IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 4, 2018, accepted June 12, 2018, date of publication June 22, 2018, date of current version July 19, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2849690

Describing Local Reference Frames for 3-D Motion Trajectory Recognition

ZHANPENG SHAO[1], YOUFU LI[2], (Senior Member, IEEE), YAO GUO[2], AND XIAOLONG ZHOU[1]

[1]College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China
[2]Department of Mechanical and Biomedical Engineering, City University of Hong Kong, Hong Kong

Corresponding author: Xiaolong Zhou (zxl@zjut.edu.cn)

The work was supported by the National Natural Science Foundation of China under Grant 61603341, Grant 61673329, Grant U1509207, and Grant 61325019.

ABSTRACT Motion trajectories tracked from the points of interest can provide the key relevant features for characterizing the motion patterns in video. As the increasing number of 3-D vision sensors rises, the 3-D motion trajectories that serve as motion representations have been applied successfully to video retrieval and analysis, scene understanding, motion recognition, and so on, in existing works. Most of these works use raw data of motion trajectories directly or draw simple geometric quantities to describe the motion trajectories, whereas these simple descriptions are not intrinsically complete as they cannot feature the orientation changes of moving points along the 3-D motion trajectories. In principle, orientation changes of a single moving point in 3-D space have to been obtained by resorting to high-order derivatives, but the high-order derivatives would result in high sensitivity to noise. This paper tackles the problem by describing the local reference frames along 3-D motion trajectories, while we consider a motion trajectory as a temporal sequence of local reference frames. The maximal blurred segment of the noisy discrete curves is employed to estimate the local reference frames without high-order derivatives involved, and the local reference frame contains complete information of positions and orientations in the 3-D Euclidean space. To describe such local reference frames, we use the rotations and local square root velocities of local reference frames as the proposed descriptor to characterize the position and orientation changes of the moving points along the motion trajectories. In the experiments, we evaluate the effectiveness of the proposed descriptor by applying it to the gesture recognition on two large benchmark data sets that contain hand motion trajectories. The results show that our proposed descriptor can achieve superior performance compared to the existing descriptors and state-of-the-art methods in the 3-D motion trajectory recognition.

INDEX TERMS Motion trajectory, gesture and activity recognition, local reference frame, maximal blurred segment.

I. INTRODUCTION

As increasing number of 3D videos arise, 3D motion trajectories tracked from points of interests can provide a compact and informative clue for motion characterization. They could serve as an effective feature to retrieve and match motions in video [1], to recognize hand gestures in human-machine interaction [2], to imitate human actions for robots [3], and so on. Among these studies, 3D motion trajectories of body parts and skeleton joints were used to represent gestures and human actions in video. In most cases, motion trajectories are often described directly by raw trajectory data, such as absolute positions and relative position changes of moving points. These raw data suffer from variance to view changes (e.g., rotation, translation and scaling in 3D space) and sensitivity to noise. As such, we focus on addressing these challenges by deriving an effective and invariant descriptor with sufficient discriminability to provide substantial advantages over raw data in trajectory recognition tasks.

A variety of existing works tried to propose effective descriptions for 3D motion trajectories which include point motion trajectories [4], [6], [7] and rigid body motion trajectories (position vectors and orientation vectors) [8], [9] in 3D space. These descriptors showed very good properties in noise robustness and view invariance, and achieved superior performance when applying them in trajectory matching, retrieval, and recognition. It can be observed that a complete description for 3D motion trajectories should be able to fully characterize spatio-temporal patterns by considering both the

positions and orientations of moving points. For rigid body motion trajectories, their descriptors are usually obtained by considering motions of more than one reference points on rigid bodies [8], [9]. In many applications, however, we can only obtain a single point trajectory of a moving object in stereo vision tracking systems, such as 3D motion trajectories of blobs obtained by tracking moving body parts [10]. In this case, high-order derivatives [6], [7] have to be involved for obtaining local reference frames to derive coordinate-free invariant descriptors. Such invariant descriptors are closely related to local curvature and torsion [6] that are a kind of complete descriptors for 3D point trajectories, being able to characterize both the position and orientation changes of moving points. Nevertheless, high-order derivatives could bring a sensitivity to noise in the trajectory. Inspired by our previous work [4], [5], and [17], we are trying to estimate a local reference frame at each point along motion trajectories without computing high-order derivatives, as shown in Figure 1. We then describe such local reference frames to capture the position and orientation changes of moving points along motion trajectories, while we see a point motion trajectory as a temporal sequence of local reference frames along time instances in this paper.

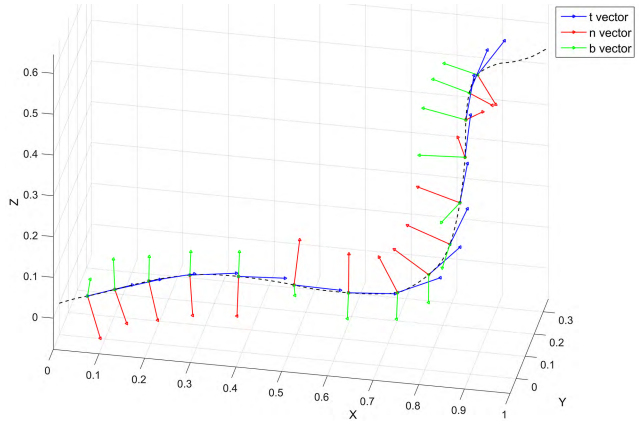


FIGURE 1. Example of a temporal sequence of local reference frames along a motion trajectory, where {t, n, b} represent the tangent vector, normal vector, and binormal vector of each local reference frame, respectively.

