Human Action Recognition Using Silhouette Histogram

Chaur-Heh Hsieh, *Ping S. Huang, and Ming-Da Tang

Department of Computer and Communication Engineering
Ming Chuan University
Taiyuan 333, Taiwan, ROC
*Department of Electronic Engineering
Ming Chuan University
Taiyuan 333, Taiwan, ROC

{hsiehch, pshuang}@mail.mcu.edu.tw, s7166076@ss24.mcu.edu.tw

Abstract
This paper presents a method for human action recognition using silhouette histogram. The human silhouette is obtained by using the background subtraction method and then mapped into three different polar coordinate systems that characterize three parts of a human figure respectively. Each polar coordinate system is quantized by partitioning it into several cells with different radii and angles. The evaluation applied at the Weizmann dataset shows that our method can achieve better performance than the existing silhouette-based system.

Keywords: silhouette, polar coordinate system, classifier.

1 Introduction
Human action recognition from videos is an important research area of computer vision. It can be applied to a variety of application domains, such as video surveillance, video retrieval and human-computer interaction systems. In the past, many human action recognition methods have been proposed and they are mainly divided into two kinds of approaches: shape-based and flow-based. The first method utilizes human silhouettes as features which are commonly extracted by background subtraction [1-5]. The second method adopts motion information from human movement such as optical flow [6-8]. Classification techniques that are usually used for feature matching consist of Nearest-Neighbor (N-N) [1, 6], Support Vector Machines (SVM) [9] and Hidden Markov Models (HMM) [2, 9, 10].

The silhouette-based method aims to recognize actions by characterizing the shapes of the actor’s silhouette through space-time. The method is very popular since it is computationally efficient and robust to variations in clothing and lighting [11]. Nevertheless, it often suffers from the artifacts such as shadows, which will reduce the recognition accuracy. The flow-based method computes the optical field between adjacent frames and uses that as features for action recognition. This is suitable for recognizing small objects [6]. However, it is computationally expensive, and the optical field is a very coarse feature and thus different actions may exhibit similar flows over short periods of time [11].


Hsiao et al. [1] propose the temporal-state shape context (TSSC) that can capture local space-time shape characteristics effectively. This can eliminate time warping effects in the local feature methods by segmenting the video sequence into several temporal states defined by soft fuzzy intervals. The distance between two action clips is defined as the summation of shape context distances in all frames. Besides estimating the distance measure between two actions, TSSC method can also provide spatial consistency information.

Motivated by the TSSC approach, this paper presents a human action recognition method by using three silhouette histograms. The human silhouette extracted by background subtraction is mapped into three different polar coordinate systems that characterize three parts of a human figure respectively. Each polar coordinate system is quantized by partitioning it into several cells with different radii and angles.

This paper is organized as follows. Section 2 introduces the proposed method. Experimental results and discussion are explained in Section 3. Finally conclusions are given in Section 4.

2 Proposed Method

2.1 Overview of Proposed System
The proposed system includes four main processes as shown in Figure 1. First, the human silhouette is extracted from the input video by background subtraction method [11]. Then, the extracted silhouette is mapped into three polar coordinate systems that characterize three parts of a human figure respectively. The largest circle covers the motion of the human body and the other two circles are to include the effect arms and legs have on the human action/silhouette. That is why two of the centres are between the shoulders and between the hips, respectively. Each polar coordinate system is quantized by partitioning...
it into several cells with different radii and angles. By counting the number of pixels fallen into each cell from the silhouette at a particular frame, the silhouette histogram of the frame can be obtained. By collecting a sequence of silhouette histograms, a video clip is thus generated and used to describe the human action. Based on the silhouette histogram descriptor, an action classifier is trained and then used to recognize the action of an input video clip.

2.2 Silhouette Extraction
We assume video sequences are captured from a still camera. Thus, background subtraction is applied to obtain the human silhouette from each frame. The well-known Gaussian Mixture Model (GMM) [11] is used to update the background. The method of background subtraction is given by

\[ O_{i}(x, y) = \begin{cases} 1, & |F_{i}(x, y) - B_{i}(x, y)| > T \\ 0, & \text{otherwise} \end{cases} \]  

where \( B_{i}(x, y) \) is the background image, \( F_{i}(x, y) \) is the input frame and \( T \) is the threshold value. Background refresh is achieved by

\[ B_{i}(x, y) = (1 - \alpha) \cdot B_{i}(x, y) + \alpha \cdot F_{i}(x, y) \]  

where \( 1 \leq \alpha \leq 0 \).

