

# Robust Modeling of Heterogeneous Gestures Using Localized Parsers

Guangqi Ye

Computational Interaction and Robotics Lab  
Department of Computer Science  
The Johns Hopkins University

# Gesture for HCI

- Vision-based interaction: new interaction concept; non-intrusive
- Power of gestures  
Natural, intuitive, rich interaction medium

- Example application

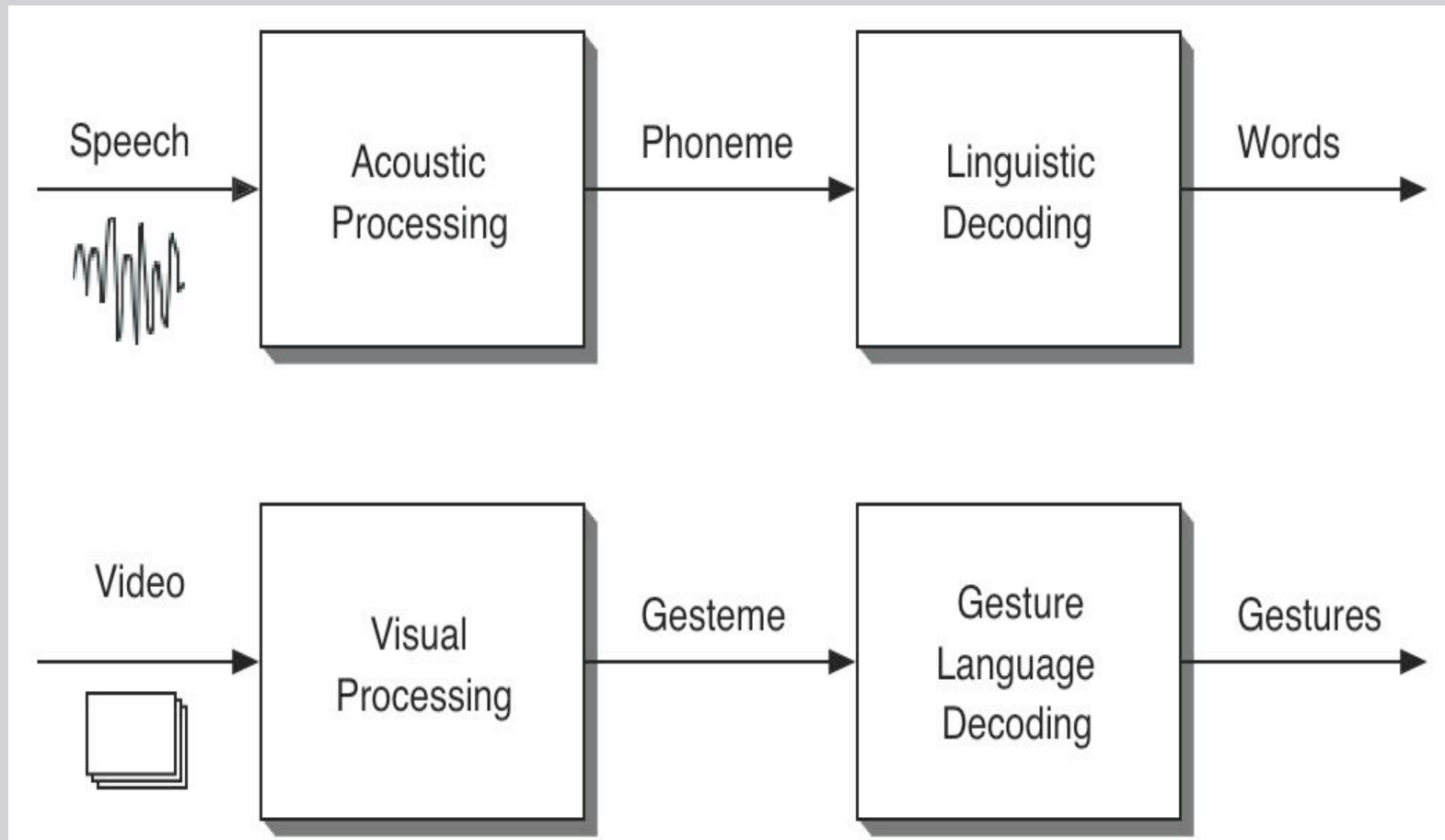
Virtual Design Panel

Tele-operation

Training and assistance in surgical sites



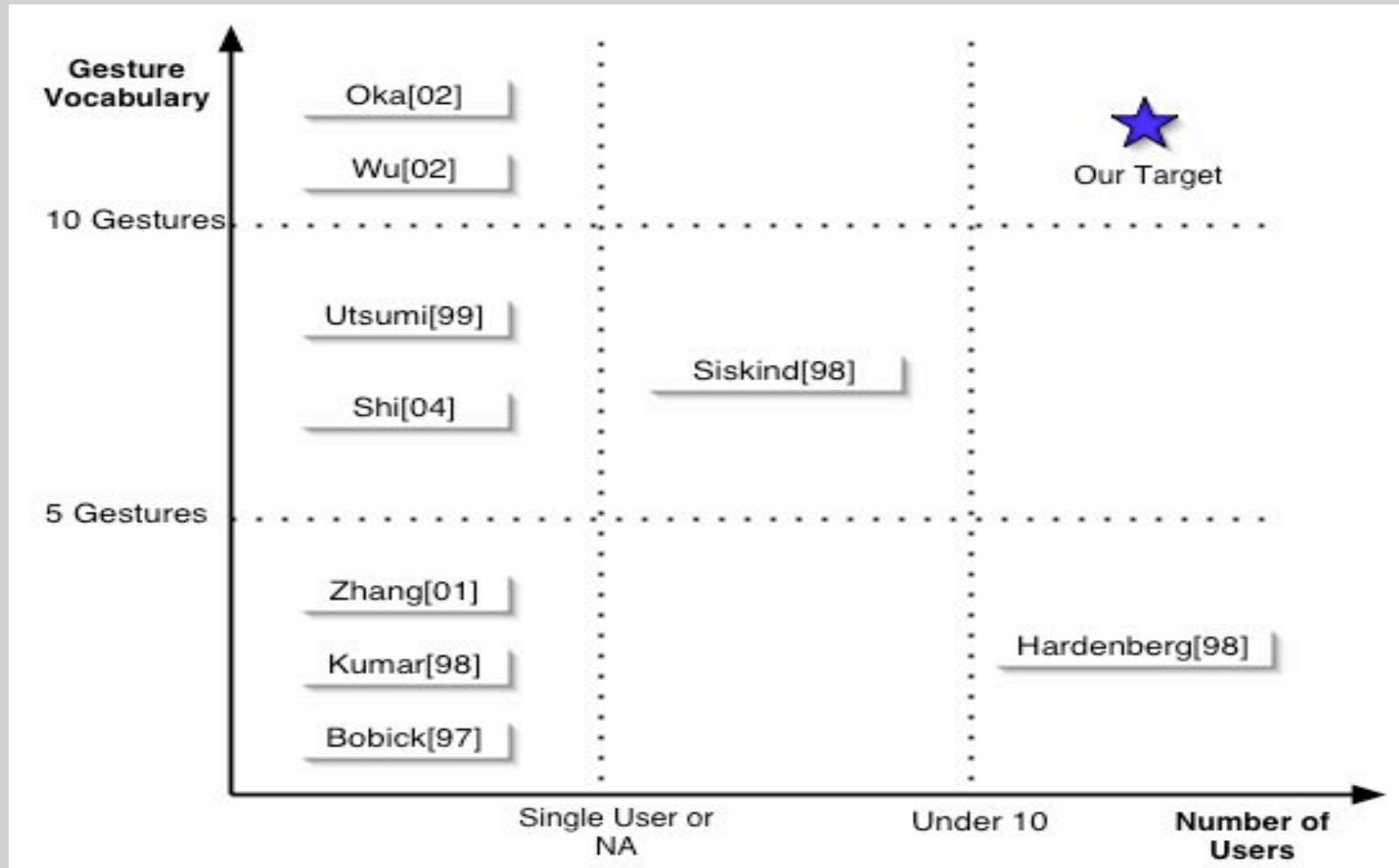
# Analog Between Speech and Gesture Processing



