

# Human Body Language Understanding with 3D Sensors

**Zhengyou Zhang**

Microsoft Research

[zhang@microsoft.com](mailto:zhang@microsoft.com)

<http://research.microsoft.com/~zhang/>

# Kinect Sensor



infra-red  
projector

RGB  
camera

infra-red  
camera



Microphones  
Motor  
USB



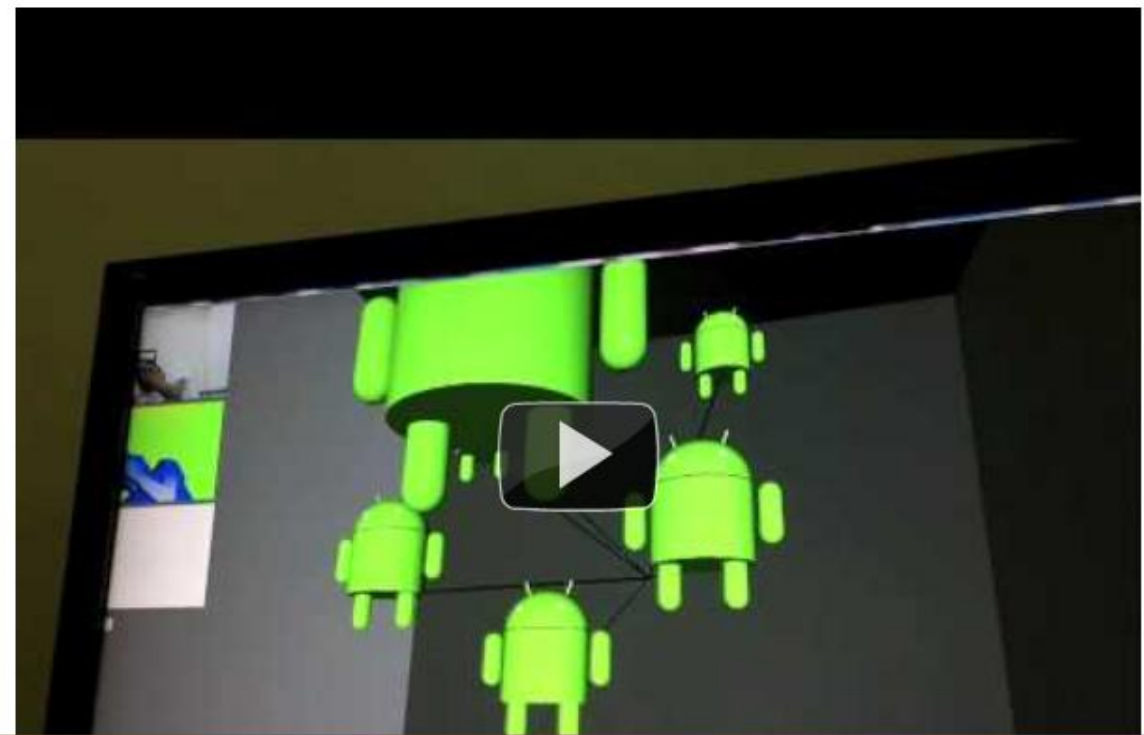
# KINECTHACKS

- HOME
- FORUMS
- FAQ
- GUIDES
- TOOLS AND RESOURCES
- ABOUT

## Crazy Head Tracking Androids

December 4th, 2010 Madhav K 4 Comments and 6 Reactions

Sittiphol Phanvilai @ Hua Lampong Co.,Ltd has implemented head tracking using Kinect and creates crazy 3D effects with android dolls. This is a modification of their earlier Kinect VR project which had spheres instead of androids.



Search



### FORUMS



### FACEBOOK

[Sign Up](#)

Create an account or [log in](#) to see what your friends like.

[Kinect Hacking on Facebook](#)

765 people like Kinect Hacking

Dan

Graham

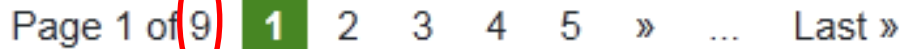
Eric

Stephan

Lawrence

# ~ 90 Projects as of 12/04/2010

24 pages as of 2/10/2011



Page 1 of 9 1 2 3 4 5 » ... Last »

Every few hours new applications are emerging for the Kinect and creating new phenomenon that is nothing short of revolutionary.

