Introduction

• We present a flexible method for single frame hand gesture recognition by fusing information from color and depth images.
• Depth makes it possible to obtain a reliable bounding box of the hand regardless of light and distance changes.
• Our method extracts common patch-level features, and fuses them by means of kernel descriptors.

Efficient Match Kernels for SVM

The match kernel of two images $I_i$ and $I_j$ for kernel SVM can be written as

$$K(X_i, X_j) = \sum_{a \in X_i} \sum_{b \in X_j} k(a, b),$$

where $k(a, b) = \phi(a)^T \phi(b)$.

The infinite-dimensional kernel vector $\phi(a)$ is approximated by a D-dimensional vector $\phi(a) = Hv_a$.

We construct $H$ by extracting the visual words of the patch, hence

$$k(a, b) = \psi(a)^T \psi(b) = (Hv_a)^T Hv_b = (H^T \phi(a)) \cdot (H^T H)^{-1} \cdot (H^T \phi(b)) = k_z(a)^T \cdot K_{zz} \cdot k_z(b).$$

Color and depth descriptors are concatenated into an image-level kernel feature vector.

Experimental Result

We have evaluated our method on two datasets: ASL FingerSpelling Dataset [10] and our own dataset.

• FingerSpelling dataset: 500 color images for each of 5 users are obtained for each sign.
• Our dataset: It consists of 24 static signs of 5 users at 5 different distances.

<table>
<thead>
<tr>
<th>Method</th>
<th>#training samples</th>
<th>Overall Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pugeault [10]</td>
<td>1250</td>
<td>75 %</td>
</tr>
<tr>
<td>Our Approach (NP)</td>
<td>40</td>
<td>77.39 ± 0.13 %</td>
</tr>
<tr>
<td>Our Approach (IP)</td>
<td>40</td>
<td>88.94 ± 0.39 %</td>
</tr>
</tbody>
</table>

Comparison with [10]

<table>
<thead>
<tr>
<th>Method</th>
<th>#training samples</th>
<th>Overall Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pugeault [10]</td>
<td>1250</td>
<td>75 %</td>
</tr>
<tr>
<td>Our Approach (NP)</td>
<td>40</td>
<td>77.39 ± 0.13 %</td>
</tr>
<tr>
<td>Our Approach (IP)</td>
<td>40</td>
<td>88.94 ± 0.39 %</td>
</tr>
</tbody>
</table>

Take-home Message

1. Hand can be easily tracked by depth camera, i.e., we can easily obtain a reliable bounding box of the hand.
2. Patch-based approaches can be used here, so we can achieve high accuracy with efficient algorithm.
3. The method works in a normal indoor setting (0.6m~2m).

Contact Us

Xiaolong ZHU: xzlzhu@cs.hku.hk
Dr. Kwan-Yee K. Wong: kykwong@cs.hku.hk