Appearance-based Human Gesture Recognition using Multimodal Features for Human Computer Interaction

Luo Dan\textsuperscript{(a,b)} Hua Gao\textsuperscript{(b)} Hazim Kemal Ekenel\textsuperscript{(b)} Jun Ohya\textsuperscript{(a)}

\textsuperscript{(a)}GITS, Waseda University \textsuperscript{(b)}CVHCI, Karlsruhe Institute of Technology
Introduction

- Control of consumer electronics
- Interaction with visualization systems
- Control of mechanical systems
- Computer games
Challenges

• Different components of human gestures
• Wide variety of signs (ambiguous)
• Variable appearance/clothing
• Unconstrained illumination
• Local-body Occlusions
Two cyber gloves and three pohelmus 3SPACE-position trackers are used as input devices. 4942 isolated signs from two signers 3312 in the test set 78.1%
Hidden Conditional Random Fields for Gesture Recognition
Sy Bor Wang, CVPR2006

A 3D cylindrical body model,

**Stereo Camera**
Head Gesture Dataset
16 signer
A total of 152 head nods, 11 head shakes and 159 junk sequences

**Arm Gesture Dataset**
13 signer 6 classes
90 gestures for per class

<table>
<thead>
<tr>
<th>Models</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM $\omega = 0$</td>
<td>65.33</td>
</tr>
<tr>
<td>CRF $\omega = 0$</td>
<td>66.53</td>
</tr>
<tr>
<td>CRF $\omega = 1$</td>
<td>68.24</td>
</tr>
<tr>
<td>HCRF (multi-class) $\omega = 0$</td>
<td>71.88</td>
</tr>
<tr>
<td>HCRF (multi-class) $\omega = 1$</td>
<td>85.25</td>
</tr>
</tbody>
</table>
Recognizing temporal trajectories using the condensation algorithm
M. J. Black and Jepson, FG1998
System Overview

Input → Face Detection → Skin Segment → Feature Extraction → Recognition
Multimodal Feature

MCT Face Detector \(\rightarrow\) Face Region \(\rightarrow\) Color Train Data

Input Images \(\rightarrow\) HSV Images \(\rightarrow\) Gaussian Mixture Model \(\rightarrow\) Image Segmentation \(\rightarrow\) Output Face Blob Hand Blob

Face Detection \(\rightarrow\) Skin Segment \(\rightarrow\) Feature Extraction \(\rightarrow\) Recognition
Hand Feature

Disgust

Excite

Nervous

Input → Face Detection → Skin Segment → Feature Extraction → Recognition
Facial feature

7 Face expression subject [Training Dataset: “FEEDTUM”]

happy  surprised  fear  disgust  angry  sad  neutral

5 DCT coefficients from 64 blocks
Facial appearance feature vector (5×64=320 dimensional)
Facial feature

Expression Subspace-Expression Trajectories

Face Region by MCT → LDA Project → 7 expression Sub-space From FEED → 6 dimensional features
Facial Feature

Disgust

Excite

Happy

Input → Face Detection → Skin Segment → Feature Extraction → Recognition
Feature Combination

• Hand feature
  Hand Location

• Facial feature
  6 dimensional vectors

• Two different combination strategies
  - The first one is at feature level by combining the feature vectors extracted from face and hands. A statistical method can be used afterwards to select the most discriminative features for classification.
  - The second one is at decision level by combining the classification scores of each modality.
The sample set

\[ S_t = (\mu, \phi^l, \alpha^l, \rho^l, \phi^r, \alpha^r, \rho^r) \]

Prediction

\[ \mu_{t+1} = \mu_t \]

\[ \phi^i_t = \frac{1 - \sqrt{y}}{\sqrt{y}}, \quad y \in [0,1] \]

\[ \alpha^i_{t+1} = \alpha^i_t + N(\sigma_{\alpha}) \]

\[ \rho^i_{t+1} = \rho^i_t + N(\sigma_{\rho}) \]

Updating

\[ p(Z_{t,i} \mid s_t) = \frac{1}{\sqrt{2 \pi}} \exp \frac{-\sum_{j=0}^{\omega-1} (x_{(t-j),i} - \alpha^z m^{(\mu \mu)}_{(\phi^z \rho^z),i})^2}{2(\omega - 1)} \]
The sample set in each state