- Quote from KinectHacks.net

# 3D Video Capture

# Music Video

# **Navigational Aids for the Visually Impaired**

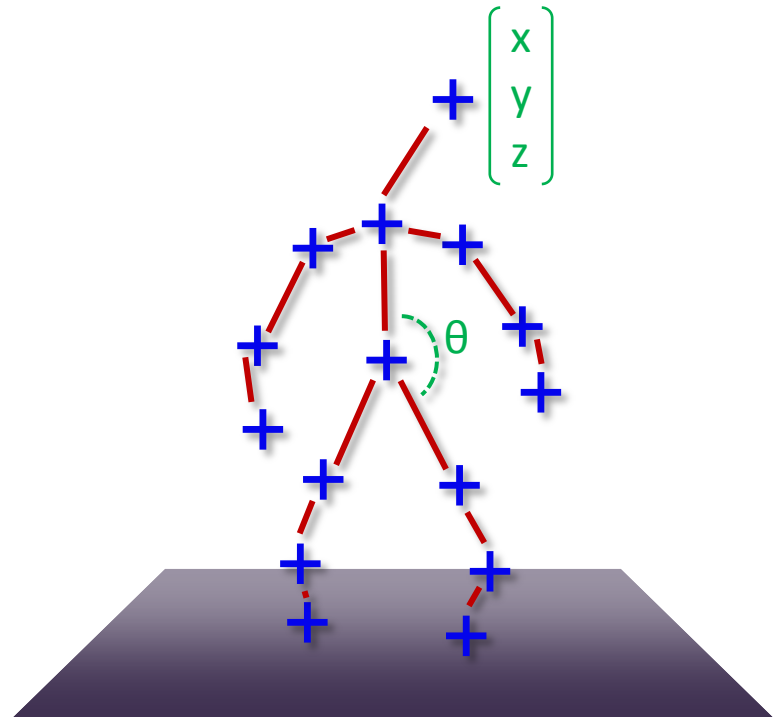
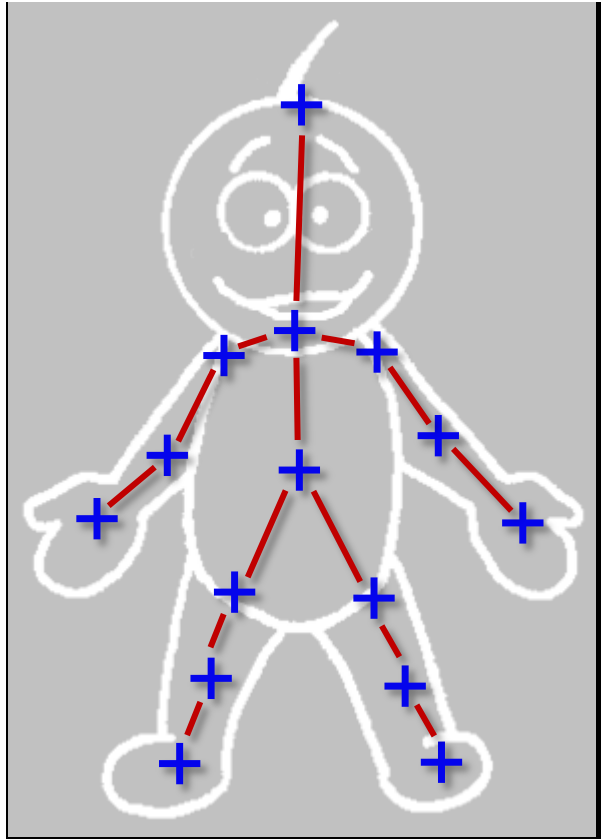


# **HUMAN BODY LANGUAGE UNDERSTANDING**

Jamie Shotton, Andrew Blake, Kinect Team

# **SKELETAL TRACKING**

# Human pose estimation



Kinect tracks 20 body joints in real time.

