MSM-HOG: A Flexible Trajectory Descriptor for
Rigid Body Motion Recognition

Yao Guo1, You-Fu Li∗1,2, Senior Member, IEEE and Zhanpeng Shao3

Abstract—This paper proposes a flexible descriptor for representing 6-D rigid body motion trajectories, which not only shows strong invariances and descriptive ability but also achieves satisfactory results in both recognition accuracy and efficiency. 6-D rigid body motion trajectories are first transformed into the Multi-layer Self-similarity Matrices (MSM) representation. The MSM is the combination of the square similarity matrices in three layers, which captures both local and global spatiotemporal features of the trajectories. Next, the Histogram of Oriented Gradients (HOG) features extracted from the MSM representation are concatenated as the final MSM-HOG trajectory descriptor. Then we train the Support Vector Machine (SVM) classifier with the linear kernel for multiclass motion recognition. Finally, rigid body motion recognition experiments on two public datasets are conducted to verify the effectiveness and efficiency of the proposed method.

I. INTRODUCTION

One of the prospective topics in human-robot interaction is that intelligent robots are able to understand and recognize motion behaviors of different objects from their visual observations. Motion trajectories in 3-D space offer the concise expressions and informative clues in the motion characterization. Trajectory-based motion analysis has been extensively investigated in various applications, including human-machine interaction [1], robotics [2] and leisure industry [3]. One of the main challenges in trajectory-based motion recognition is that raw trajectories representing the same motion may exist various variations. On the one hand, rigid transformations (rotation, translation, and scaling) mainly originate from the changing of the vision capturing system. On the other hand, the local and global nonanalytical variations (shape, execution rate, length and etc.) are produced while the same motion is performed by different individuals. Hence, an invariant descriptor for representing motion trajectories can offer substantial advantages over raw data in the description, thus to achieving effective trajectory-based motion recognition.

In the past decades, the invariant descriptions for 3-D point trajectories have attracted extensive attentions. The previous works can be categorized into the following groups: 1) shape descriptors [4], [5], [6]; 2) transformation-based descriptors [7], [8]; 3) kinematic-based descriptors [9], [11], [10]. The proposed transformation-based and shape descriptors are the extensions of 2-D curves, in which the local features and temporal clues in motion trajectories are neglected. To overcome this, a number of kinematic-based descriptors have been proposed by calculating differential or integral invariants, which have achieved successful applications in 3-D trajectory matching and recognition. However, the motion behavior of an object cannot be uniquely characterized without considering the 3-D rotations over time. In recent years, the invariant descriptions for 6-D rigid body motion trajectories (3-D positions and 3-D rotations of a body over time) are pervasively investigated [12], [13], [14]. However, the output descriptors for a trajectory are generated as time sequences with the same dimensions as the original trajectory, and the template matching method is commonly utilized in the classification stage. Therefore, these methods consume high cost in time (pair-wise warping and comparison) and space (storing training descriptors).

Inspired by these, this paper proposes a novel framework for rigid body motion recognition by devising a flexible trajectory descriptor, which balances the tradeoff between recognition performance and computational cost. 6-D rigid body motion trajectories are firstly represented by Multi-layer Self-similarity Matrices (MSM). The MSM representation captures the local and global spatiotemporal features at trajectory and component levels, which has shown strong invariances and descriptive ability. As the similarity matrices can be viewed as gray-scale images [15], the well-proven Histogram of Oriented Gradients (HOG) [16] features extracted from the MSM representation are concatenated as the final MSM-HOG trajectory descriptor. The MSM-HOG descriptors for different trajectories are of the same dimensions, which are the input of the classifier. Finally, the Support Vector Machine (SVM) with a linear kernel is applied for multiclass recognition tasks. Fig. 1 shows the pipeline of the proposed approach.

The rest of this paper is organized as follows. 6-D rigid body motion trajectory is defined section II. In section III, the MSM representation and the invariant properties are introduced. In addition, the construction of the final MSM-HOG descriptor is given. Finally, the trajectory-based motion recognition experiments are conducted in section IV.

II. PROBLEM STATEMENT

A. 6-D rigid body motion trajectory

A 6-D rigid body motion trajectory recorded in the global coordinate system \{E\} is demonstrated in Fig. 2, which contains the 3-D position vectors of a reference point on

This work was supported by a grant from the Research Grant Council of Hong Kong [CityU 11205015] and NSFC [61673329].