2.3 Polar Transform
To be able to effectively describe the human shape, the Cartesian coordinate system is transformed into the polar coordinate system through the following equations:

\[ r_{i} = \sqrt{(x_{i} - x_{c})^2 + (y_{i} - y_{c})^2} \]  

\[ \theta_{i} = \tan^{-1}\left(\frac{y_{i} - y_{c}}{x_{i} - x_{c}}\right) \]  

where \( (x_{i}, y_{i}) \) is the coordinate of silhouette pixels in the Cartesian coordinate system. \( (r_{i}, \theta_{i}) \) is the radius and the angle in the polar coordinate system. \( (x_{c}, y_{c}) \) is the centre of the silhouette. The centre of the silhouette can be calculated by

\[ x_{c} = \frac{1}{N} \sum_{i=1}^{N} x_{i} \]  

\[ y_{c} = \frac{1}{N} \sum_{i=1}^{N} y_{i} \]  

where \( N \) is the total number of pixels.

The existing approaches [1, 4, 5] often use a single polar coordinate system to describe the human posture. However, our investigation indicates that the single coordinate is not enough to discriminate different postures with small difference. In this work, we design a method which contains three polar coordinate systems (three circles) defined as:

- C1: Circle that encloses the whole human body.
- C2: Circle that encloses the upper part of a body.
- C3: Circle that encloses the lower part of a body.

To verify the effectiveness of the proposed three-circle method, it is applied to describe two different actions in the Weizmann dataset: wave1 (one hand waving) and wave2 (two hands waving), as shown in Figure 2. Weizmann dataset is widely used by related research papers. The silhouette histograms obtained by C1 and C2 are shown in Figure 3 and Figure 4, respectively. We can see that two histograms from C1 are very similar such that the discriminability of action types is poor. On the contrary, superior discriminability is demonstrated by two histograms from C2.

2.4 Histogram Computation
To generate the histogram of each polar coordinate system, we partition each polar coordinate space into K cells by uniformly dividing each radius into \( n \) parts, and angles into \( m \) orientations such that \( K = m \times n \). Therefore, K-bins are obtained for each histogram by

\[ H(k) = \# \left\{ P_{i,j} : P_{i,j} \in cell_{k} \right\} \]  

where the ranges of \( \theta_{i} \) and \( r_{i} \) are defined by

\[ \frac{2\pi i}{m} \leq \theta_{i} \leq \frac{2\pi (i+1)}{m}, \quad i = 0, 1, \ldots, m - 1 \]  

and

\[ \frac{j r_{\max}}{n} \leq r_{i} \leq \frac{j+1 r_{\max}}{n}, \quad j = 0, 1, \ldots, n - 1. \]  

(\( r_{\max} \) is the radius of circle C1, C2, or C3.)
Figure 4: Silhouette histograms obtained by C2 for two waving actions.

Figure 5: Human silhouettes of different actions and each is mapped onto three polar coordinate systems.

Generally, the recognition accuracy is increased with the adding of the number of cells. However, it would also raise the time complexity. In our experiments, we chose m=8 and n=3, and a histogram with 24-bins is obtained for each polar coordinate system. The procedures for calculating the silhouette histogram can be organized into the following steps. First, we compute the centre of human silhouette and divide the silhouette into an upper part and a lower part according to the centre position. Then the centres of upper silhouette and lower silhouette are individually computed. Those three centre positions are taken as origins for the respective polar coordinate systems. Second, we compute the heights of all human silhouettes, which are used to calculate the radius of C1. The radius of C2 or C3 is half of C1 radius. The third step is to apply Equation (6) to compute three histograms separately for each human silhouette.

Figure 5 shows 3 frames of different actions and each is mapped onto three polar coordinate systems. Each histogram count is divided by the total number of pixels fallen into that coordinate space. The normalization is used to avoid the bias of different silhouettes. Finally, the normalized histograms are concatenated into a feature vector, which is used to describe the human posture at a particular time instant.