This paper proposes a new approach to describe the position and orientation changes of moving points along 3D point trajectories. First, at each point we estimate a local reference frame independent of view changes and scaling by employing the maximal blurred segment [4], [5], [17] of discrete curves. Basing on such temporal sequence of local reference frames, the rotation and local square-root velocity [8] of the current frame with respect to its previous frame are computed to form our proposed descriptor characterizing the position and orientation changes of moving points along motion trajectories. To evaluate the effectiveness of the proposed descriptor, we apply it to gesture recognition. An overview of the recognition pipeline based on the proposed descriptor can be found in Figure 2. The proposed descriptor shows superior recognition performance on very noisy and large gesture datasets, IP (InteractPlay) dataset and MSR-12 dataset, collected by tracking body parts with a stereo vision system and Kinect sensor, respectively.

A. RELATED WORKS ON ESTIMATION OF LOCAL REFERENCE FRAMES

The estimation of local reference frames refers to approximate their basis vectors on each point of digital lines and objects, for instance, tangent and normal vectors. In discrete geometry, tangent estimation has many applications [14], [15], such as the length estimation and curvature estimation of a digital curve. Among those estimation techniques, parametric curve fitting and digital line segments are most available common methods to obtain instant frames and their curvatures and torsions. Particularly, digital line segments are more acceptable for motion trajectories since that they do not require the point density and clean point data, showing more robustness to noisy digital curves [11]. As such, Nguyen and Debled-Rennesson [16], [17] proposes a blurred segment approach to estimate the curvatures and torsions for 3D digital shapes and curves. By using these estimation techniques of curvatures and torsions, the basis vectors for local reference frames can be obtained straightforward since that they are closely related among curvatures, torsions, and local reference frames. Thus, we employ a typical 3D maximal blurred segment [16], [17] to estimate the local reference frames along 3D motion trajectories.

II. DESCRIBING LOCAL REFERENCE FRAMES

A. ESTIMATION OF LOCAL REFERENCE FRAMES

A 3D motion trajectory is a set of position vectors of a moving object in 3D Euclidean space. Normally, it can be represented by a set of triple parametric functions with respect to the time t, $\Gamma\ (t) = \{x(t), y(t), z(t) | t \in [1, N]\}$, where N is the trajectory length.

Estimating the instant local reference frame at each point is the key problem to propose our approach for motion trajectory description. In 3D Euclidean space, Frenet-Serret frame is a special moving frame which describes the kinematic properties of a particle along a motion trajectory [5], where each point of a parametric trajectory is associated with a set of triple orthogonal unit vectors to describe its dynamic properties: 1) the tangent vector $\mathbf{t}$; 2) normal vector $\mathbf{n}$; and 3) binormal vector $\mathbf{b}$, as shown in Figure 1. The three vectors of the Frenet-Serret frame ($\mathbf{t}$, $\mathbf{n}$, $\mathbf{b}$) at point $p_t \in \Gamma$ are an orthogonal basis constructed from the Gram-Schmidt process to the vectors and their derivatives, defined as follows,

$$F(t) = \{\mathbf{t}(t), \mathbf{n}(t), \mathbf{b}(t)\}, \tag{1}$$

where,

$$\begin{cases} \mathbf{t}(t) = \dfrac{\Gamma'(t)}{\|\Gamma'(t)\|} \\ \mathbf{n}(t) = \dfrac{\mathbf{t}'(t)}{\|\mathbf{t}'(t)\|} = \dfrac{\Gamma'(t) \times \left(\Gamma''(t) \times \Gamma'(t)\right)}{\|\Gamma'(t)\| \, \|\Gamma''(t) \times \Gamma'(t)\|} \\ \mathbf{b}(t) = \mathbf{t}(t) \times \mathbf{n}(t) \end{cases} \tag{2}$$