1Yao Guo and You-Fu Li are with the department of Mechanical and Biomedical Engineering, City University of Hong Kong, 83 Tat Chee Avenue, Hong Kong (*Corresponding email: meyfli@cityu.edu.hk)
2You-Fu Li is also with Shenzhen Research Institute, City University of Hong Kong, Shenzhen, China
3Zhanpeng Shao is with Department of Computer Science and Technology, Zhejiang University of Technology, China.

978-1-5386-2682-5/17/$31.00 ©2017 IEEE 4298
the object and the 3-D rotations of this rigid body over time. Various representations (Euler angles, rotation matrix, quaternion, and etc.) exist to express the 3-D rotations. According to Euler’s rotation theorem[17], any representation of the 3-D rotation is equivalent to a rotation by an angle $\beta$ along a fixed Euler axis $\hat{e}$. Then the rotation vector, parameterizing a 3-D rotation in a 3-D Euclidean space, can be defined by $\mathbf{r} = \beta \hat{e}$. In this paper, we introduce the concept of angular trajectory $\gamma_a$, in which the rotation vector $\mathbf{r}$ can be regarded as the translational velocity of this angular trajectory. To sum up, a 6-D rigid body trajectory $\gamma \in \mathbb{R}^{6 \times N}$ is the combination of the 3-D angular trajectory $\gamma_a(t) = \int_0^N r(t) dt = [x_a(t), y_a(t), z_a(t)]^T$ and the 3-D point trajectory $\gamma_p(t) = \mathbf{p}(t) = [x_p(t), y_p(t), z_p(t)]^T$ as

$$\gamma(t) = [\gamma_p(t)^T, \gamma_a(t)^T]^T, \ t \in [1, N]$$  (1)

In the following, we use $\gamma(t)$ represents the vector at time step $t$, and $\gamma$ indicates the temporal trajectory sequence.

In this paper, the component of a trajectory indicates the data vectors that represent in one dimension. Especially, a 6-D rigid body motion trajectory contains six components $\{x_p, y_p, z_p, x_a, y_a, z_a\}$. It should be pointed out that the components of a trajectory will be different under rotational variations in 3-D space.

B. Challenges in Trajectory-based Motion Recognition

The main challenge in trajectory-based motion recognition is that raw trajectories always suffer various variations while being observed by different visual systems or performed by different individuals.

**Translation:** The translation in 3-D space is equivalent to add a constant position vector $\mathbf{p}_0 = [x_0, y_0, z_0]^T$ on the 3-D point trajectory as $V_{\text{trans}}(\mathbf{p}(t)) = \mathbf{p}(t) + \mathbf{p}_0$.

**Scaling:** It can be simplified as the zoom in and zoom out of the 3-D point trajectory, which is $V_{\text{scale}}(\mathbf{p}(t)) = \alpha \mathbf{p}(t)$, where $\alpha$ is a scalar. It should be pointed that the 3-D angular trajectories are not influenced under translation and scaling in 3-D space.

**Rotation:** Considering the vision capture system is rotated with a rotation matrix $\Gamma$ and we have $\Gamma \in SO(3)$, $\Gamma^T \Gamma = I$, and $\det(\Gamma) = 1$. Accordingly, the point trajectory and the angular trajectory will be simultaneously rotated. Accordingly, the rotated trajectory is:

$$V_{\text{rot}}(\gamma) = [(\Gamma \gamma_p)^T, (\Gamma \gamma_a)^T]^T$$

**Others variations:** In the real-world applications, the trajectories representing the same motion may exist various nonanalytical variations while being performed by different individuals, such as differences in shape, execution rate, trajectory length.

To solve the aforementioned challenges, a lot of invariant descriptions for 6-D rigid body motion trajectories have been proposed in recent years, including the ISA [12], EFS [13], and RRV [14] descriptors. However, these works share the similar framework that the kinematic-based local invariants of trajectories are calculated, which means that the output descriptor sequences are of the same dimension as the raw trajectories. Moreover, the local invariants are sensitive to local perturbations. Accordingly, the template matching method incorporating nearest neighbor classifier is used for trajectory-based motion recognition. Although they have achieved the satisfactory recognition performance in different applications, they are more or less time-consuming and space-consuming. To this end, this paper aims to propose an effective and efficient recognition framework by proposing a flexible statistical descriptor. The final output trajectory descriptors for different trajectories are of the same dimensions, which is beneficial for incorporating the computational efficient recognition method.