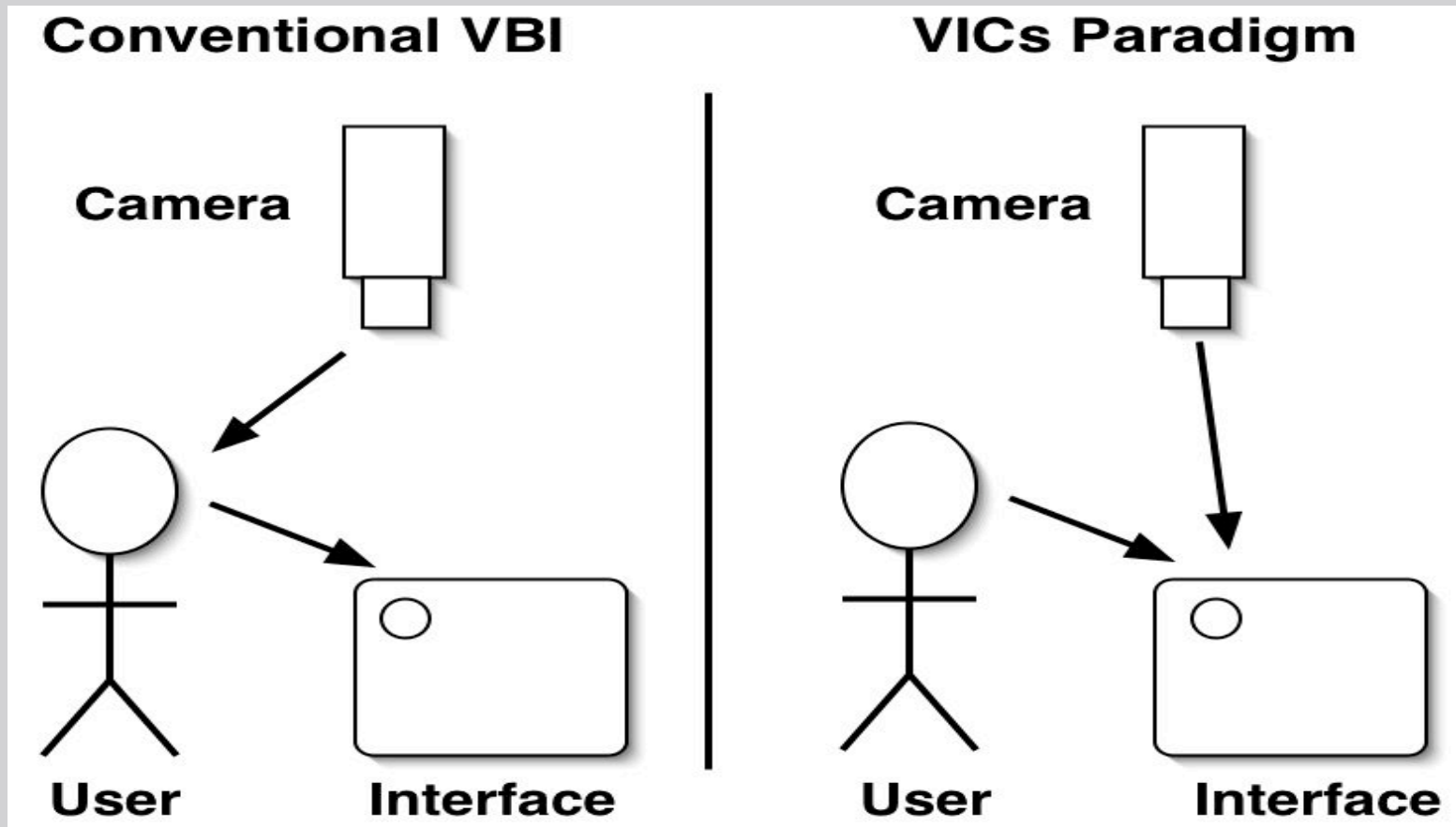
# Targeted Problems

- Analysis: mostly tracking-based
  - Our approach:** using localized parser
- Recognition
  - Limited vocabulary and number of users
  - Single modality: posture / dynamic / tracking
  - Continuous gesture recognition
  - Our contribution:** relatively large vocabulary; multiple users; model continuous multimodal gestures