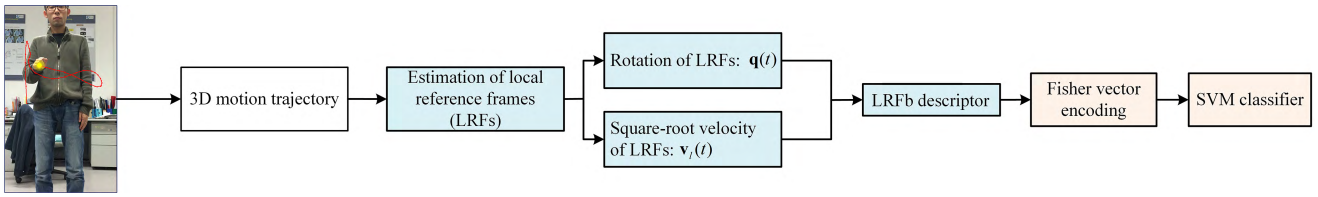


FIGURE 2. Overview of the proposed approach for 3D motion trajectory recognition. Input can be a 3D motion trajectory of the center of any particular blob obtained by vision tracking. By using the maximal blurred segments of discrete curves, we estimate the local reference frames (LRFs) along the motion trajectory. As the motion trajectory is considered as a temporal sequence of LRFs, the rotations and square-root velocities of the LRFs are calculated to form the proposed descriptor (LRFb). Finally, Fisher vector encoding is employed to encode such descriptor into a representation, which is input into a simple linear SVM classifier to do motion trajectory recognition.

In this manner, all the points in a 3D motion trajectory can be represented by their unit vectors, $(\mathbf{t}, \mathbf{n}, \mathbf{b})$. In this paper, we employ a 3D maximal blurred segment [14], [17] to estimate Frenet-Serret frames as our local reference frames, avoiding high-order derivatives.

3D maximal blurred segment (MBS) of noise curves [5], [14], [17] is used to decompose a discrete noisy trajectory into some consecutive overlapped minimally thin blurred segments via eliminating and bypassing those noisy points. Next, a number of pairs of the corresponding left and right key points of the blurred segments nearby each reference point are obtained, so that the local reference frames can be constructed by those non-collinear triple points. The detailed description of MBS [5] is briefly recalled as follows.

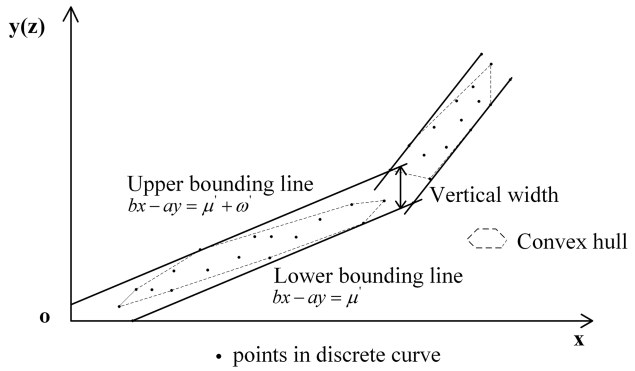


FIGURE 3. Optimal bounding line *D* of two successive blurred segments of a discrete curve in *OXY* plane [4].

Throughout this paper, we use $\Gamma\ (i, j)$ to denote a segmented trajectory from time instance i to j of the trajectory Γ. As the motion trajectory can be seen as a discrete curve, we segment a known 3D discrete curve into a number of 3D discrete lines (blurred segments) $D_{3D}(a, b, c, \mu, \mu', \omega, \omega')$ [11] of the width v that is to control the segmentation level of a sequence of points $\Gamma\ (i, j)$ in the 3D discrete curve such that $\omega' - 1/\max\ (|a|, |b|) \leq v$ on the plane OXY and $\omega - 1/\max\ (|a|, |c|) \leq v$ on the plane OXZ. A toy example of blurred segments of a discrete curve on the plane OXY is shown in Figure 3, where there are two consecutive blurred segments and the width is determined by the dynamical thickness estimation of convex hulls [12] that consists of a set of successive discrete points as shown in Figure 3.

Faure *et al.* [14] and Nguyen and Debled-Rennesson [17] further proposed the concept of MBS of width v, which means each segmented 3D discrete line for a discrete curve cannot be extend neither at right side nor at left side for given width v. Basing on this concept, we can decompose a 3D discrete trajectory Γ into a sequence of intercrossed maximal blurred segments (MBSs) of width v with m length:

$$MBS_v(\Gamma) = \{MBS(B_1, E_1, v), \ldots, MBS(B_m, E_m, v)\}, \quad (3)$$

with $B_1 < B_2 < \cdots < B_m$ and $E_1 < E_2 < \cdots < E_m$. $\{B_i, E_i\}\,|i \in [1, m]$ denotes the beginning and ending positions of a maximal blurred segment of the discrete trajectory. Given a sequence of maximal blurred segments $MBS_v(\Gamma)$, to estimate the Frenet-Serret frame at a point, we first denote the estimated key points of the discrete trajectory with $\Gamma\ (B_i), \Gamma\ (E_i)|i \in [1, m]$ from the MBSs of the discrete trajectory. Then let $\{R(t) \in \{B_i, E_i\}, L(t) \in \{B_i, E_i\}\,|t = 1 \cdots N\}$ record a sequence of positions of the estimated right nearest key points and left nearest key points at each reference point $\Gamma\ (t)$ such that: $L(t) < t < R(t)$, where we assume these triple points $\{\Gamma\ (R(k)), \Gamma\ (t), \Gamma\ (L(t))|t = 1 \cdots N\}$ to be always not collinear. We then approximate the osculating circle at $\Gamma\ (t)$ using the circumcircle of the triangle bounded by these triple points. Let $C(t)$ be the center of the circumcircle.