### III. Motion Trajectory Description

#### A. Trajectory Preprocessing

Firstly, raw trajectories are always contaminated by noise points, so Kalman smoothing is used to remain the essential shape of a trajectory as well as smooth the noisy points.

Next, the number of sampling frames $N_i$ of a trajectory $\gamma_i$ will be re-sized to a constant $M$ in an interpolation and resampling principle. Each 3-D point/angular trajectory will be firstly interpolated to $3 \times kN_i$ with cubic spline interpolation method. Then the output trajectory will be sampled $M$...
frames in a uniform manner. The angular trajectory share the same resampling frame indexes as point trajectory.

As mentioned above, the components of a trajectory are not invariant to rotation. At last, the Singular Value Decomposition (SVD) method is adopted to remove the rotational variations of the components. A rotated 6-D rigid body motion trajectory is \( \mathbf{\gamma} = [\mathbf{\Gamma}\mathbf{\gamma}_p, \mathbf{\Gamma}\mathbf{\gamma}_a]^{T} \), where \( \mathbf{\Gamma} \) is a rotation matrix. Let denote \( \mathbf{\gamma}_p \) be the center of the point trajectory. Then we have \( \mathbf{X} = \mathbf{X}_p - \mathbf{\gamma}_p \in \mathbb{R}^{3 \times M} \). According to the key idea of SVD method, \( \mathbf{X} \) can be decomposed as \( \text{svd}(\mathbf{X}) = \mathbf{U}\Sigma\mathbf{V}^{T} \). As is well known, \( \mathbf{X}^{T}\mathbf{X} = \Sigma\mathbf{V}^{T}\mathbf{V}\Sigma^{T} \). Therefore, the rotational variance of the 3-D point trajectory can be removed as \( \mathbf{\hat{\gamma}}_p = \mathbf{U}^{T}\mathbf{\gamma}_p \). Simultaneously, we also have \( \mathbf{\hat{\gamma}}_a = \mathbf{U}^{T}\mathbf{\gamma}_a \) so as to maintain the internal relations between the point trajectory and the angular trajectory.

### B. Multi-layer Self-similarity Matrices Representation

In the following, the Multi-layer Self-similarity Matrices for representing 6-D rigid body motion trajectories will be introduced, which consists of a set of similarity matrices calculated in three layers.

1) **Layer 1:** The first layer contains two standard Self-Similarity Matrix (SSM) [15], which calculates the pair-wise Euclidean distance between any two points within the 3-D point trajectory \( \mathbf{\gamma}_p \) or 3-D angular trajectory \( \mathbf{\gamma}_a \) itself. Taking the SSM of \( \mathbf{\gamma}_p \) as the example:

\[
S(\mathbf{\gamma}_p) = [s_{ij}]_{i,j=1,...,M} = \begin{bmatrix}
  s_{11} & \cdots & s_{1M} \\
  \vdots & \ddots & \vdots \\
  s_{M1} & \cdots & s_{MM}
\end{bmatrix}
\] (2)

where \( s_{ij} = ||\mathbf{\gamma}_p(i) - \mathbf{\gamma}_p(j)||_2 \) indicates \( l_2 \)-norm between two 3-D position vectors at frame \( i \) and \( j \). The MSM representation at layer 1 can be expressed as \( \mathcal{F}_{L1} = \{S(\mathbf{\gamma}_p), S(\mathbf{\gamma}_a)\} \). It should be pointed out that the min-max normalization is adopted on each similarity matrix to normalize the values in the matrix belong to [0, 1]. In other words, the similarity matrix can be viewed as a gray-scale image after mapping to [0, 255].