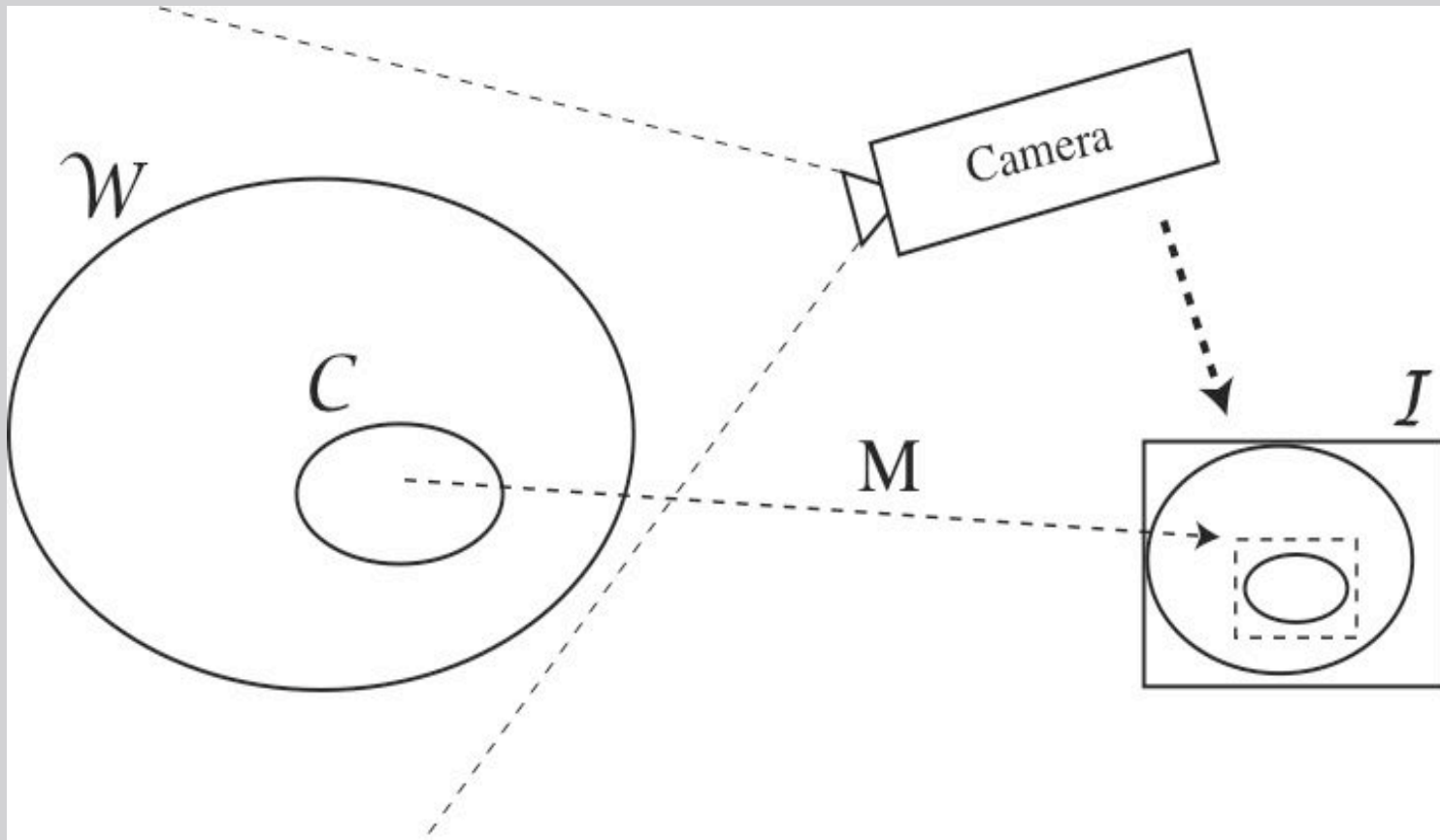
# Related Research Overview



# Visual Interaction Cues(VICs) Paradigm



# VICs Principle: Sited Interaction

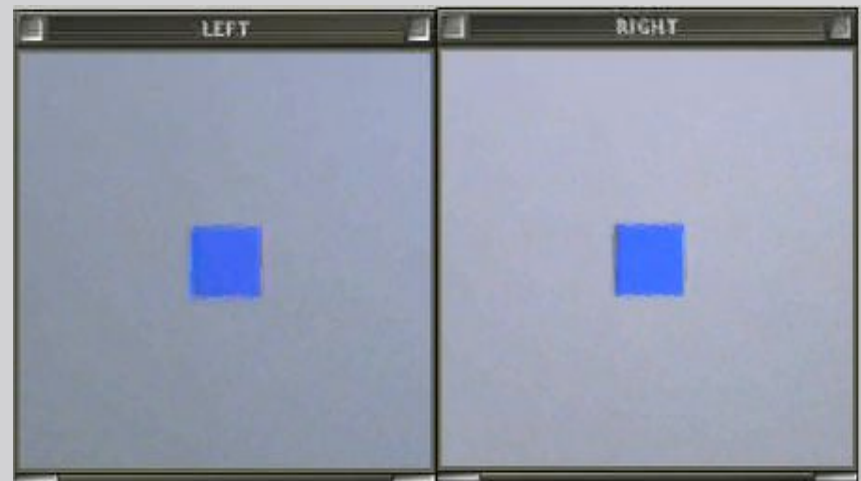


