Applying Vision to Intelligent Human-Computer Interaction

Guangqi Ye

Department of Computer Science
The Johns Hopkins University
Baltimore, MD 21218
October 21, 2005
Vision for Natural HCI

• Advantages
  Affordable, non-intrusive, rich info.
• Crucial in multimodal interface
  Speech/gesture system
• Vision-Based HCI: 3D interface, natural

HandVu and ARToolkit by M. Turk, et. al
Motivation

• Haptics + Vision
  Remove constant contact limit.
• Gestures for vision-based HCI
  Intuitive with representation power
  Applications: 3D VE, tele-op., surgical
• Addressed problems
  Visual data collection
  Analysis, model and recognition
Outline

- Vision/Haptics system
- Modular framework for VBHCI
- 4DT platform
- Novel scheme for hand motion capture
- Modeling composite gestures
- Human factors experiment
Vision + Haptics

- 3D Registration via visual tracking
  Remove limitation of constant contact
- Different passive objects to generate various sensation
Vision: Hand Segmentation

- Model background: color histograms
- Foreground detection: histogram matching
- Skin modeling: Gaussian model on Hue
Vision: Fingertip Tracking

- Fingertip detection: model-based

- Tracking: prediction (Kalman) + local detection
Haptics Module

- 3-D registration:

\[ H \tilde{P} = H_c R C \tilde{P} + \tilde{t} \]

- Interaction simulation

\[ F = F_{GC} + \Lambda_{gain} f \]

- Examples: planes, buttons
Experimetal Results

• System: Pentimum III PC, 12fps
Vision + Haptics: Video
Outline

- Vision/Haptics system
- Modular framework for VBHCI
- 4DT platform
- Novel scheme for hand motion capture
- Modeling composite gestures
- Human factors experiment
- Conclusions
Visual Modeling of Gestures: General Framework

- Gesture generation

- Gesture recognition

[Diagram showing the process of gesture generation and recognition]
Related Research in Modeling Gestures for HCI

![Graph showing gesture vocabulary and number of users]
Targeted Problems

• Analysis: mostly tracking-based
  Our approach: using localized parser

• Model: single modality (static/dynamic)
  Our model: coherent multimodal framework

• Recognition: Limited vocabulary/users
  Our contributions: large-scale experiment
Visual Interaction Cues (VICs) Paradigm

- Site-centered interaction
  Example: cell phone buttons
VICs State Mode

- Extend interaction functionality
  3D gestures
VICs Principle: Sited Interaction

• Component mapping
Localized Parsers

- Low-level Parsers
  - Motion, shape
- Learning-Based Modeling
  - Neural Networks, HMMs
System Architecture

- Application Layer
- Message Processing Layer
- I/O Layer
  - VICs Processing
  - Interface Rendering
- System Layer
Outline

- Vision/Haptics system
- Modular framework for VBHCI
- 4DT platform
- Novel scheme for hand motion capture
- Modeling composite gestures
- Human factors experiment
- Conclusions
4D-Touchpad System

- Geometric calib.
  Homography-based

- Chromatic calib.
  Affine model for appearance transform
System Calibration Example

![Image showing system calibration example](image-url)
Hand Detection

• Foreground segmentation
  Image difference

• Modeling skin color
  Thresholding in YUV space
  Training: 16 users, 98% accuracy

• Hand region detection
  Merge skin pixels in segmented foreground
Hand Detection Example
Integrated into Existing Interface

- Shape parser + state-based gesture modeling

Mouse Control of Standard Applications
Outline

- Vision/Haptics system
- Modular framework for VBHCI
- 4DT platform
- **Novel scheme for hand motion capture**
- Modeling composite gestures
- Human factors experiment
- Conclusions
Efficient Motion Capture of 3D Gesture

- Capturing shape and motion in local space
- Appearance feature volume
  Region-based stereo matching

\[ F_{x,y,z} = \text{Match}(I_l(x_0+x,y_0+y), I_r(x_0+x+z,y_0+y)) \]

- Motion: differencing appearance
Appearance Feature Example
Posture Modeling Using 3D Feature

• Model 1: 3-layer neural networks
  Input: raw feature
  NN: 20 hidden nodes

<table>
<thead>
<tr>
<th>Posture</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pick</td>
<td>99.97%</td>
<td>99.18%</td>
</tr>
<tr>
<td>Push</td>
<td>100.00%</td>
<td>99.93%</td>
</tr>
<tr>
<td>Press-Left</td>
<td>100.00%</td>
<td>99.89%</td>
</tr>
<tr>
<td>Press-Right</td>
<td>100.00%</td>
<td>99.96%</td>
</tr>
<tr>
<td>Stop</td>
<td>100.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Grab</td>
<td>100.00%</td>
<td>99.82%</td>
</tr>
<tr>
<td>Drop</td>
<td>100.00%</td>
<td>99.82%</td>
</tr>
<tr>
<td>Silence</td>
<td>99.98%</td>
<td>98.56%</td>
</tr>
</tbody>
</table>
Posture Modeling Using 3D Feature

- Model 2: histogram-based ML

Input: vector quantization, 96 clusters

\[ v = \arg \max_{v_i} P(s|v_i) \]