Compared with the existing descriptors, the SSM transforms the original 3-D trajectories into a 2-D image, which captures both local and global spatiotemporal information. Moreover, it has been shown that the SSM representation is invariant to rigid transformations [15]. However, the SSM representation still has two limitations: 1) insufficient descriptive ability; 2) motion patterns at the component level are not readily noticeable. To illustrate the aforementioned limitations, a toy example is given in Fig. 3. As can be observed, the corresponding SSMs for two obviously different trajectories \( \mathbf{\gamma}_p1 \) and \( \mathbf{\gamma}_p2 \) are similar. Moreover, the periodical patterns of \( \mathbf{\gamma}_p1 \) can not be captured in \( S(\mathbf{\gamma}_p1) \).

2) **Layer 2:** The second layer is composed of the Self-Component-Similarity Matrices (SCSM), which calculates the pair-wise distance within each component of a rigid body motion trajectory \( \mathbf{\gamma} \). Taking \( \mathbf{x}_p \) as the example, the SCSM of \( \mathbf{x}_p \) can be expressed as:

\[
S(\mathbf{x}_p) = [s_{ij}]_{i,j=1,...,M} \quad (3)
\]

where \( s_{ij} = ||\mathbf{x}_p(i) - \mathbf{x}_p(j)||_2 \) indicates the Euclidean distance between two points in component \( \mathbf{x}_p \) at time \( i \) and \( j \). Then the MSM representation for \( \mathbf{\gamma} \) at layer 2 is \( \mathcal{F}_{L2} = \{S(\mathbf{x}_p), S(\mathbf{y}_p), S(\mathbf{z}_p), S(\mathbf{x}_a), S(\mathbf{y}_a), S(\mathbf{z}_a)\} \). Compared with the similarities at the trajectory level, the self-component-similarities aim to capture more latent motion patterns at the component level, thus to improving the discriminability in the description. It is important to note that the rotational normalization is the key step before calculating the SCSMs.

3) **Layer 3:** Except for the self-component similarities, the cross-similarities between any two components additionally compose the third layer, which can be denoted as the Cross-Components-Similarity Matrix (CCSM). For instance, the CCSM of components \( \mathbf{x}_p \) and \( \mathbf{y}_p \) can be written as:

\[
C(\mathbf{x}_p, \mathbf{y}_p) = [c_{ij}]_{i,j=1,...,M} = \begin{bmatrix}
  c_{11} & \cdots & c_{1M} \\
  \vdots & \ddots & \vdots \\
  c_{M1} & \cdots & c_{MM}
\end{bmatrix}
\] (4)

where \( c_{ij} = ||\mathbf{x}_p(i) - \mathbf{y}_p(j)||_2 \) is the \( l_2 \)-norm between \( \mathbf{x}_p \) and \( \mathbf{y}_p \) components at frame \( i \) and \( j \). The entries with smaller values in the CCSM indicate higher correlations between two components and vice versa. The third layer can depict the internal synergies of two components. It should be noted that the CCSM is not the symmetrical matrix as SSM and SCSM. In specific, the third layer can be written as \( \mathcal{F}_{L3} = \{C(\mathbf{x}_p, \mathbf{y}_p), C(\mathbf{x}_p, \mathbf{z}_p), C(\mathbf{y}_p, \mathbf{z}_p), C(\mathbf{x}_a, \mathbf{y}_a), C(\mathbf{x}_a, \mathbf{z}_a), C(\mathbf{y}_a, \mathbf{z}_a)\} \).

4) **The MSM representation:** In conclusion, the MSM representation for a 6-D rigid body motion trajectory is \( \mathcal{F} = \{\mathcal{F}_{L1}, \mathcal{F}_{L2}, \mathcal{F}_{L3}\} \). The entry with the maximum value in each layer will be selected so as to normalize the values in the similarity matrices belong to [0, 1]. Compared to the existing kinematic-based descriptors only designed for 3-D or 6-D trajectories, the proposed MSM representation can be easily extended to high dimensional trajectories. To our knowledge,
this is the first work incorporating the self-similarities and cross-similarities at the component level within a trajectory. This simple mechanism can significantly improve the inter-class discriminability than the standard SSM representation.

C. Important Properties of MSM representation

The entries in the similarity matrices are the $l_2$-norm between two vectors, which means that the MSM representation is invariant to translation. The first layer in the MSM representation calculates the pair-wise Euclidean distance between the 3-D position vectors, which is invariant to rotation. In addition, the SVD-based method empowers the rotational invariance of components. Along this line, the self-component-similarities and cross-component-similarities are also invariant to rotation. To remove the scaling, the entries in the similarity matrices will be normalized to $[0,1]$ in a min-max criterion.