Then, we define the norm vector at $\Gamma\ (t)$ as $\mathbf{n}(t) = \overrightarrow{\Gamma\ (t)C(t)} \Big/ \left|\overrightarrow{\Gamma\ (t)C(t)}\right|$. The unit tangent vector $\mathbf{t}(t)$ is the unit vector that is tangent with the osculating circle at point $\Gamma\ (t)$. Then the binormal vector $\mathbf{b}(t)$ is obtained straightforwardly by cross product: $\mathbf{b}(t) = \mathbf{t}(t) \times \mathbf{n}(t)$. The local Frenet-Serret frame is then defined as, $F(t) = \{\mathbf{t}(t), \mathbf{n}(t), \mathbf{b}(t)\}$.

B. ROTATION AND SQUARE-ROOT VELOCITY OF LOCAL REFERENCE FRAMES

Basing on a temporal sequence of local reference frames, we use a combined vector of quaternion and local square-root velocity to describe the position and orientation changes of the local reference frame at each time instance,

$$\mathbf{s}_t = \left[\mathbf{q}(t), {}_{\{\mathbf{R}\}}\mathbf{v}_l(t)\right], \{\mathbf{s}_t\} \in \mathbb{R}^{7\times N}, \quad (4)$$

where the quaternion $\mathbf{q}(t)=[q_w(t), q_x(t), q_y(t), q_z(t)]$, $\mathbf{q} \in \mathbb{R}^{4\times N}$ describes the rotation of the current local frame with respect to the previous frame. ${}_{\{\mathbf{R}\}}\mathbf{v}_l(t) = \mathbf{R}(t)^{\mathrm{T}}\mathbf{v}_g(t)$ denotes the local square-root velocity that is obtained by projecting square-root velocity vector $\mathbf{v}_g(t)$ of each local reference frame in the world coordinate into the relative velocity vector with respect to the local reference frame. The unit vector $\mathbf{v}_g(t)$ is the SRVF [20], which is an elastic metric based representation with many advantages and defined as

$$\mathbf{v}_g(t)=\frac{\dot{\Gamma}(t)}{\sqrt{\|\dot{\Gamma}(t)\|}}, \tag{5}$$

where $\dot{\Gamma}(t)$ denotes the first-order derivative with respect to t.

According to Euler's rotation theorem, a sequence of rotations can be equivalent to a single rotation by a given angle β about a unit vector $\hat{\boldsymbol{w}}$. The unit quaternion can provide a simple way for encoding such axis-angle representation in four numbers,

$$\mathbf{q} = \left[\cos\left(\frac{\beta}{2}\right), \hat{\boldsymbol{w}}^{\mathrm{T}} \sin\left(\frac{\beta}{2}\right)\right], \tag{6}$$

where the unit vectors $\hat{\boldsymbol{w}}$ can be obtained by,

$$\hat{\boldsymbol{w}} = \frac{1}{2\sin\beta}\begin{bmatrix}\mathbf{R}(3,2)-\mathbf{R}(2,3)\\ \mathbf{R}(1,3)-\mathbf{R}(3,1)\\ \mathbf{R}(2,1)-\mathbf{R}(1,2)\end{bmatrix}, \tag{7}$$

where $\beta = \arccos\left(\frac{\mathrm{trace}(\mathbf{R})-1}{2}\right)$ and $\mathrm{trace}(\mathbf{R})$ is the sum of diagonal elements of $\mathbf{R}$. Suppose there are a sequence of local reference frames $F(t), t \in [1, N]$ for a motion trajectory, the rotation between a pair of time-adjacent local reference frames can be represented by a rotation matrix $\mathbf{R} \in \mathbb{R}^{3\times 3}$, which can be explicitly calculated by

$$\mathbf{R}(t) = F^{-1}(t-1)F(t). \tag{8}$$

And therefore, by (4-8) we can obtain the local reference frame based (LRFb) descriptor, $\mathbf{S} = \{\mathbf{s}_t\}$.

III. EXPERIMENTS

This section applies the proposed descriptor to human gesture recognition for evaluating its effectiveness in motion trajectory characterization and recognition, where two large and noisy benchmark datasets for gesture recognition, InteractPlay [10] and the MSRC-12 [21], are employed. In experiments, we first explore the effects on recognition performance when applying different individual descriptors on the InteractPlay dataset. Then, we compare our method with the previous state-of-the-art methods and analyze their performance. Finally, we extend our method to full body gesture recognition on MSRC-12 dataset when only two joint trajectories of the left and right hands are used, and evaluate the potential of our method in action recognition, while other methods applied on MSR-12 dataset have utilized all 20 joint trajectories as full body gesture representation.

The recognition performance is evaluated by the mean classification accuracy, which represents the average of successful classification rates for all classes in a dataset.