# Localized Parsers

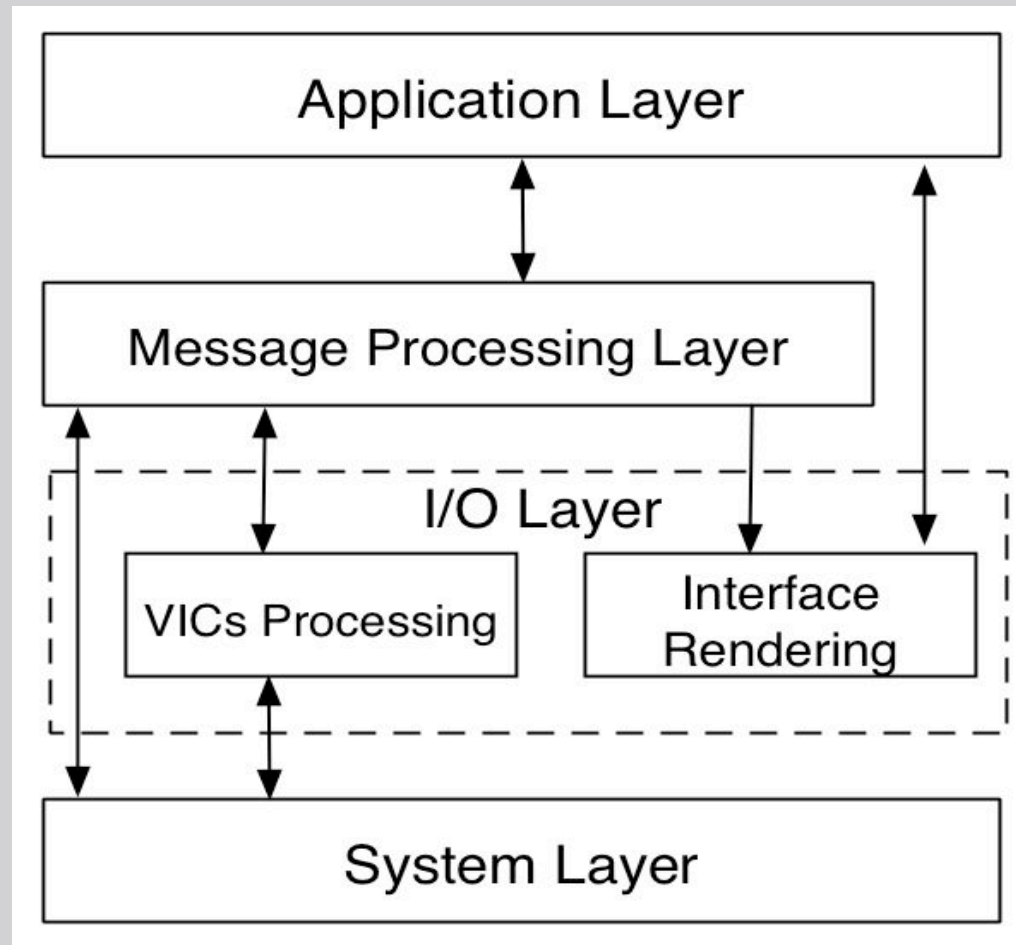
- Low-level parsers  
Motion, shape



- Learning-based modeling  
Neural Network  
Hidden Markov Models  
Finite-State Machines

