Human-human interaction recognition has attracted increasing attention in recent years due to its wide applications in computer vision fields. However, it still remains a challenge due to mutual occlusion, various subject appearance or body size and complex context. In this paper, a novel feature descriptor based on spatial relationship and semantic motion trend similarity between body parts is proposed for human-human interaction recognition.

2. Feature Extraction

The moving direction of each joint is firstly quantified into several semantic words and then the motion trend similarity is captured by histogram intersection.

2.1 Motion trend feature

- Frame direction: $v_i = [x, y, z], x = \sum_{j=1}^{m} x_j, y = \sum_{j=1}^{m} y_j, z = \sum_{j=1}^{m} z_j$
- Soft direction encoding depending on cosine similarity and displacement:
  
  $\text{cos}\theta_j(t) = \frac{v_i \cdot v_j}{\|v_i\| \|v_j\|}, j \in [1, m]$
  
  $\text{bin}_i_{\text{fract}} = \text{bin}_i_{\text{fract}} + \text{Dis}(t) \times \text{cos}\theta_j(t)$
  
  $\text{bin}_i_{\text{angular}} = \text{bin}_i_{\text{angular}} + \text{Dis}(t) \times \text{cos}\theta_j(t)$

2.2 Motion trend similarity

$$\text{Intersection}(b^m_i, b^m_j) = \min\{|b^m_i|, |b^m_j|\}$$

$$H = \sum_{p=1}^{m} \sum_{q=1}^{m} \text{Intersection}(b^m_p, b^m_q)$$

$$K_{\text{type}} = \sum_{p=1}^{m} K(H(p), H(q))$$

Where,

$$K(H(i), H(j)) = \sum_{k=1}^{m} \text{Intersection}(b^m_i, b^m_j) = H(i) \cdot H(j)$$

Table 1: Recognition Accuracy (%) on SBU dataset.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed Method</td>
<td>Test One</td>
<td>Test Two</td>
<td>Cross Subjects Test</td>
<td></td>
</tr>
<tr>
<td></td>
<td>90.58</td>
<td>90.28</td>
<td>92.50</td>
<td>91.12</td>
</tr>
<tr>
<td></td>
<td>80.3</td>
<td>89.4</td>
<td>86.9</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 1 Confusion matrixes

Fig. 2 Comparison between categories