A. IMPLEMENTATIONS

1) FEATURE ENCODING AND CLASSIFIER

In order to apply a linear classifier on the proposed descriptor, we employ Fisher vector (FV) encoding [18] with a temporal pyramid manner [19] to encode the descriptors into representations with same lengths, where the fisher vectors are aggregated along a temporal pyramid dimension to constitute the fisher codes. More specifically, we recursively partition motion trajectories into a pyramid from 0 to Z scale, where at z- th scale there are 2^z segments in temporal dimension. We then do FV encoding on all the segments, and obtain local FVs on each segment. Thus, each representation is the concatenation of these local FVs from all the segments, and its size is $2KD^*\sum_{i=0}^{Z} 2^i$, where D is the descriptor length of $\mathbf{s}_t$ and K is the number of mixture components in the GMM model for Fisher vector encoding. With such FV encoded representations, we train a linear SVM as the classifier to recognize human gestures using the LIBSVM library [22]. In the training of the SVM classifier, we use the linear kernel and leave all the parameters in default as set in the library.

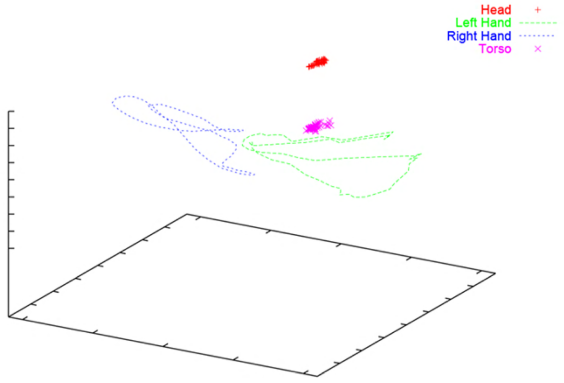
2) PARAMETER SETTINGS

All the provided experiments are conducted on trajectory-based gesture datasets. Hand trajectories tracked by stereo vision systems and Kinect sensors are provided in two datasets, respectively. We estimate the local reference frames, and extract the LRFb descriptors for each hand trajectory. The extracted LRFb descriptors of the left and right hand trajectories are concatenated along the temporal dimension into a set of feature descriptors. As the estimation of the local reference frames is based on discrete trajectories, we digitalize each motion trajectory with 1000 grids. Maximal blurred segment of the digitalized trajectory with width $v = 8$ on all datasets is then performed. In the Fisher vector encoding, we empirically use 64 mixture components to learn a GMM model. For the temporal pyramid, we subdivide hand trajectories at 3 scales in the temporal dimension, and do average pooling on FV codes of each temporal partition. Hence, the final representation for each gesture is an encoded vector of size $2^*K^*D^*15$, where $K = 64$, $D = 14$ for two hand motion trajectories in the experiments.

B. 3D INTERACTPLAY DATASET (IP)

The IP dataset [10] is a hand gesture dataset consisted of 3D hand trajectories tracked by a stereo vision system. This dataset contains 16 gestures from 22 persons and provides 5 sessions and 10 recordings. There is a total of 16000 gesture samples. The dataset contains 3D motion trajectories of the head, torso, and two hands. Figure 4(a) shows an example of the swim gesture sequence from one camera, and the corresponding trajectories of four blobs are shown in Figure 4(b). As shown in the figure, the two hand trajectories play a major role in performing gestures, and we use only hand trajectories to extract the LRFb descriptors. We follow the evaluation

(a)



(b)

FIGURE 4. (a) Example of a "swim" gesture instance from the IP dataset. From top-left to bottom-right, a frame-by-frame decomposition of a "swim" gesture instance from the point of view of the right camera, where the use of gloves with distinct colors are used to facilitate visual tracking. (b) 3D motion trajectories of the centers of particular blobs (head, torso, left hand and right hand) for a "swim" gesture. [10].

protocol in [2] and [10], where the half samples performed by 10 persons are for training, and the remaining half samples are for testing.

We first compare our experimental result with the results using existing popular descriptors on this dataset. Those existing descriptors are generated as time sequences and share the same encoding procedure as the proposed descriptor. To have a fair comparison, the linear SVM classifier is used for all the compared descriptors. As Table 1 summarizes those results, it can be observed that the best performance of 88.95% is achieved by the proposed descriptor. To see more

TABLE 1. Comparison results with existing descriptors on IP dataset.

Descriptor Used (with SVM)	Accuracy(%)
Differential Invariants [6]	41.50
Integral Invariants [4]	53.58
SRVF [20]	72.64
Multiscale SSM [23]	86.87
3D shape context [13]	65.83
LRFb (Our descriptor)	**88.95**

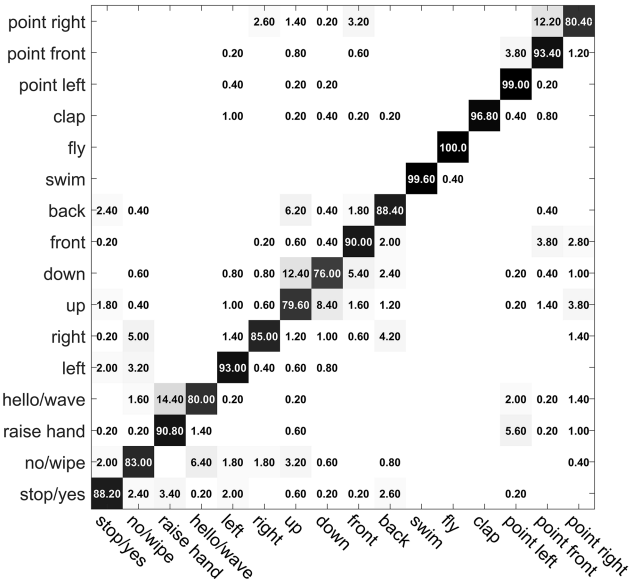


FIGURE 5. The confusion matrix of the proposed method for IP dataset.