D. MSM-HOG trajectory descriptor

As the similarity matrices in the MSM representation can be viewed as gray-scale images. At last, extracting the local image descriptors from MSM can construct the final motion trajectory descriptor in a reduced dimensional space as the final input to the classifier. The Histogram of Oriented Gradients (HOG) is a gradient-based statistical representation, in which the gradient-based mechanism is advantageous in capturing edge or local shape information [16] and the statistical features are beneficial for analyzing the nonanalytical variations (speed, shape, length and etc.) of the intra-class trajectories. In practice, an image will be divided into some small connected cells, and the gradient values of each cell can be calculated. The next step of calculation is creating the cell histograms. Each pixel within a cell is voted for oriented histogram based on the gradient values in the first step. Finally, an output histogram $H \in \mathbb{R}^{1 \times K}$, concatenating all the cell histograms, is derived for representing the given image.

To sum up, the final trajectory descriptor $\mathcal{H} = [H_1, H_2, \cdots, H_i, \cdots]$ is to concatenate the histograms extracted from different similarity matrices in the MSM representation, which is denoted as the MSM-HOG descriptor in this paper.

IV. EXPERIMENTS

A. Motion Recognition Methods

The proposed MSM-HOG descriptors for different trajectory samples are of the same dimension, which are the final input to the final classifier. Recall that the previous works typically use template matching method for classification, it is time consuming due to the pair-wise warping and comparison, and it is also space consuming owing to the storage of all training descriptors in the classification stage. Motivated by this, we train a linear SVM classifier with the MSM-HOG descriptors for efficient multiclass trajectory-based recognition, in which only the pre-trained model parameters and support vectors are stored.

B. Experimental Implementation

1) Parameter settings: In the preprocessing step, the resampling frames is set to $M = 64$, which means that the similarity matrices in the MSM representation are of size $64 \times 64$. For HOG descriptor, the direction of the gradient is discretized into $\beta = 12$ “signed” bins. In addition, we experimentally select $16 \times 16$ pixels in each cell. By choosing these two parameters, the final MSM-HOG descriptor for each similarity matrix is $H_i \in \mathbb{R}^{1 \times 192}$. In this paper, the motion recognition experiments on the CharTrj and AUSLAN2 datasets are conducted so as to evaluate the effectiveness and efficiency of the proposed method. To illustrate the invariance of the proposed descriptor, the trajectory samples in two datasets are randomly scaled, translated and rotated.

2) Character trajectory (CharTrj) dataset: The CharTrj dataset [19] consists of 2,858 character samples and is divided into 20 classes. Each sample records a 3-D pen tip velocity sequence captured at 200 Hz. Given a random initial position, the 3-D character trajectory can be acquired from velocity sequence. In the CharTrj dataset, the angular trajectory $\gamma_a$ equals to $[0, 0, 0]^T$ at all time steps. Fig. 5(a) plots four trajectory samples in CharTrj dataset. To show the comparative results, this paper follows the existing works [20] to randomly choose half of the samples for training and the remaining samples for testing. The experiments are repeated for 100 times, and the average recognition results will be reported.
C. Experimental Results

1) Evaluation of the MSM-HOG descriptor: Firstly, the significance of three layers in the MSM representation are evaluated by means of trajectory-based motion recognition experiments. The average recognition results with different combinations of layers are listed in Table I. As can be observed, the proposed SCSSMs (Layer 2) and CCSMs (Layer 3) can provide superior recognition accuracies than the standard SSM (Layer 1). Moreover, the full MSM representation, which is the combination of three layers, achieves the best performance in two motion trajectory recognition tasks, which emphasizes that the full MSM-HOG descriptor is advantageous in capturing the motion patterns at the component level thus to improving the descriptive ability.