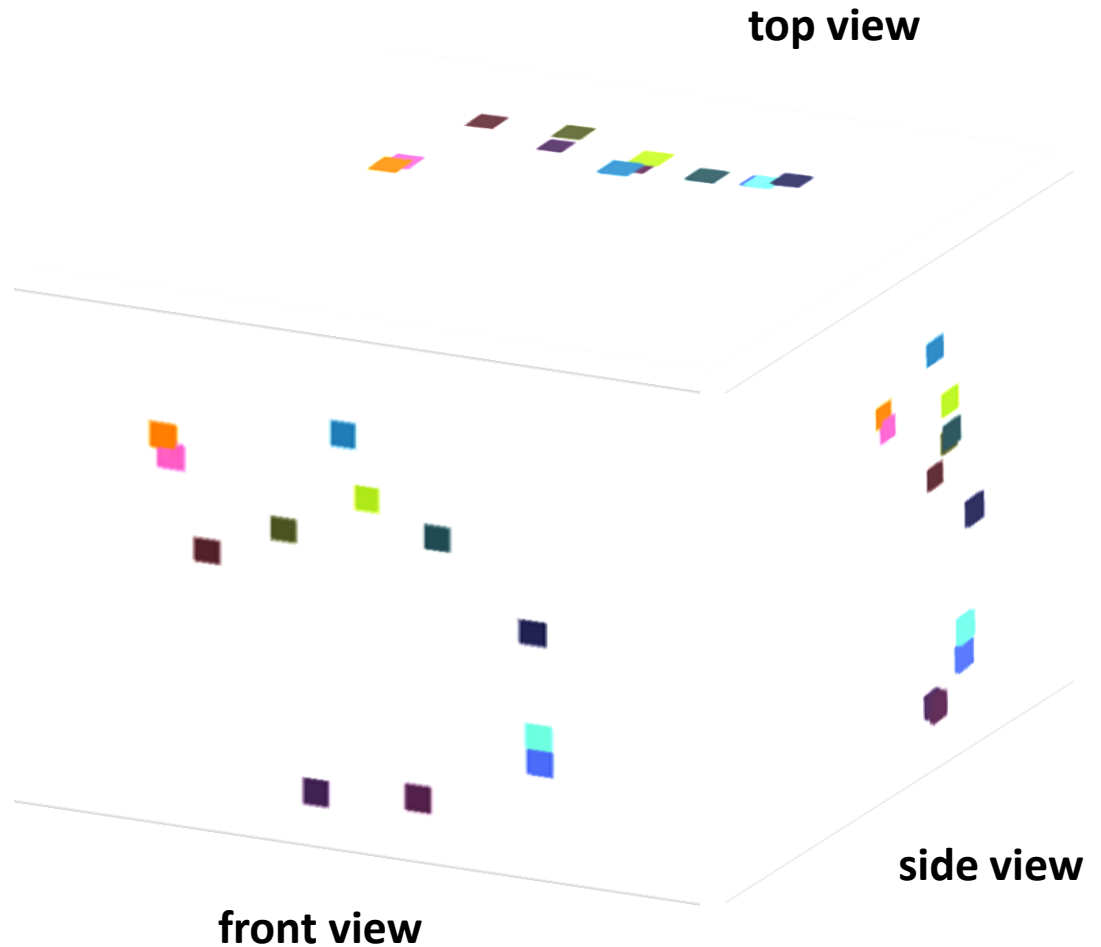
# Skeletal Tracking



input depth image



inferred body parts &  
overlaid joint hypotheses



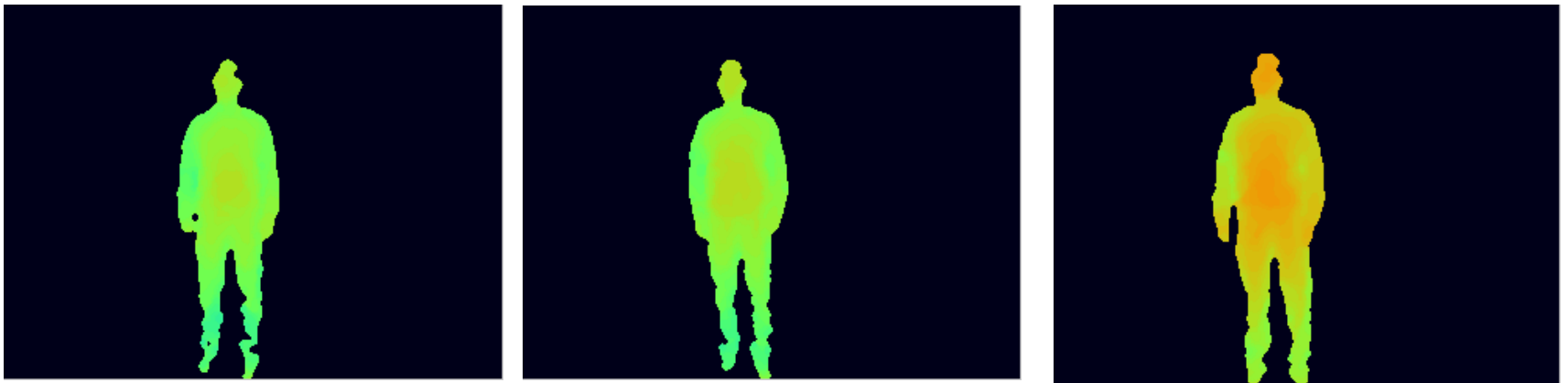
3D joint hypotheses

Wanqing Li, Zhengyou Zhang, Zicheng Liu

# **HUMAN ACTION RECOGNITION**

# The Problem

- Recognize actions from sequences of depth maps

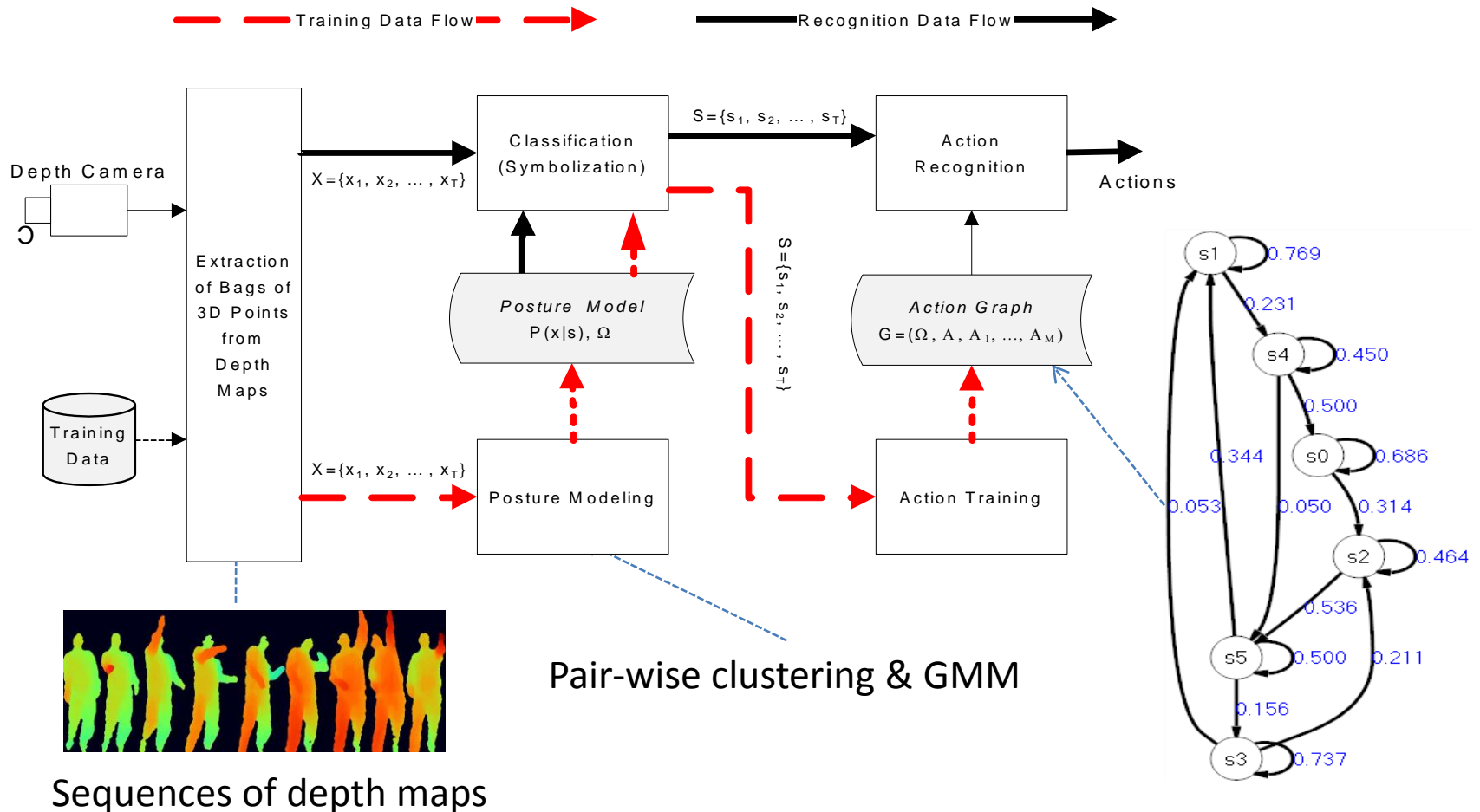


- Issues to address
  - large amount of data
  - Coarse and noisy depth measurement

[Tennis Swing](#)

# Method - Action Graphs

- Node: Salient posture
- Path: Action



# Posture Modeling

- 3D representative points are sampled from each depth map → A Bag of Points (BoPs)
  - Projection based
- Distribution of the 3D points for each posture
  - GMM
- Distances between two depth maps
  - Hausdorff distance between the two BoPs



# Experimental Results

- Data Collection
  - Depth camera using structured infrared light
  - Depth map resolution 640x480 pixels
  - 20 Actions
    - Movement of arms, legs, torso and coordination of them
  - 7 Subjects
    - Each subject performed each action 3 times

# 20 Actions

- 20 actions
  - 10 with one hand, 2 with two hands, 2 with one leg
  - 6 with whole body

<b>High-arm wave</b>	<b>Two hand wave</b>
<b>Horizontal-arm wave</b>	<b>Side-boxing</b>
<b>Hammer</b>	<b>Bend</b>
<b>Hand catch</b>	<b>Forward-kick</b>
<b>Forward punch</b>	<b>Side-kick</b>
<b>High throw</b>	<b>Jogging</b>
<b>Draw x</b>	<b>Tennis swing</b>
<b>Draw tick</b>	<b>Tennis swing</b>
<b>Draw Circle (Clockwise)</b>	<b>Golf-swing</b>
<b>Hand clap</b>	<b>Pickup &amp; throw</b>

