuja, B Ghanem, Action recognition using discriminative structured trajectory groups. Appl. Comput. Vision, 899–906 (2015)
29. E Vig, M Dorr, D Cox, Space-variant descriptor sampling for action recognition based on saliency and eye movements. Eur. Conf. Comput. Vision, 84–97 (2012)
30. Y Yi, Y Lin, Human action recognition with salient trajectories. Signal Process. **93**(11), 2932–2941 (2013)
31. I Jargalsaikhan et al., Action recognition based on sparse motion trajectories. IEEE Int. Conf. Image Process., 3982–3985 (2014)
32. G Farnebäck, Two-frame motion estimation based on polynomial expansion. Scand. Conf. Image Anal., 363–370 (2003)
33. A Saade, F Krzakala, Spectral clustering of graphs with the Bethe hessian. Int. Conf. Neural Inf. Process. Syst., 406–414 (2014)
34. MD Rodriguez, J Ahmed, M Shah, Action MACH a spatio-temporal maximum average correlation height filter for action recognition. Comput. Vision Patt. Recog., 1–8 (2008)
35. J Liu, J Luo, M Shah, Recognizing realistic actions from videos. Comput. Vision Patt. Recog., 1996–2003 (2009)
36. H Kuehne et al., HMDB: A large video database for human motion recognition. IEEE Int. Conf. Comput. Vision, 2556–2563 (2011)
37. JFE Frank, A review of multi-instance learning assumptions. Knowl. Eng. Rev. **25**(1), 1–25 (2010)
38. F Orabona, L Jie, B Caputo, Online-batch strongly convex multi kernel learning. Comput. Vision Patt. Recog. **119**(5), 787–794 (2010)
39. X Wang, C Qi, Action recognition using edge trajectories and motion acceleration descriptor. Mach. Vision Appl. **27**(6), 861–875 (2016)
40. X Peng, Y Qiao, Q Peng, Motion boundary based sampling and 3D co-occurrence descriptors for action recognition. Image Vision Comput. **32**(9), 616–628 (2014)
41. J Cho et al., Robust action recognition using local motion and group sparsity. Pattern Recogn. **47**(5), 1813–1825 (2014)