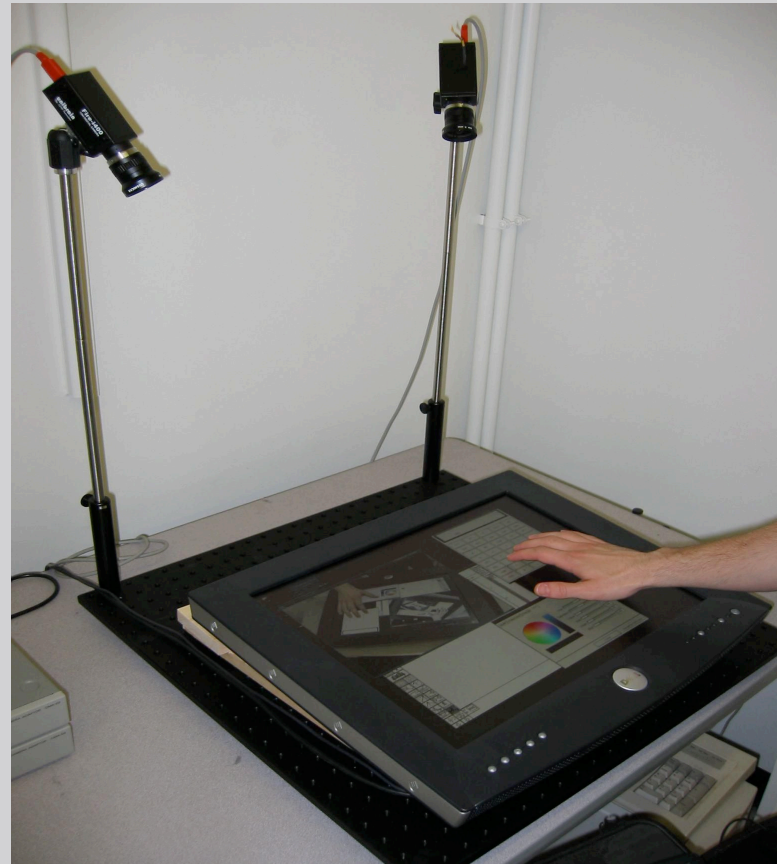


# System Architecture

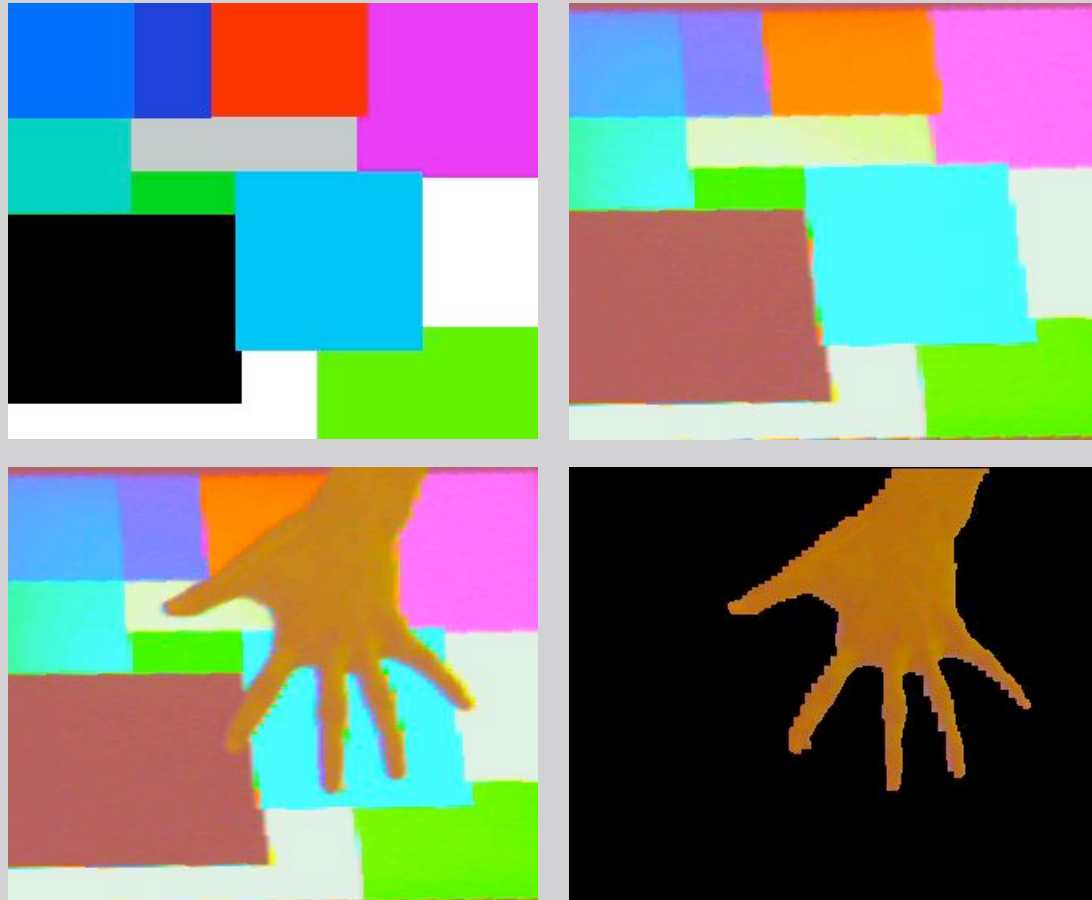


# 4D-Touchpad System

- Geometric calib.  
Homography-based
- Chromatic calib.  
Affine model for  
appearance transform