# Three Test Actions Sets

- Due to consideration of the computational cost, the 20 actions are divided into three subsets:

Action Set One (AS1)	Action Set Two (AS2)	Action Set Three (AS3)
Horizontal-arm wave	High-arm wave	High throw
Hammer	Hand catch	Forward kick
Forward punch	Draw x	Side kick
High throw	Draw tick	Jogging
Hand clap	Draw circle	Tennis Serving
Bend	Two hand wave	Tennis swing
Tennis serve	Forward kick	Golf swing
Pickup & throw	Side-boxing	Pickup & throw

# Recognition Accuracy using 3D BoP

Action Set	1/3 samples as training	2/3 samples as training	½ subjects' samples as training
AS1	89.5%	93.4%	72.9%
AS2	89.0%	92.9%	71.9%
AS3	96.3%	96.3%	79.2%
overall	91.6%	94.2%	74.7%

# Comparison to 2D Silhouettes

- 2D silhouettes were obtained from the xy-projections
  - which is close to silhouettes from a 2D image
- 80 2D points were sampled from the contour of each 2D silhouette.
- Using
  - the same number of postures
  - the same number of Gaussian components and
  - the same number of training samples

# Recognition Accuracy using 2D Silhouettes

Action Set	1/3 samples as training	2/3 samples as training	½ subjects' samples as training
AS1	79.5%	81.3%	36.3%
AS2	82.2%	88.7%	48.9%
AS3	83.3%	89.5%	45.8%
overall	81.7%	86.5%	43.7%

*vs. 3D Bag of Points*

overall	91.6%	94.2%	74.7%
---------	-------	-------	-------

***Recognition with 3D is much more accurate!***

Zhou Ren, Junsong Yuan, Zhengyou Zhang

# **HAND GESTURE RECOGNITION**

## Challenges

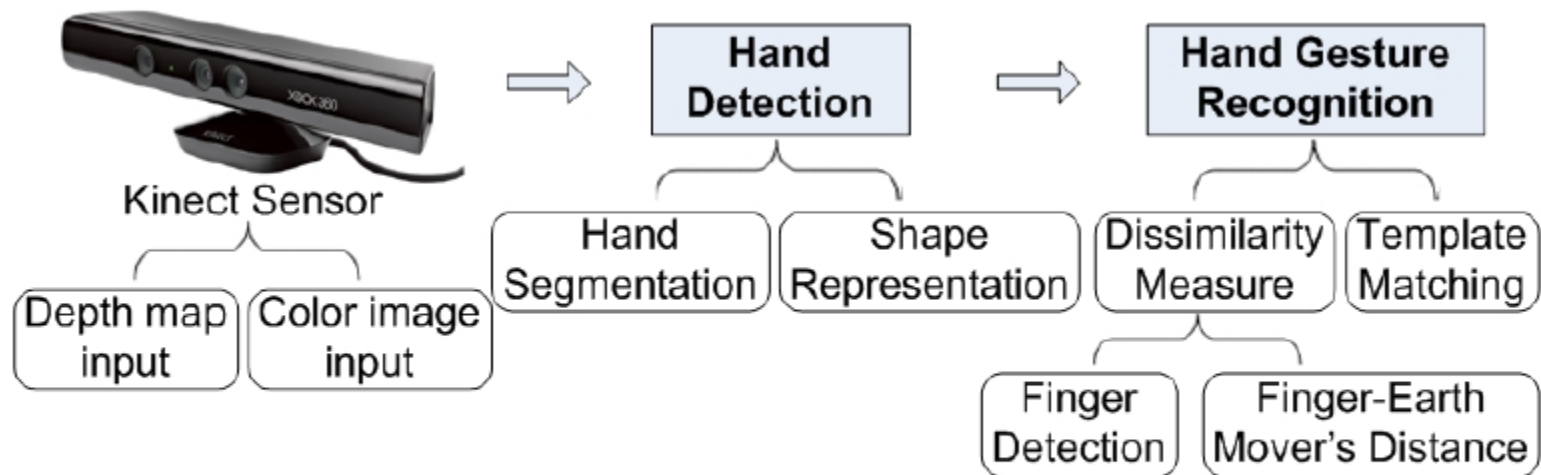


Figure 1: Some challenging cases for hand gesture recognition with depth cameras: the first and the second hands have the same gesture while the third hand confuses the recognition.