deeply the results, the confusion matrix is given in Figure 5. We observe that most gestures are classified very well. Some mistakes occur between "up" and "down" gestures, since that these two gestures involve a same trajectory shape but different moving directions with respect to the hip. Likewise, some confusion occurs between "point right" and "point front". We then compare our method (LRFb+SVM) with current state-of-the-art methods we can find in recognizing gestures on this dataset, and summarize their results in Table 2. Our proposed descriptor with a linear SVM achieves the best performance among all methods compared.

TABLE 2. Comparison results with current state-of-the-art methods on IP dataset.

Methods	Accuracy(%)
IOHMM [10]	74.00
Continuous HMMs [24]	75.00
HDCRF [25]	80.80
MvMF-HMM [2]	85.82
Multiscale SSM [23]	86.87
LRFb+SVM	**88.95**

C. MSRC-12 DATASET

MSRC-12 dataset [21] is a large gesture dataset recorded by Microsoft Kinect, where 16 classes of gestures are collected from 30 subjects. One gesture is performed several times by an individual in each sequence. In total, there are 594 sequences consisted of 6244 gesture instances.

TABLE 3. Gesture classes in the MSRC-12 dataset and the number of annotated gesture instances from each class.

Metaphoric Gest.	No.	Iconic Gest.	No.
Start system	508	Duck	500
Push right	522	Goggles	508
Wind it up	649	Shoot	511
Bow	507	Throw	515
Had enough	508	Change weapon	498
Beat both	516	Kick	502

Each sequence contains tracks of 20 joints estimated using Kinect pose estimation pipeline and is captured at a sample rate of 30HZ with 2cm accuracy in joint positions. We segment these gesture sequences based on the labeled action points provided by [26] getting 6244 gesture instances. Table 3 lists 12 action classes in the dataset and the number of annotated action instances of each class. We follow the evaluation protocol of 50% cross-subject in [26], in which we split the segmented dataset into training set and testing set, performed by odd subjects and even subjects, respectively. As most existing works use the 20 joint trajectories for gesture recognition on this dataset, we instead only extract the left and right hand trajectories as the motion trajectories used in recognition since that two hand trajectories play a main role in most of the gestures in the MSRC-12 dataset.

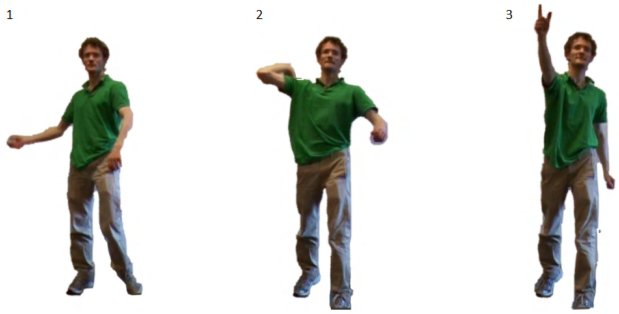


FIGURE 6. Example gesture "Throw an object" from MSRC-12 dataset.

TABLE 4. Comparison results with current state-of-the-art methods on MSRC-12 dataset.

Methods	Accuracy(%)
Cov3DJ [26]	91.60
Ellis. et al. [28]	88.70
ConvNets [27]	**93.12**
LRFb+SVM	**90.38**

Figure 6 shows an example gesture from this dataset. Table 4 lists our result and existing state-of-the-art results. The confusion matrix achieved by our method is shown in Figure 7. Our method achieves a competitive result of 90.38%, compared with the best result which however requires all 20 joint trajectories of the full body. We achieve such competitive results with only two hand trajectories. Insights into our performance can be obtained by examining the confusion matrix, where our method distinguishes all gestures very well. However, the best method [27] cannot have

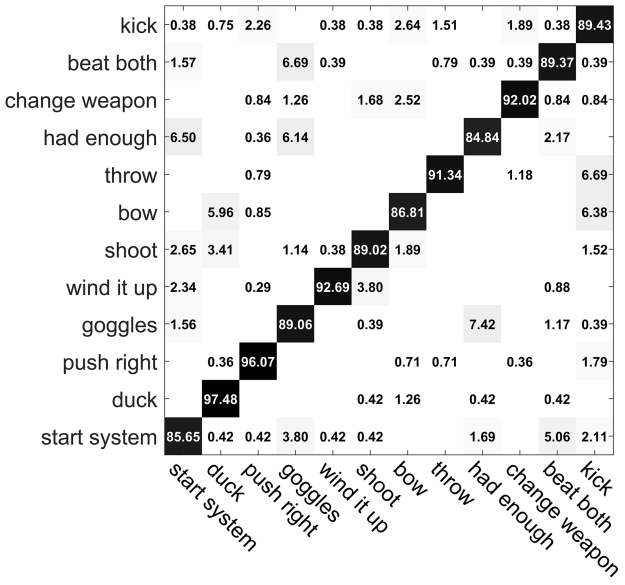


FIGURE 7. The confusion matrix of the proposed method for MSRC-12 dataset.

a good classification between "goggles" and "had enough", and achieved a classification accuracy of 72%.