<table>
<thead>
<tr>
<th>Posture</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pick</td>
<td>96.95%</td>
<td>97.50%</td>
</tr>
<tr>
<td>Push</td>
<td>96.98%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Press-Left</td>
<td>100.00%</td>
<td>94.83%</td>
</tr>
<tr>
<td>Press-Right</td>
<td>99.07%</td>
<td>98.15%</td>
</tr>
<tr>
<td>Stop</td>
<td>99.80%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Grab</td>
<td>98.28%</td>
<td>95.00%</td>
</tr>
<tr>
<td>Drop</td>
<td>100.00%</td>
<td>98.85%</td>
</tr>
<tr>
<td>Silence</td>
<td>98.90%</td>
<td>98.68%</td>
</tr>
</tbody>
</table>
Dynamic Gesture Modeling

- Hidden Markov Models
  Input: VQ, 96 symbols
  Extension: modeling stop state \( p(s_T) \)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Twist</td>
<td>96.30</td>
<td>81.48</td>
<td>100.00</td>
<td>85.19</td>
</tr>
<tr>
<td>Twist-Anti</td>
<td>93.62</td>
<td>93.10</td>
<td>100.00</td>
<td>93.10</td>
</tr>
<tr>
<td>Flip</td>
<td>100.00</td>
<td>96.43</td>
<td>100.00</td>
<td>96.43</td>
</tr>
<tr>
<td>Negative</td>
<td></td>
<td>79.58</td>
<td></td>
<td>98.79</td>
</tr>
<tr>
<td>Overall</td>
<td>96.64</td>
<td>81.05</td>
<td>100.00</td>
<td>97.89</td>
</tr>
</tbody>
</table>
Outline

- Vision/Haptics system
- Modular framework for VBHCI
- 4DT platform
- Novel scheme for hand motion capture
  - Modeling composite gestures
- Human factors experiment
- Conclusions
Model Multimodal Gestures

• Low-level gesture as Gesture Words
  3 classes: static, dynamic, parameterized

• High-level gesture
  Sequence of GWords

• Bigram model to capture constraints

\[
P(W|S) \propto P(W) \cdot P(S|W)
\]

\[
= \pi(v_1) \prod_{i=2}^{p} A(v_i|v_{i-1}) \cdot \prod_{i=1}^{p} B(S_i|v_i)
\]
Example Model
Learning and Inference

- Learning the bigram: maximum likelihood
- Inference: greedy-choice for online
  Choose path with maximum $p(v_t|v_{t-1})p(s_t|v_t)$
Outline

- Vision/Haptics system
- Modular framework for VBHCI
- 4DT platform
- Novel scheme for hand motion capture
- Modeling composite gestures
- Human factors experiment
- Conclusions
Human Factors Experiment

• Gesture vocabulary: 14 gesture words
  Multi-Modal: posture, parameterized and dynamic gestures
  9 possible gesture sentences

• Data collecting
  16 volunteers, including 7 female
  5 training and 3 testing sequences

• Gesture cuing: video + text
Example Video Cuing

TEST MOVE

1. Enter
2. PICK
3. MOVE
4. DROP
5. Retract
Modeling Parameterized Gesture

• Three Gestures: moving, rotate, resize

• Region tracking on segmented image
  Pyramid SSD tracker: $X' = R(\Theta)X + T$
  Template: 150 x 150

• Evaluation
  Average residual error: 5.5/6.0/6.7 pixels
## Composite Gesture Modeling Result

<table>
<thead>
<tr>
<th>Gesture</th>
<th>Sequences</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pushing</td>
<td>35</td>
<td>97.14%</td>
</tr>
<tr>
<td>Twisting</td>
<td>34</td>
<td>100.00%</td>
</tr>
<tr>
<td>Twisting-Anti</td>
<td>28</td>
<td>96.42%</td>
</tr>
<tr>
<td>Dropping</td>
<td>29</td>
<td>96.55%</td>
</tr>
<tr>
<td>Flipping</td>
<td>32</td>
<td>96.89%</td>
</tr>
<tr>
<td>Moving</td>
<td>35</td>
<td>94.29%</td>
</tr>
<tr>
<td>Rotating</td>
<td>27</td>
<td>92.59%</td>
</tr>
<tr>
<td>Stopping</td>
<td>33</td>
<td>100.00%</td>
</tr>
<tr>
<td>Resizing</td>
<td>30</td>
<td>96.67%</td>
</tr>
<tr>
<td>Total</td>
<td>283</td>
<td>96.47%</td>
</tr>
</tbody>
</table>
User Feedback on Gesture-based Interface

• Gesture vocabulary
  Easy to learn: 100% agree

• Fatigue compared to GUI with mouse
  50%: comparable, 38%: more tired, 12% less

• Overall convenience compared to GUI with mouse
  44%: more comfortable
  44%: comparable
  12%: more awkward
Contributions

• Vision+Haptics: novel multimodal interface
• VICs/4DT: a new framework for VBHCI and data collection
• Efficient motion capture for gesture analysis
• Heterogeneous gestures modeling
• Large-scale gesture experiments
Acknowledgement

• Dr. G. Hager
• Dr. D. Burschka, J. Corso, A. Okamura
• Dr. J. Eisner, R. Etienne-cummings, I. Shafran
• CIRL Lab
  X. Dai, L. Lu, S. Lee, M. Dewan, N. Howie, H. Lin, S. Seshanami
• Haptics Explorative Lab
  J. Abott, P. Marayong
Thanks