The resolution of depth map is low



## System of Kinect-based gesture recognition



**Figure 2:** The framework of our real-life hand gesture recognition system.

**Key Modules:** Hand segmentation and representation, Dissimilarity Measure (Finger Detection and FEMD)

# Hand Segmentation & Representation

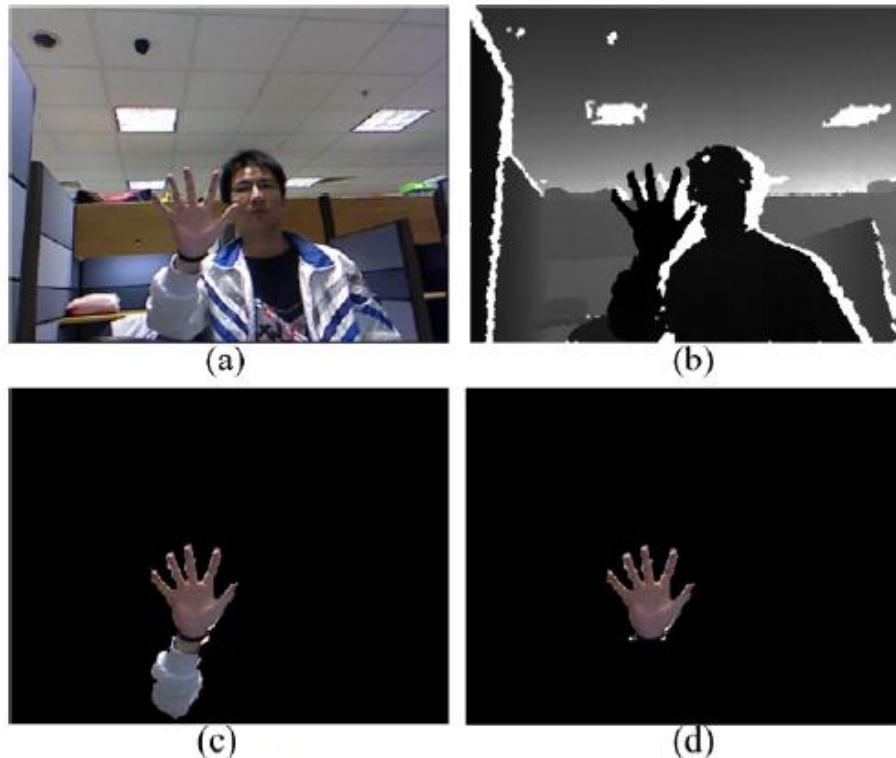


Figure 3: Hand segmentation process. (a). The RGB color image captured by Kinect Sensor; (b). The depth map captured by Kinect Sensor; (c). The area segmented using depth information; (d). The hand shape segmented using RGB information.

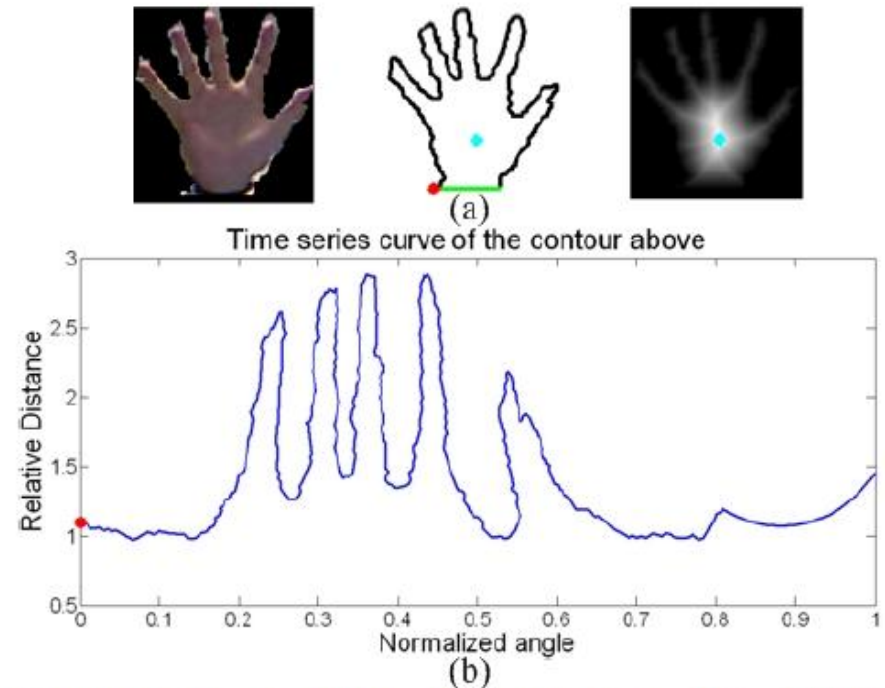


Figure 4: Hand shape representation. (a). On the contour of the segmented hand, the green line is the detection of the black belt; the red point is the initial point; the cyan point is the center point detected by Distance Transform; (b). The time-series curve of the shape above.

