A Shape Descriptor for Real Time 3D Foot Pose Estimation

HO-GEUN SONG, HA-SUNG KOO
Department of Computer & Information Engineering
Hanseo University
360 Daegok-ri Haemi-myun Seosan-si Chungcheongnam-do
KOREA
hksong@hanseo.ac.kr, hskoo@hanseo.ac.kr

Abstract: This paper proposes the effective shape descriptor for 3D foot pose estimation. To reduce processing time, silhouette-based foot image database is built and meta information which involves the 3D pose of the foot is appended to the database. And we proposed a modified Centroid Contour Distance whose size of the feature space is small and performance of pose estimation is better than the others. In order to analyze performance of the descriptor, we evaluate time and spatial complexity with retrieval accuracy, and then compare with the previous methods. Experimental results show that the proposed descriptor is more effective than the previous methods on feature extraction time and pose estimation accuracy.

Key-Words: Shape descriptor, Pose estimation, Image retrieval, 3D foot pose, and Silhouette based database

1 Introduction

An application for human body tracking needs motion analysis technology, and is used in the field of special effects in cinema, game, animation and perceptual interface, etc. [1] The motion tracking technology should be robust within the given environment conditions and fast for real time processing.

Shape descriptors for many applications such as human motion tracking, human pose estimation, object recognition and augmented reality are active research themes in the content-based image retrieval systems [1][2][3]. The related researches have recently studied to estimate 3D object pose from the 2.5D database when a 2D shape is given to the systems. The shape descriptor for 2D shape information of an object is indexed for the database with object pose information to reduce the processing time especially for real-time system[4]. And this allows us to cancel out difficulties for the previous methods which require time consuming initialization, tracking and geometrical computation. However those systems require not only fast processing scheme for real time process but also accurate decision capacity for real applications.

A shape is an important clue to represent projected images of a 3D object. However, the projected shape depends on occlusion, noise, and other environmental changes [5][6][7][8] as well as geometrical distortions. For that reason, the shape descriptor for 3D application should be chosen to hold robustness for those geometrical transformations and to require short processing time when the object shape is retrieved from database.

In general, shape descriptor is divided into two categories, e.g. region-based and contour-based approach. The region-based approach is available for both all single connected regions and multiple regions, and whose retrieval time is fast with noise robustness. However, they show poor performance for some distorted object. Those approaches involve Hu invariant moments, Zernike moments, axis projection, and binary sequences [8][9][10][11] etc. The contour-based method represents an external closed shape of 2D object. The approaches are robust for some distortion, but sensitive to noise. The methods include Chain-code, Centroid contour distances (CCD), Fourier descriptor, B-spline, and Shape contexts[12][13][14][15][16] etc.

Therefore this paper proposes the effective shape descriptor for 3D foot pose estimation. To reduce processing time, silhouette-based foot image database is built and meta information which involves the 3D pose of the foot is appended to the database. And we proposed a modified Centroid Contour Distance whose size of the feature space is small and performance of pose estimation is better than the others. In order to analyze performance of the descriptor, we evaluate time and spatial complexity with retrieval accuracy, and then compare with the previous methods. Experimental results show that the proposed descriptor is more effective than the previous

methods on feature extraction time and pose estimation accuracy.

2 Proposed method.

Proposed shape descriptor for 3D foot pose estimation has three groups of feature value whose size of the feature space is small and performance of pose estimation is better than the previous methods.

2.1 Centroid Contour Distance

When contour points \( P(p_1, p_2, p_3, \ldots, p_n) \) of an object and the coordinate set of the \( P, f(x, y) \), are given, centroid of the object \( C(X_c, Y_c) \) is defined as:

\[
X_c = \sum_x \sum_y f(x, y)x / \sum_x \sum_y f(x, y) \\
Y_c = \sum_x \sum_y f(x, y)y / \sum_x \sum_y f(x, y)
\]

(1)

(2)

If we consider \( r \) as distance between the centroid \( C \) and arbitrary contour point \( p' \) at angle \( \alpha \), we can define the set of distance \( R(r_1, r_2, \ldots, r_{360}) \) from each \( \alpha \) as shown in Fig. 1. Then the set \( R \) is defined as Centroid Contour Distance (CCD).

Fig. 1 Centroid contour distance

2.3 Angular segmentation for refinement.

As we described above, the CCD has 360 distance feature values for an object. However, we can reduce the number of the features because of its discrete, local and redundant characteristics in digital images.