3 Experimental Results

Our algorithm is evaluated with Weizmann dataset that includes nine persons performing ten actions. Figure 6 shows 8 sample frames from this dataset with different actions. The result is compared to that of the TSSC method [1]. The video format is 180x144 with the frame rate of 25 fps. We adopt the nearest neighbor classifier to identify different action types. The leave-one-out [13] evaluation method is used, which means one clip of a video is used for testing, and the remaining clips of all videos are included into the training set. This process is repeated until all video clips are tested.

The video clip is a fixed-length video data, say ten frames. The video clips are generated by dividing a complete video into a designate number of clips. Neighbouring clips can be overlapped with a fixed length, called the overlapped length. In our experiments, we set the clip length as 10 frames, and the overlapped length as 5 frames. Totally, 961 clips are generated from 90 videos. Therefore, the feature dimension for each video clip which contains 10 frames is 720 (24x3x10). To improve the computation efficiency of classification, the principal component analysis (PCA) is adopted to reduce the dimension of a feature vector. The recognition performance is evaluated by calculating confusion matrices shown in Table 1 and Table 2, which are generated from the TSSC method and our proposed method respectively. The results are obtained with a clip length of 10 and an overlapped length of 5.

Table 1: Confusion matrix of TSSC Method (N-N)
It can be seen that, for TSSC method, 45 clips are identified incorrectly, which achieves a recognition rate of 95.31%. However, our method can accomplish a higher recognition accuracy of 98.33%. Notice that the blend and skip actions have caused more errors for both methods.

Also, we have investigated the effects of overlapped length and clip length and the results are shown in Table 3 and Table 4, respectively. Table 3 indicates that the more the frames are overlapped, the better the recognition accuracy for both methods can achieve. The same conclusion is obtained for the clip length. The results listed in two tables show that our proposed method can achieve better recognition accuracy under various conditions of clip length and overlapped length.

<table>
<thead>
<tr>
<th>Overlapped Frames</th>
<th>TSSC</th>
<th>Proposed method</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>97.96%</td>
<td>98.91%</td>
</tr>
<tr>
<td>5</td>
<td>95.31%</td>
<td>98.33%</td>
</tr>
<tr>
<td>3</td>
<td>92.90%</td>
<td>96.59%</td>
</tr>
</tbody>
</table>

Table 3: Recognition rates under different number of overlapped frames

<table>
<thead>
<tr>
<th>Clip length</th>
<th>TSSC</th>
<th>Proposed method</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>95.31%</td>
<td>98.33%</td>
</tr>
<tr>
<td>15</td>
<td>97.35%</td>
<td>99.08%</td>
</tr>
<tr>
<td>20</td>
<td>98.59%</td>
<td>99.48%</td>
</tr>
</tbody>
</table>

Table 4: Recognition rates under different clip lengths

We argue that the leave-one-out test on clips is not very objective, because the parts of a testing video clip are included in the training set. Here, we design a test method called leave-n-out by taking one video for test and the remaining videos are used for training. Then n clips are extracted from the test video. Each video clip is evaluated separately, and the majority voting scheme is then applied to get the final decision. Table 5 is the TSSC result by using the leave-n-out method and Table 6 shows the result of our method. The results indicate that for TSSC the recognition accuracy is degraded significantly, while our method can still keep superior performance.

4 Conclusions

In this paper, we have developed an action recognition method which employs three polar coordinate systems to characterize different parts of the human posture. The human silhouette is mapped into a feature vector which contains three histograms. Experimental results show that our method can obtain superior performance than the TSSC method. Using the Weizmann dataset, the proposed method achieves a recognition rate of 98.33%. However, like the other silhouette-based approaches, the proposed method requires the extraction of a complete silhouette. Although the robustness to incomplete silhouettes has been raised by our method, the problem has not been completely solved. Furthermore, since the experiments are currently done within a well-controlled environment, more realistic outdoor scenes need to be tested. They will all be investigated in the future.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments and suggestions. This research is supported in part by the National Science Council, Taiwan under the grant NSC 99-2632-E-130-001-MY3.

5 References


