Grasp Type Revisited: A Modern Perspective on A Classical Feature for Vision
Yezhou Yang1, Cornelia Fermüller1, Yi Li2, Yiannis Aloimonos1
1Computer Vision Lab, University of Maryland, College Park, 2 NICTA and ANU

Why Grasp Type?
The grasp type contains fine-grain information about human action, and the recognition of grasp type is essential for a more detailed analysis of human action. It can help an intelligent system predict human action intention and is an essential component in the characterization of human actions of manipulation.

Motivation

Figure 1: (a) Rest or Extension on the handlebar vs. (b) Firmly power cylindrical grasping the handlebar.

Figure 2: (a) Power Hook Grasping a knife vs. (b) Precision Lumbrical Grasping a knife. (c) A natural reaction when seeing scene (b) is to open the hand to receive the knife.

Figure 3: Example of a Psychophysics experiment [Zhang et.al. VSS 2015].

Human Grasp Types
Seven fundamental grasp categories are: PoC: Power Cylindrical; PoS: Power Spherical; PoH: Power Hook; PrP: Precision Pinch; PrT: Precision Tripod; PrL: Precision Lumbrical; RoE: Rest or Extension.

Human Action Intention

Figure 4: Human action intention categories.

CNN for Grasp Type Recognition
We used a five layer CNN (including the input layer and one fully-connected perception layer for regression output). The first convolutional layer has 32 filters of size $5 \times 5$ with max pooling, the second convolutional layer has 32 filters of size $5 \times 5$ with average pooling, and the third convolutional layer has 64 filters of size $5 \times 5$ with average pooling, respectively. Each convolutional layer convolves its input with a bank of filters, then applies a point-wise non-linearity and max or average pooling operation. The final fully-connected perception layer has 7 regression outputs. The fully-connected perception layer applies linear filters to its input, then applies point-wise non-linearity. Our system considers 7 grasp type classes.

Manipulation Action Segmentation using Grasp Type Evolution
Grasp type evolution can help with a finer segmentation of manipulation actions:

Figure 6: Grasp type evolution (right hand) in an action.

Figure 7: Hand grasp type recognition along timeline and video segmentation results compared with ground truth segments.

In the 8 test clips, there are 18 ground truth segmentation key frames, and 14 of them are successfully detected, which yields a recall of 78%. Among the 20 detected segmentation key frames, 16 are correct, which yields a precision of 80%.

More Experimental Results

<table>
<thead>
<tr>
<th>Methods</th>
<th>F-O</th>
<th>S-O</th>
<th>C</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>GL+SVM</td>
<td>.54</td>
<td>.35</td>
<td>.73</td>
<td>.59</td>
</tr>
<tr>
<td>GT+SVM</td>
<td>.61</td>
<td>.35</td>
<td>.82</td>
<td>.71</td>
</tr>
</tbody>
</table>

Table 1: Inference of Action Intention from Grasp Type:

<table>
<thead>
<tr>
<th>Methods</th>
<th>F-O</th>
<th>S-O</th>
<th>C</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>GL+SVM</td>
<td>.63</td>
<td>.35</td>
<td>.73</td>
<td>.49</td>
</tr>
<tr>
<td>GT+SVM</td>
<td>.65</td>
<td>.35</td>
<td>.82</td>
<td>.83</td>
</tr>
</tbody>
</table>

Acknowledgements
For the details of our work:
• www.umiacs.umd.edu/ yzyang/ or yzyang@cs.umd.edu