Static Hand Gesture Recognition With
*Spatial LOH Distribution Model*

Hanning Zhou, Dennis J. Lin and Thomas S. Huang

Image Formation and Processing Group
University of Illinois at Urbana-Champaign
Motivation & Objective

- Hand gestures are a natural way of communicating
  - Does not require the user to hold an input device
- Want to recognize a (relatively small) predetermined set of static hand postures
  - Needs to be fast and robust to be useful
Our Approach

- Bottom up
- Color histogram to identify the hand
  \[ L(u, v) = \frac{p(H(u,v)|skin)}{p(H(u,v)|nonskin)} \]
- Spatial local orientation histogram distribution to model a particular hand posture
- Retrieve using locality sensitive hashing
Local Orientation Histogram

- Based on SIFT [Lowe03]
- Divide the 24 x 24 window into 4 x 4 = 16 blocks
- Collect 8 bin weighted orientation histogram in each 6 pixel x 6 pixel block
- Descriptor 16 x 8 = 128 dim vector

Figure from the preprint of [Lowe03]
Spatial LOH

- Hand does not have texture
  - Difference of Gaussian pyramids do not work
- Raster-scan 24 x 24 subwindows within the hand region, overlapping by 6 pixels
- Augment each histogram with \((x,y)\) position relative to the centroid of the hand to give 130 dimensional spatial LOH feature
Spatial LOH Distribution

- Combine all the Spatial LOH features from all training images of the same hand posture
  - Posture is a particular hand configuration at a particular view angle
- Perform $k$-means clustering
  - We took $k = 30$
Clustering Results
Matching

\[
\Theta_{ex}(k) = \sum_{i=1}^{n} \sum_{j=1}^{30} \frac{1}{\eta(s_i, m_{j,k})}
\]

- Extract all the subwindows from the input image and compute the augmented LOH feature \(s_i\)
- Compare each window against each mean feature \(m_{j,k}\) of posture \(k\)
LSH to Speed Retrieval

- Use locality sensitive hashing [Gionis99] to find an approximate nearest neighbor set $\Gamma$

$$\Theta_{apr}(k) = \sum_{i=1}^{n} \sum_{m_{j,l} \in \Gamma} \delta(l - k) \frac{1}{\eta(s_i, m_{j,l})}$$

- Reduce the number of distance evaluation by a factor of 8~10
Locality Sensitive Hashing

- Create several hash tables by thresholding random dimensions at random thresholds
- To find neighbors, look up the query point in all the hash tables and take the union of the result
- Sub-linear lookup
Experimental Setup

- Two training sets: 80 images each
  - 8 different gestures: customary Chinese sign language for digit 1 ~ 8
  - Two examples of five different views

- Test sets 1 and 2:
  - Hand of two different subjects
  - Same setting as training set

- Test set 3:
  - Hand of the first subject
  - Different settings (taken 2 years ago)
Sample Images

Test Set 1

Test Set 2

Test Set 3
Results (1)

- Some example of good retrieval results
Some incorrect recognition results
## Results (3)

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Rank 1</th>
<th>Rank 1-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Set 1</td>
<td>69%</td>
<td>91%</td>
</tr>
<tr>
<td>Test Set 2</td>
<td>85%</td>
<td>93%</td>
</tr>
<tr>
<td>Test Set 3</td>
<td>65%</td>
<td>74%</td>
</tr>
</tbody>
</table>
Comments

- Pretty fast
  - 1/5 sec per frame on an Intel 1.3G laptop
- Could be more robust
  - Need better model of the distribution
  - Need more training data
    - Some way to (semi-)automatically acquire labeled data
Summary

- Bottom up approach
- Spatial local orientation histogram distribution
  - Purely local feature
- Locality sensitive hashing for retrieval
- 65 – 85% accuracy in 1/5 second