# System Calibration Example



# Hand Detection

- Foreground segmentation

Image difference

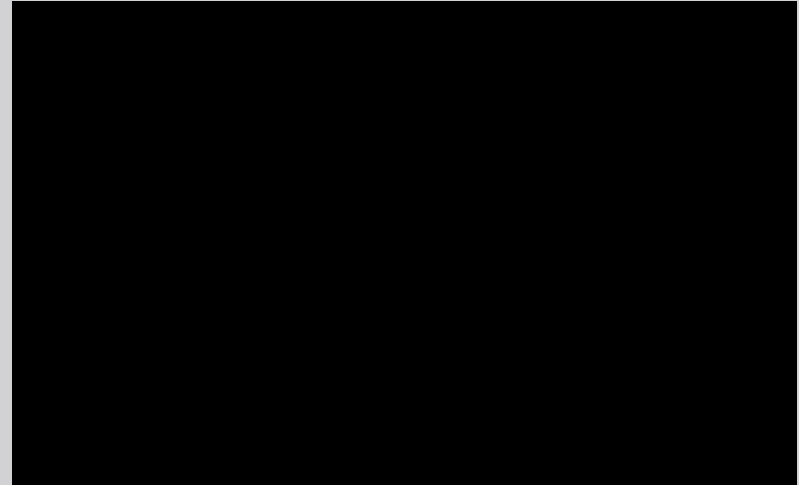
- Hand region detection

Train linear color model of skin in YUV space

Data collection: hand sequences of 16 people

Training: over 98% correct

# Hand Detection Example



# Integrated into Existing Interface

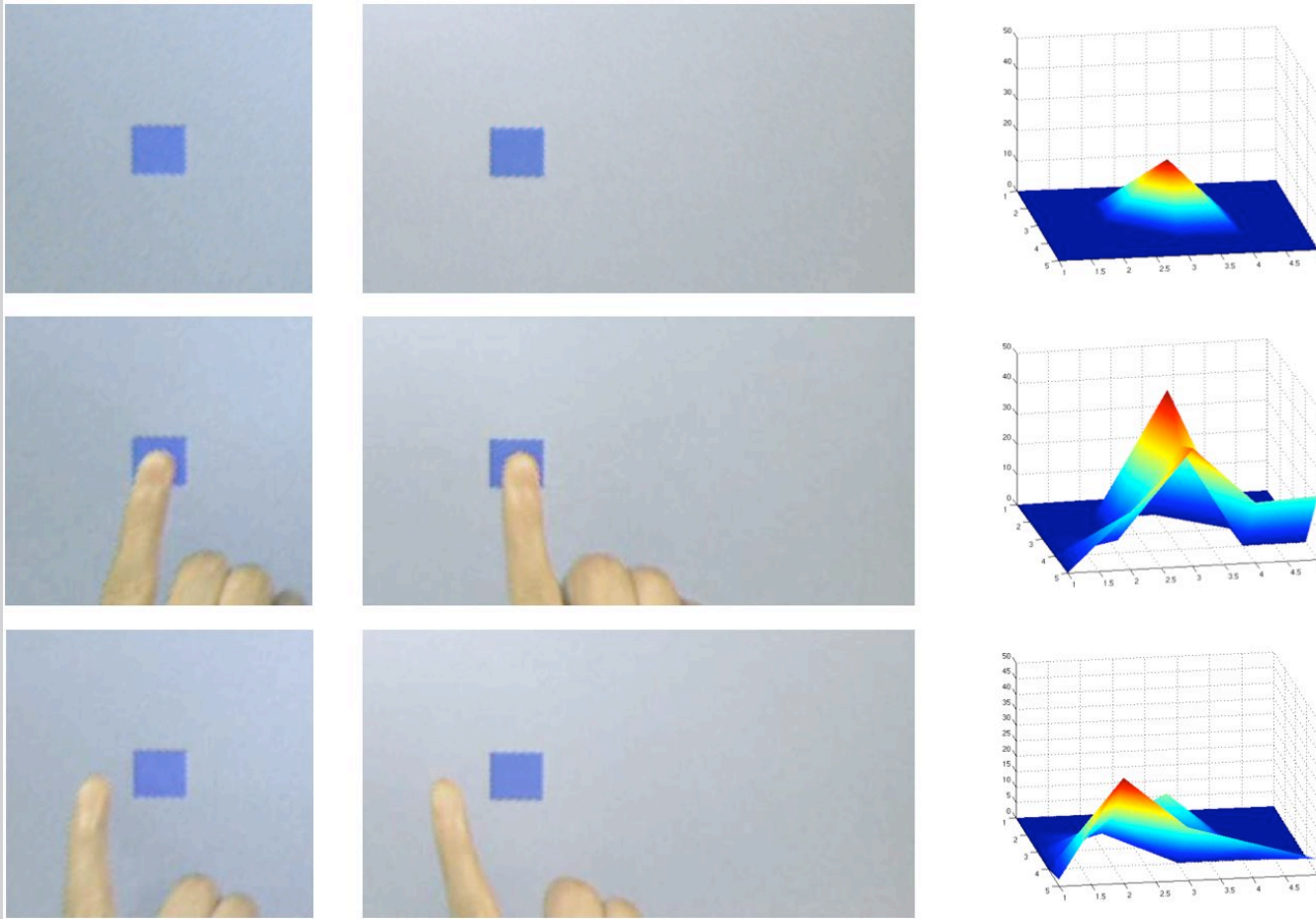
- Shape parser + state-based gesture modeling

**Mouse Control of  
Standard Applications**

# Describing 3D Appearance and Motion

- Capturing shape and motion in local space
- Appearance feature: region-based stereo matching
- Motion: differencing appearance

# Local Feature Example



# Multi-User Gesture Experiment

- Gesture vocabulary: 13 gestures  
Multi-Modal: posture, parameterized and dynamic gestures
- 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