# Finger Detection via shape decomposition

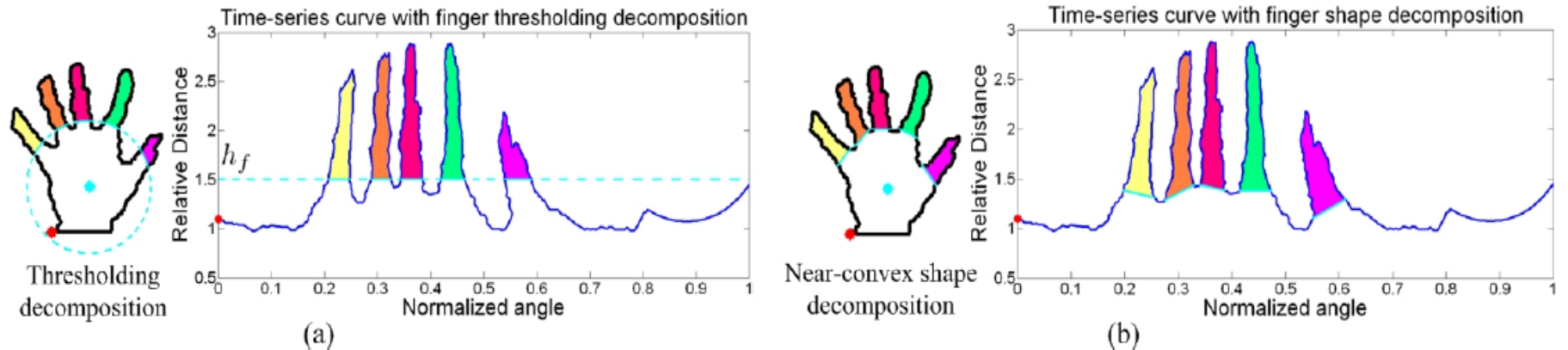


Figure 6: Illustration of the two proposed finger detection methods: (a). Thresholding decomposition uses a height threshold  $h_f$  in the time-series curve to detect fingers, which means to decompose the shape with a circle, thus information is inevitably lost; (b). Near-convex decomposition decomposes the hand into several near-convex parts that are fingers and the palm. The finger decomposition of (b) is more accurate and robust.

$$\begin{aligned} \min \quad & \alpha \| \mathbf{x} \|_0 + (1 - \alpha) \mathbf{w}^\top \mathbf{x}, \\ \text{s.t.} \quad & \mathbf{A} \mathbf{x} \geq \mathbf{1}, \quad \mathbf{x}^\top \mathbf{B} \mathbf{x} = 0, \quad \mathbf{x} \in \{0, 1\}^{\overline{n}} \end{aligned}$$