When we divide entire 360 angular region \( A \) by arbitrary angle \( \alpha' \), number of the angular bin is \( m = 360/\alpha' \).

Therefore, angular bin set \( A \) is defined as:

\[
A = \{ A_1, A_2, A_3, \ldots, A^m \}
\]

(3)

where \( m \) is the number of the bins. (Fig. 2)

Fig. 2 Angular segmentation, when \( \alpha'=45^\circ \)

2.4 Adjacent Area Distance

On the basis of above definition, we need to describe an object shape as its relational features effectively. Because 2D shape of foot has no prominent geometrical feature point, e.g. corner points.

First of all, as shown in Fig. 3, we accumulate difference of corresponding distance \( r \) for each adjacent angular bin. It called as adjacent area distance (AAD) and defined as:

\[
AAD_i = \sum_{l=0}^{q} \| A'_j - A'^{next}_i \|_2 \quad i = 0, 1, \ldots, p
\]

(4)

where \( p \) represents the number of angular bins and \( q \) describes grouping size of the bins, angle \( \alpha' \).

The AAD is one of statistical features that stand for relative variance of the CCD features within the local area. Therefore, if an object has regular shape with small relative variance, the AAD shows small value. Otherwise, e.g. irregular shape, the AAD shows large one.

Fig. 3 Adjacent Area Difference

2.5 Symmetric Area Distance

In spite of advantage of the AAD, it does not represent global feature of an object. In other words, we need to discriminate between variances with large and small feature value.

Therefore, we accumulate difference of corresponding distance \( r \) for each symmetric angular bin as shown in Fig. 4. And it called symmetric area distance and defined as:

\[
SAD_i = \sum_{l=0}^{q} \| A'_j - A'^{symmetric}_i \|_2 + \| A'_j - A'^{symmetric}_i \|_2 \quad i = 0, 1, \ldots, p'
\]

(4)

where \( p' \) is a half of number of angular bins, \( q \) describes grouping size of the bins, and adding number 1 in denominator is a minimum constant to avoid dividing by zero, respectively.

Fig. 4 Symmetric Area Difference

2.6 Self-Area Accumulation

Both AAD and SAD are relational feature values to describe the shape of an object. However, the shape of an object should be represented by the feature itself.
Therefore, distance \( r \) for each angular bin is accumulated and defined as:

\[
SAA_i = \sum_{j=0}^{p} A'_j, \quad i = 0, 1, \ldots, p
\]  

(5)

where \( p \) denotes the number of angular bins and \( q \) represents the grouping size, as stated above.

In this paper, we set the grouping size of the bin to 20 for AAD and SAD, and define the size as 10 for SAA. Then number of AAD, SAD and SAA are 18, 9 and 36, respectively. As a result, the proposed shape descriptor has total 63 feature values for an object while the CCD has 360 values.

3. Experimental Conditions and the Results

In order to analyze performance of the proposed descriptor, we evaluate time and spatial complexity with retrieval accuracy, and then compare with the previous methods.

3.1 Experimental Environment

Our experiments are performed on IBM personal computer with Pentium 4 Dual core 3.2 Ghz CPU and 2GB memory. And we write a program with Visual Studio 2005.

Total flow of our experiment is as followed. First, when a query foot image is given, feature distances between the query and each database image are calculated. Second, we select a database image whose feature values are approximately matched with the query image among the candidate images. Third, 3D pose of query foot image is estimated by using the meta information in the database.

3.2 Database for Foot Pose Estimation

To reduce the processing time, silhouette-based foot image database is built and additional information is appended to the database. The additional information includes not only 3D pose of the foot but also pre-estimated proposed feature values.

For this, we make 3D foot model as in Fig. 5 and create total 13,500 silhouette image database as shown in Fig. 6. Each rotational scope of 3D axis and the unit angle for the rotation when we make the database are explained in Table 1.

### Table 1. Scope of axis rotation

<table>
<thead>
<tr>
<th>Axis</th>
<th>Classified number</th>
<th>Scope of axis rotation (unit angle)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>20</td>
<td>(-40^\circ \sim +40^\circ) (every 4(^\circ))</td>
</tr>
<tr>
<td>Y</td>
<td>45</td>
<td>(0^\circ \sim 180^\circ) (every 4(^\circ))</td>
</tr>
<tr>
<td>Z</td>
<td>15</td>
<td>(-30^\circ \sim +30^\circ) (every 4(^\circ))</td>
</tr>
</tbody>
</table>

3.3 Spatial and Time Complexity

In order to analyze the performance of our method, we choose and compare with typical 6 shape descriptors, namely chain code, Hu moments, binary sequences, axis projection and shape context.