# Posture Classification 1

- 3-layer neural network with 800 input nodes and 20 hidden nodes

<b>Posture</b>	<b>Training</b>	<b>Testing</b>
Pick	99.97%	99.18%
Push	100.00%	99.93%
Twist	99.99%	99.96%
Twist-Anti	100.00%	99.89%
Stop	100.00%	100.00%
Resize	100.00%	99.82%
Drop	100.00%	99.82%
Negative	99.98%	98.93%

# Posture Classification 2

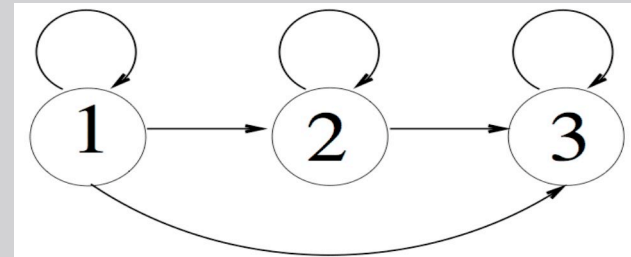
- Vector Quantization: using 96 clusters.
- Recognition: histogram plus Bayesian rule

<b>Posture</b>	<b>Training</b>	<b>Testing</b>
Pick	96.88%	97.43%
Push	96.98%	100.00%
Twist	100.00%	100.00%
Twist-Anti	100.00%	93.10%
Stop	99.80%	100.00%
Resize	98.28%	93.33%
Drop	100.00%	98.85%
Negative	91.02%	98.03%

# HMM Modeling of Dynamic Gesture

- 3-state forward HMM  
Ending state modeling

$$p(s_N = s_i)$$



- Training: 54 twists, 47 twist-anti and 149 flips, 100.0% correct.
- Testing: 84 valid and 578 invalid, 97.73% correct.

# Modeling Parameterized Gesture

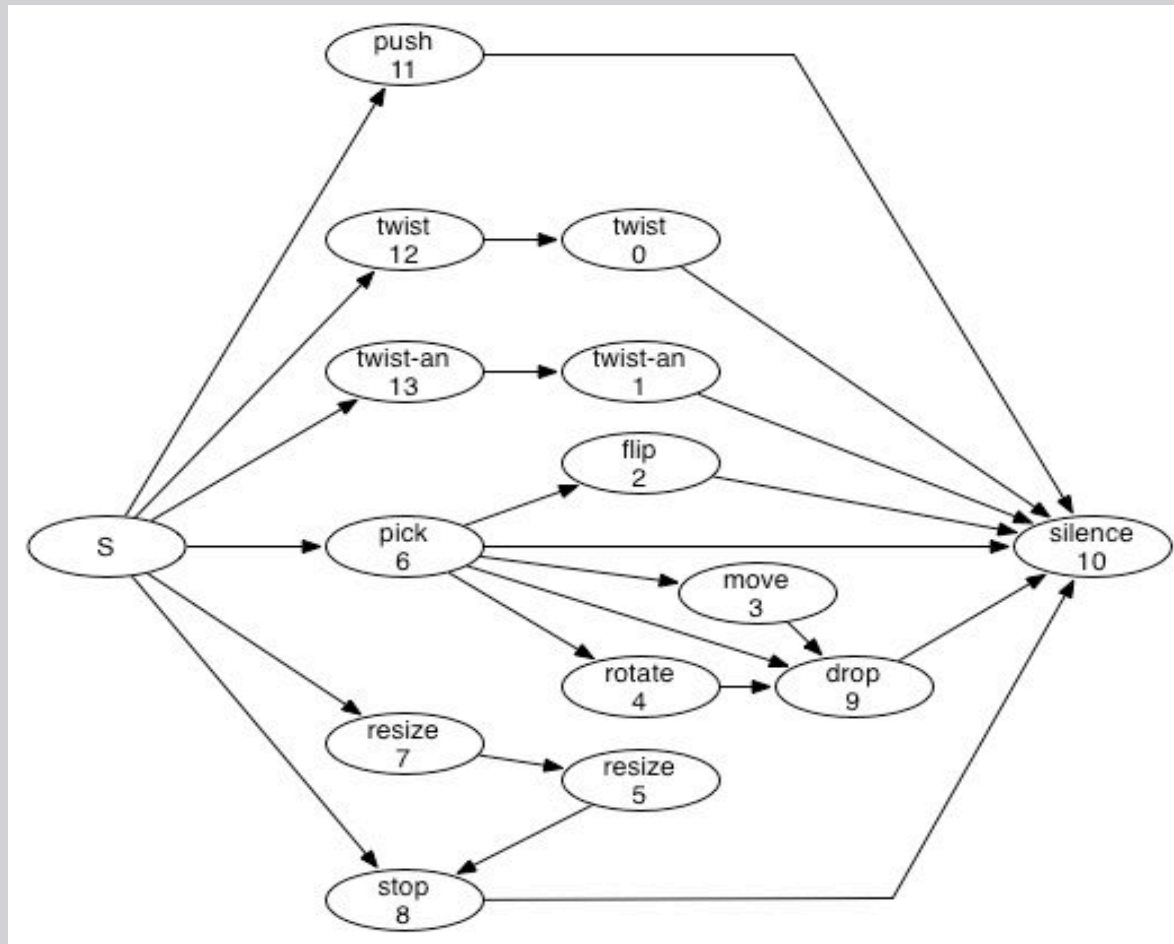
- Two gestures: moving and rotating
- Using localized tracking  
Pyramid SSD tracker:  $X' = R(\Theta)X + T$   
Template: 150 x 150
- Evaluation  
Average residual error: 5.8 pixel for moving and 6.30 for rotating gesture.