<table>
<thead>
<tr>
<th>Layers</th>
<th>Accuracy (%) CharTrj</th>
<th>Accuracy (%) AUSLAN2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layer1</td>
<td>96.24</td>
<td>90.69</td>
</tr>
<tr>
<td>Layer2</td>
<td>97.13</td>
<td>91.34</td>
</tr>
<tr>
<td>Layer3</td>
<td>97.01</td>
<td>91.42</td>
</tr>
<tr>
<td>Layer1+2</td>
<td>98.07</td>
<td>92.58</td>
</tr>
<tr>
<td>Layer1+3</td>
<td>98.25</td>
<td>93.03</td>
</tr>
<tr>
<td>Layer2+3</td>
<td>97.89</td>
<td>92.96</td>
</tr>
<tr>
<td>Layer1+2+3</td>
<td>98.66</td>
<td>93.31</td>
</tr>
</tbody>
</table>

2) Comparison results: Next, the comparison results with the existing descriptors and methods on two datasets are presented in Table II.

For the CharTrj dataset, we firstly compare the results provide by 3-D kinematic-based trajectory descriptors (SIG, AII, SRVF), which take advantageous of the temporal information. The comparison results by the existing 6-D rigid body motion trajectory descriptors (ISA, EFS, RRV) are also given, in which only the translational invariants are utilized in the description. It is important to note that the translational invariant part in the RRV descriptor is equivalent to the SRVF descriptor. The mean recognition accuracies by different methods on CharTrj dataset are reported in the Table II. As can be observed, the classification accuracy provided by the MSM-HOG descriptor with SVM classifier significantly outperforms other methods on CharTrj dataset. In addition, the result by SSM-HOG descriptor is also slightly better than those using kinematic-based descriptors. The highest recognition rate achieves 98.66% on CharTrj dataset while using MSM-HOG descriptor, which exceeds the second best (98.00%[20]) by 0.66%. As for the AUSLAN2 dataset, the recognition experiments are repeated 252 times and the average accuracies are given. Moreover, the comparison results by the ISA, EFS, RRV and SSM-HOG descriptors are reported. It should be noted that the results by other state-of-the-art works are not provided due to the existing works have different experimental settings. As listed in the Table II, the proposed MSM-HOG descriptor outperforms the existing descriptors in terms of...
recognition accuracies. The well-designed kinematics-based RRV descriptor with template matching method achieves the second best classification rate. As can be observed, the highest recognition accuracy result reaches 93.31%, which exceeds the second best result (92.46%) by 0.85%.

<table>
<thead>
<tr>
<th>Methods</th>
<th>CharTrj dataset</th>
<th>AUSLAN2 dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy(%)</td>
<td>Time cost per test sample(ms)</td>
</tr>
<tr>
<td>EFS(DTW+1-NN)</td>
<td>88.55</td>
<td>501</td>
</tr>
<tr>
<td>SIG(DTW+1-NN)</td>
<td>89.69</td>
<td>482</td>
</tr>
<tr>
<td>ISA(DTW+1-NN)</td>
<td>90.76</td>
<td>532</td>
</tr>
<tr>
<td>All(DTW+1-NN)</td>
<td>94.01</td>
<td>510</td>
</tr>
<tr>
<td>SRVF(DTW+1-NN)</td>
<td>95.22</td>
<td>467</td>
</tr>
<tr>
<td>RRV(DTW+1-NN)</td>
<td>95.22</td>
<td>467</td>
</tr>
<tr>
<td>SSM-HOG(SVM)</td>
<td>96.24</td>
<td>18</td>
</tr>
<tr>
<td>Grabocka et al. [20]</td>
<td>98.00</td>
<td>-</td>
</tr>
<tr>
<td>MSM-HOG(SVM)</td>
<td>98.66</td>
<td>43</td>
</tr>
</tbody>
</table>

V. CONCLUSION

This paper proposes a novel framework for 6-D rigid body motion trajectory recognition by calculating the multilayer self-similarity matrices (MSM) at the trajectory level and component level, respectively. The main contribution of this paper is that we explore the self-similarities and cross-similarities at the component level in the MSM representation, which enables us to capture more discriminative patterns thus to improving the descriptive power. Next, the local HOG features extracting from the MSM representation are concatenated as the final trajectory descriptor, MSM-HOG. Finally, we train a linear SVM classifier for achieving efficient multiclass recognition task. The effectiveness of the our approach is evaluated by means of trajectory-based motion recognition experiments, and it can be concluded that our approach can achieve satisfactory result in both recognition accuracy and computational efficiency.

REFERENCES