IV. CONCLUSION AND FUTURE DIRECTIONS

This paper proposes an effective approach for motion trajectory description and recognition. We build the descriptor on the estimated local reference frame at each point along motion trajectories. Such local reference frame based descriptor intrinsically owns invariant properties under scaling, translations, and rotations attributed to the view independence of local reference frames. As the estimation of local reference frames is based on maximal blurred segment which can automatically bypass noisy points, the descriptor also can show robustness to noise data. While this descriptor could capture the spatio-temporal position and orientation changes of a point trajectory, its use in trajectory classification can help distinguish very well complicated motion patterns. The experimental results in large benchmarks show our descriptor is an effective descriptor for characterizing a point motion trajectory in 3D space.

By observing the experimental results, those gestures with similar trajectory shapes tend to confuse each other. That is because the proposed descriptor only considers trajectory shapes without taking their relative directions into account. We believe that the relative directions with respect to a key reference point in a particular scenario can be easily incorporated our descriptor in the future work.

REFERENCES

[1] M. W. Chao, C. H. Lin, J. Assa, and T. Y. Lee, "Human motion retrieval from hand-drawn sketch," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 5, pp. 729–740, May 2012.
[2] J. Beh, D. K. Han, R. Durasiwami, and H. Ko, "Hidden Markov model on a unit hypersphere space for gesture trajectory recognition," *Pattern Recognit. Lett.*, vol. 36, pp. 144–153, Jan. 2014.

[3] J. Aleotti, A. Cionini, L. Fontanili, and S. Caselli, "Arm gesture recognition and humanoid imitation using functional principal component analysis," in *Proc. IEEE Int. Conf. Intell. Robot. Syst.*, Tokyo, Japan, Nov. 2013, pp. 3752–3758.
[4] Z. Shao and Y. F. Li, "Integral invariants for space motion trajectory matching and recognition," *Pattern Recognit.*, vol. 48, no. 8, pp. 2418–2432, 2015.
[5] Z. Shao and Y. Li, "On integral invariants for effective 3-D motion trajectory matching and recognition," *IEEE Trans. Cybern.*, vol. 46, no. 2, pp. 511–523, Feb. 2016.
[6] D. Shandong Wu and Y. F. Li, "Flexible signature descriptions for adaptive motion trajectory representation, perception and recognition," *Pattern Recognit.*, vol. 42, no. 1, pp. 194–214, 2009.
[7] I. F. Bashir, A. A. Khokhar, and D. Schonfeld, "View-invariant motion trajectory-based activity classification and recognition," *Multimedia Syst.*, vol. 12, no. 1, pp. 45–54, 2006.
[8] Y. Guo, Y. F. Li, and Z. Shao, "RRV: A spatiotemporal descriptor for rigid body motion recognition," *IEEE Trans. Cybern.*, vol. 48, no. 5, pp. 1513–1525, May 2018.
[9] M. Vochten, T. De Laet, and J. De Schutter, "Comparison of rigid body motion trajectory descriptors for motion representation and recognition," in *Proc. IEEE Int. Conf. Robot. Autom.*, Seattle, WA, USA, May 2015, pp. 3010–3017.
[10] J. Agnès O. Bernier, and M. Sébastien "HMM and IOHMM for the recognition of mono- and bi-manual 3D hand gestures," in *Proc. Brit. Mach. Vis. Conf.*, London, U.K., 2004, pp. 1–10.
[11] I. Debled-Rennesson, F. Feschet, and J. Rouyer-Degli, "Optimal blurred segments decomposition of noisy shapes in linear time," *Comput. Graph.*, vol. 30, no. 1, pp. 30–36, 2006.
[12] L. Buzer, "A simple algorithm for digital line recognition in the general case," *Pattern Recognit.*, vol. 40, no. 6, pp. 1675–1684, 2007.
[13] A. Frome, D. Huber, R. Kolluri, T. Bülow, and J. Malik, "Recognizing objects in range data using regional point descriptors," in *Proc. Eur. Conf. Comput. Vis.*, Prague, Czech Republic, 2004, pp. 224–237.
[14] A. Faure, L. Buzer, and F. Feschet, "Tangential cover for thick digital curves," *Pattern Recognit.*, vol. 42, no. 10, pp. 2279–2287, 2009.
[15] F. D. Vieilleville and J.-O. Lachaud, "Comparison and improvement of tangent estimators on digital curves," *Pattern Recognit.*, vol. 42, no. 8, pp. 1693–1707, 2009.
[16] T. P. Nguyen and I. Debled-Rennesson, "Curvature and torsion estimators for 3D curves," in *Proc. Int. Symp. Adv. Vis. Comput.*, Las Vegas, NV, USA, 2008, pp. 688–699.
[17] T. P. Nguyen and I. Debled-Rennesson, "On the local properties of digital curves," *Int. J. Shape Model.*, vol. 14, no. 2, pp. 105–125, 2008.
[18] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proc. Eur. Conf. Comput. Vis.*, Heraklion, Greece, 2010, pp. 143–156.
[19] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3D human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 914–927, May 2014.
[20] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo, "3-D human action recognition by shape analysis of motion trajectories on Riemannian manifold," *IEEE Trans. Cybern.*, vol. 45, no. 7, pp. 1340–1352, Jul. 2014.
[21] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin, "Instructing people for training gestural interactive systems," in *Proc. ACM Annu. Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, 2012, pp. 1737–1746.
[22] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.
[23] Y. Guo, Y. F. Li, and Z. Shao, "On multiscale self-similarities description for effective three-dimensional/six-dimensional motion trajectory recognition," *IEEE Trans. Ind. Informat.*, vol. 16, no. 6, pp. 3017–3026, Dec. 2017.
[24] J. Agnès and S. Marcel, "A comparative study of two state-of-the-art sequence processing techniques for hand gesture recognition," *Comput. Vis. Image Understanding*, vol. 113, no. 4, pp. 532–543, 2009.
[25] D. Kelly, J. McDonald, and C. Markham, "Recognition of spatiotemporal gestures in sign language using gesture threshold HMMs," in *Machine Learning for Vision-Based Motion Analysis*. London, U.K.: Springer, 2011, pp. 307–348.
[26] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations," in *Proc. Int. Joint Conf. Artif. Intell.*, Beijing, China, 2013, pp. 2466–2472.
[27] P. Wang, Z. Li, Y. Hou, and W. Li, "Action recognition based on joint trajectory maps using convolutional neural networks," in *Proc. ACM Multimedia Conf.*, Amsterdam, The Netherlands, 2016, pp. 102–106.
[28] C. Ellis, S. Z. Masood, M. F. Tappen, J. J. LaViola, Jr., and R. Sukthankar, "Exploring the trade-off between accuracy and observational latency in action recognition," *Int. J. Comput. Vis.*, vol. 101, no. 3, pp. 420–436, 2013.