# Model Multimodal Gestures in Probabilistic Framework

- Define low-level gesture as Gesture Words
- High-level gesture  
Series of temporally and contextually constrained gesture words
- Using bigram model to capture constraints

$$\begin{aligned} P(W | S) &\propto P(W)P(S | W) \\ &= P(v_1) \prod_{i=2}^n P(v_i | v_{i-1}) * \prod_{i=1}^n P(S_i | v_i) \end{aligned}$$

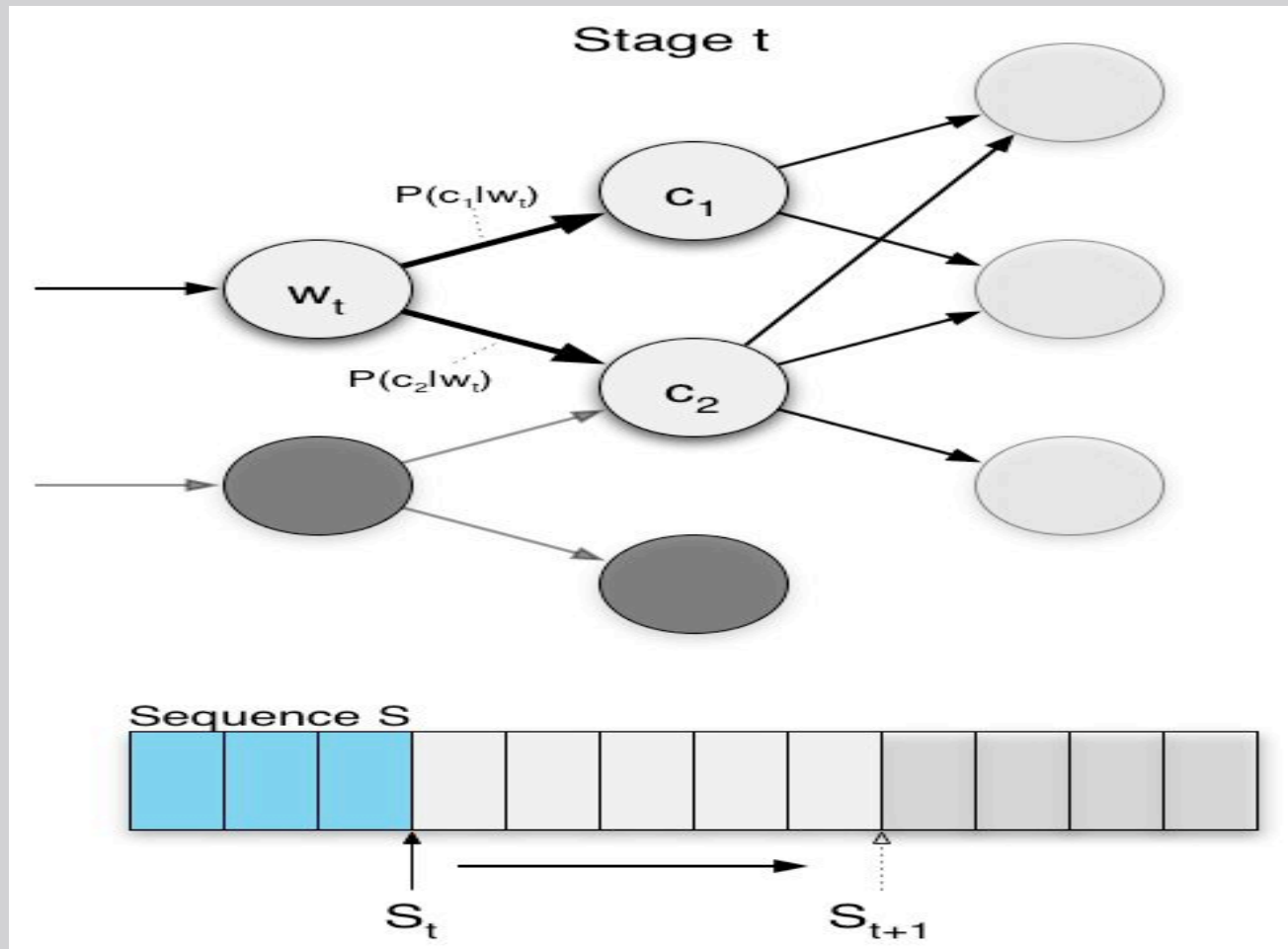
# Model Example



# Inference and Learning

- Difficulties of full search
  - Computational complexity
  - Instant user feedback
- Making greedy choice
  - Choose path with maximum  $p(v_t|v_{t-1})p(s_t|v_t)$
- Learning
  - Maximum-likelihood learning of bigram model

# Inference Example



# Testing Result

<b>Gesture</b>	<b>Sequences</b>	<b>Ratio</b>
Push	35	97.14%
Twist	34	100.00%
Twist-Anti	28	96.42%
Flip	32	96.88%
Pick	33	100.00%
Drop	29	96.55%
Move	35	97.14%
Rotate	27	100.00%
Stop	33	100.00%
Resize	30	96.67%

# Conclusion

- VICs: an efficient framework for motion capturing and gesture analysis
- Localized parsers are capable of learning relatively large gesture vocabulary of multiple users
- Heterogeneous gestures modeling in a coherent framework

# Thanks