In our experiment, spatial complexity is defined as total number of the shape feature. This complexity is proportional to number of operation when the system estimates the similarity between query image and model image. Therefore it has a big impact on real time processing. Table 2 shows the spatial complexities for each shape descriptor. As mentioned above, proposed descriptor has only 63 feature values. It is about 20% of the average spatial complexity.

### Table 2. Spatial complexity

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Feature number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chain Code</td>
<td>360</td>
</tr>
<tr>
<td>Hu Invariant Moments</td>
<td>7</td>
</tr>
<tr>
<td>Binary Sequences</td>
<td>256</td>
</tr>
<tr>
<td>Axis Projection</td>
<td>256</td>
</tr>
<tr>
<td>Shape Contexts</td>
<td>384</td>
</tr>
<tr>
<td>Proposed method</td>
<td>63</td>
</tr>
</tbody>
</table>

On the other hand, time complexity is defined as overall time for extracting the shape feature values. This complexity also includes time interval for pre-processing of the features.

Table 3 reveals the time complexities for each descriptor. Our method takes about 0.00296 seconds to extract feature values. And it is about 86% of the average time complexity.

The results show that the proposed descriptor is more effective than the previous methods on feature extraction time.

### Table 3. Time complexity
3.3 Retrieval Accuracy

Retrieval accuracy is estimated as followed.
First of all, we select 100 query images from database arbitrarily, and compute feature distance between the query and the 13,500 database images, respectively. The feature distance is given as:

\[
\text{Dist} (F_n, F_s^k) = \sum_{i=0}^{n} |F_i - F_s^k|
\]  

(6)

where \(F_n\) and \(F_s^k\) are feature vectors for query image and the \(k_{th}\) database image, respectively, and \(n\) is the length of a shape feature vector. Second, 10 candidate database images are chosen in ascending order of the distance for all the query images. Third, average pose estimation errors of the candidate images are recorded for each 3D axis. Table 4 shows the estimation errors for each and every descriptors and 3D axes. In Table 4, Rank(2) denotes average error of pose estimation with two candidate images in ascending order of feature distance, and Rank(5) and Rank(10) are also defined in the same way of Rank(2). As a result, if we estimate retrieval accuracy as a reciprocal of the average error, then the proposed method improves about 36% of the average retrieval accuracy in comparison with the other methods.

The experimental results show that the proposed descriptor is more effective than the previous methods on feature extraction time and pose estimation accuracy.

### Table 4. Retrieval accuracy

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Extraction time</th>
<th>average error of pose estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Rank(2)</td>
</tr>
<tr>
<td>Chain Code</td>
<td>0.000861</td>
<td>11.5</td>
</tr>
<tr>
<td>Hu Invariant Moments</td>
<td>0.008119</td>
<td>17</td>
</tr>
<tr>
<td>Binary Sequences</td>
<td>0.001504</td>
<td>20.5</td>
</tr>
<tr>
<td>Axis Projection</td>
<td>0.001815</td>
<td>5.5</td>
</tr>
<tr>
<td>Shape Contexts</td>
<td>0.036798</td>
<td>10.8</td>
</tr>
<tr>
<td>Proposed method</td>
<td>0.002960</td>
<td>12.7</td>
</tr>
</tbody>
</table>

4. Conclusions

Shape descriptor for 3D application should be chosen to hold robustness for geometrical transformations, and to require short processing time. To reduce processing time, total 13,500 silhouette-based foot image database is built and meta information which involves the 3D pose of the foot is appended to the database. And we propose the effective shape descriptor for 3D foot pose estimation.

In order to analyze performance of the proposed descriptor, we evaluate time and spatial complexity with retrieval accuracy, and then compare with the previous methods. According to the result of our experiment, proposed descriptor has only 63 feature values. It is about 20% of the average spatial complexity. Furthermore, our method takes about 0.00296 seconds to extract feature values. It is about 86% of the average time complexity. The results show that the proposed descriptor is more effective than the previous methods on feature extraction time. Finally, if we estimate retrieval accuracy as a reciprocal of the average error, then the proposed method improves about 36% of the average retrieval accuracy in comparison with the other methods.

The experimental results show that the proposed descriptor is more effective than the previous methods on feature extraction time and pose estimation accuracy.

References