ZHANPENG SHAO received the B.S. and M.S. degrees in mechanical engineering from Xi'an University of Technology, Xi'an, China, in 2004 and 2007, respectively, and the Ph.D. degree in computer vision from the City University of Hong Kong, Hong Kong, in 2015. From 2015 to 2016, he was a Senior Research Associate with the Shenzhen Research Institute, City University of Hong Kong. In 2016, he joined the Zhejiang University of Technology, China, where he is currently an Associate Professor with the College of Computer Science and Technology. His current research interests include computer vision, pattern recognition, machine learning, and robot sensing. He received the Best Conference Paper Award at the ICMA 2014 and ICMA 2016.

YOUFU LI (M'91–SM'01) received the B.S. and M.S. degrees in electrical engineering from the Harbin Institute of Technology, Harbin, China, and the Ph.D. degree in robotics from the Department of Engineering Science, University of Oxford, Oxford, U.K., in 1993.

From 1993 to 1995, he was a Research Staff with the Department of Computer Science, University of Wales, Aberystwyth, U.K. In 1995, he joined the City University of Hong Kong, Hong Kong, where he is currently a Professor with the Department of Mechanical and Biomedical Engineering. His current research interests include robot sensing, robot vision, 3-D vision, and visual tracking.

Dr. Li has served as an Associate Editor for the IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING. He is currently serving as an Associate Editor for the *IEEE Robotics & Automation Magazine*. He is an Editor of the IEEE Robotics & Automation Society Conference Editorial Board and the IEEE Conference on Robotics and Automation.

YAO GUO received the B.S. and M.S. degrees from Sun Yat-sen University, Guangzhou, China, in 2011 and 2014, respectively. He is currently pursuing the Ph.D. degree with the Department of Mechanical and Biomedical Engineering, City University of Hong Kong, under the supervision of Prof. Y. Li. His research interests include pattern recognition, machine learning, robot sensing, and robot vision.

XIAOLONG ZHOU received the Ph.D. degree in mechanical engineering from the Department of Mechanical and Biomedical Engineering, City University of Hong Kong, Hong Kong, in 2013. From 2015 to 2016, he was a Senior Research Fellow with the School of Computing, University of Portsmouth, Portsmouth, U.K. In 2014, he joined the Zhejiang University of Technology, Hangzhou, Zhejiang, China, where he is currently an Associate Professor with the College of Computer Science. He has authored over 50 peer-reviewed international journals and conference papers. His research interests include visual tracking, gaze estimation, 3-D reconstruction, and their applications in various fields. He serves as an ACM member. He has served as a Program Committee Member on ROBIO 2015, ICIRA 2015, SMC 2015, HSI 2016, ICIA 2016, and ROBIO 2016. He received the T.J. Tarn Best Paper Award at the ROBIO 2012 and the ICRA 2016 CEB Award for best reviewers.