Hand feature trajectories

\[ S_t = (\mu, \phi^l, \alpha^l, \rho^l, \phi^r, \alpha^r, \rho^r) \]

Facial feature trajectories

\[ S_t = (\mu, \phi^f, \alpha^f, \rho^f) \]

Hand-face feature trajectories

\[ S_t = (\mu, \phi^l, \alpha^l, \rho^l, \phi^r, \alpha^r, \rho^r, \phi^f, \alpha^f, \rho^f) \]
Experiment - dataset

• 180 video clips of 12 human gestures with facial expression (1) anger, (2) apologize, (3) appreciate, (4) desire, (5) disgust, (6) excite, (7) fear, (8) happy, (9) nervous, (10) sad, (11) so-so and (12) surprise, Selected from ASL.

• Each sign includes three phases of a gesture: prestroke, stroke and poststroke.

• 3 people perform 3 to 7 times for each gesture. 1 as test data and the other 2 as train data.

• A training set and a testing data-set for evaluation. The training set contains one recording session per person, i.e. $12 \times 3 = 36$ video clips. The rest of the clips are used for test.

• Each video clip has a spatial resolution of $640 \times 480$ pixels with a frame-rate of 25fps and it is captured by a Logitech Webcam Pro 9000 from frontal view.
Experiments

Two different combination strategies
- Feature Level
  by combining the feature vectors extracted from LDA face projection feature and hand trajectories. A statistical method (condensation) can be used afterwards to select the most discriminative features for classification.
  Feature [Face, Hand], Recognition result: 83.2%

- Decision level
  by combining the classification scores of each modality.
  Feature [Hand] | Feature [face], Recognition result: 92.6%

<table>
<thead>
<tr>
<th>Modality</th>
<th>Recognition rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hand gesture</td>
<td>85.4%</td>
</tr>
<tr>
<td>Facial expression (FE)</td>
<td>45.0%</td>
</tr>
<tr>
<td>Hand + FE (Decision fusion)</td>
<td>92.6%</td>
</tr>
<tr>
<td>Hand + FE (Feature fusion)</td>
<td>83.2%</td>
</tr>
</tbody>
</table>
## Experiments

Confusion matrix for the condensation-based classification on database (hand motion result)

<table>
<thead>
<tr>
<th></th>
<th>anger</th>
<th>apologize</th>
<th>appreciate</th>
<th>desire</th>
<th>disgust</th>
<th>excite</th>
<th>fear</th>
<th>happy</th>
<th>nervous</th>
<th>sad</th>
<th>soso</th>
<th>surprised</th>
</tr>
</thead>
<tbody>
<tr>
<td>anger</td>
<td>70.0</td>
<td>10.0</td>
<td>0.0</td>
<td>10.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>10.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>apologize</td>
<td>0.0</td>
<td>72.7</td>
<td>0.0</td>
<td>0.0</td>
<td>27.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>appreciate</td>
<td>0.0</td>
<td>0.0</td>
<td>90.9</td>
<td>0.0</td>
<td>0.0</td>
<td>9.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>desire</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>66.7</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>22.2</td>
<td>11.1</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>disgust</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>100.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>excite</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>94.7</td>
<td>0.0</td>
<td>0.0</td>
<td>5.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>fear</td>
<td>8.3</td>
<td>0.0</td>
<td>8.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>83.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>happy</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>100.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>nervous</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>30.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>70.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>sad</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>14.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>57.1</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>soso</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>100.0</td>
<td>0.0</td>
</tr>
<tr>
<td>surprised</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>3.6</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>96.4</td>
</tr>
</tbody>
</table>
Conclusion

- an appearance-based multi-modal gesture recognition framework, which combines facial expression and hand motions.

- 12 classes of human gestures with facial expression from ASL.

- Two fusion strategies: the decision fusion and feature fusion.

- Experimental results showed that the analysis of facial expression helps distinguishing ambiguous hand gestures and facial analysis improves hand gesture recognition.

- In particular, decision level fusion improves the recognition rate from 85.4% to 92.6%.
Acknowledgment

• InterACT program
  Waseda University and Karlsruhe Institute of Technology

• German Excellence Initiative
  “Concept for the Future”
Thank you!

Q?