# Distance Metric: Finger-Earth Mover's Distance

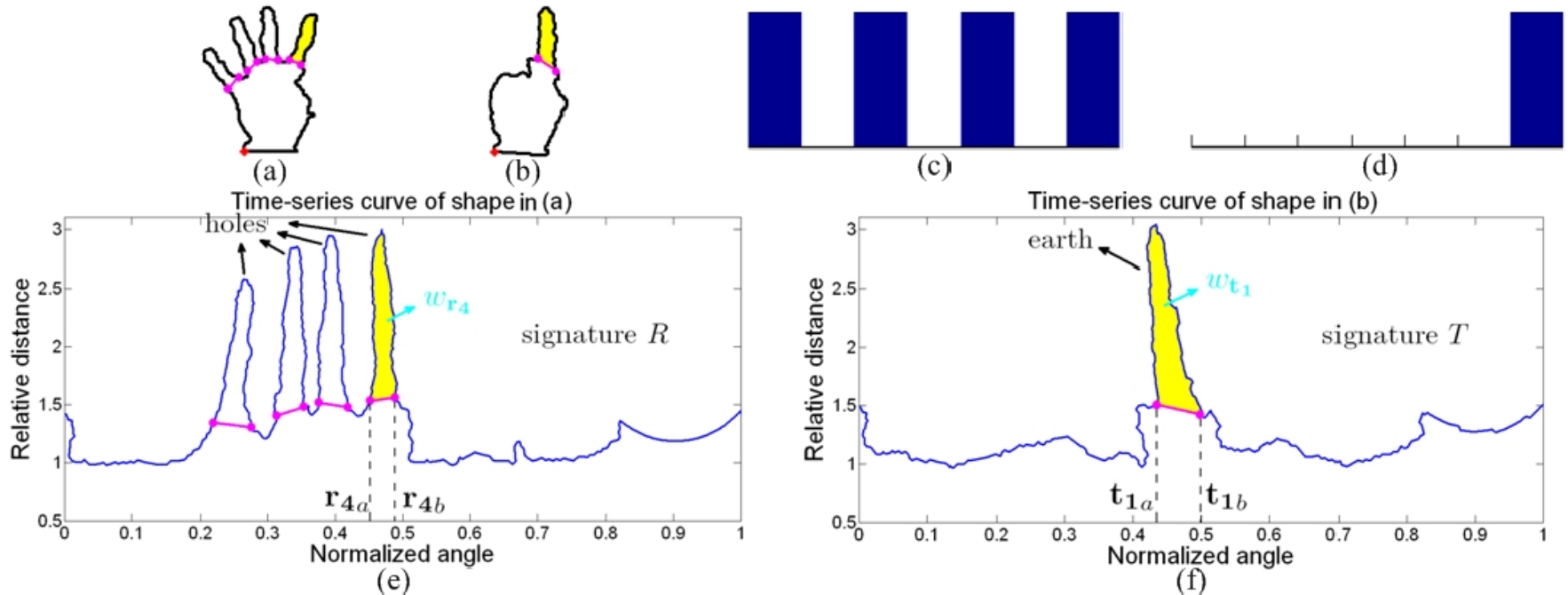


Figure 5: The motivation of using Finger-Earth Mover's Distance. (a) and (b) are two different hand shapes, whose time-series curves are shown in (e) and (f), respectively. Their major difference is the fingers. (c) and (d) are two signatures that partially match, their EMD cost is 0, however they are very different. Hence FEMD adds the penalty on empty holes. (e) and (f) are the time-series curves of the hand shapes in (a) and (b), each curve is represented as a signature with each finger as a cluster; the signature with bigger total weight serves as holes, the smaller one serves as earth piles.

FEMD vs. EMD: 1. consider global feature (finger);  
2. alleviate partial matching



# Results

- New collected dataset with Kinect camera:

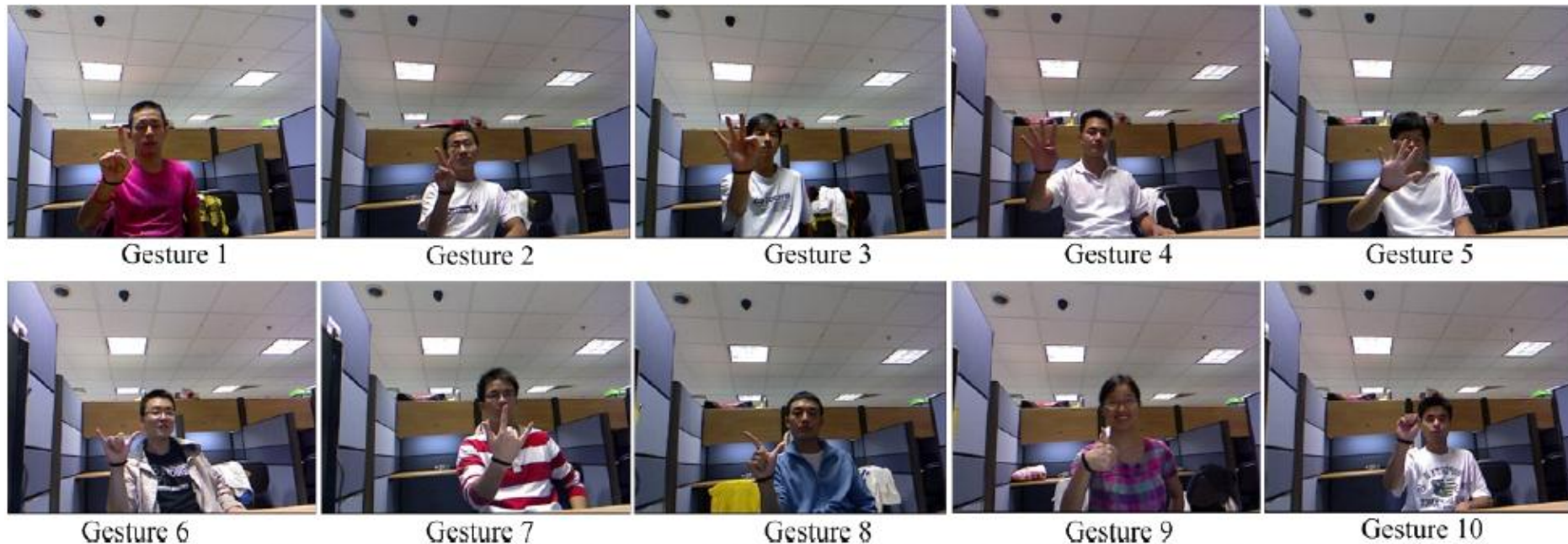


Figure 8: The color image examples for the 10 gestures in our dataset.

10 subjects \* 10 gestures/subject \* 10 cases/gesture = 1000 cases  
Contain color image and depth map  
Under uncontrolled environment

## Accuracy and efficiency

	Thresholding Decomposition+FEMD	Near-convex Decomposition+FEMD
Mean Accuracy	90.6%	93.9%
Mean Running Time	0.5004s	4.0012s

Table 1: The mean accuracy and the mean running time of the two proposed methods.

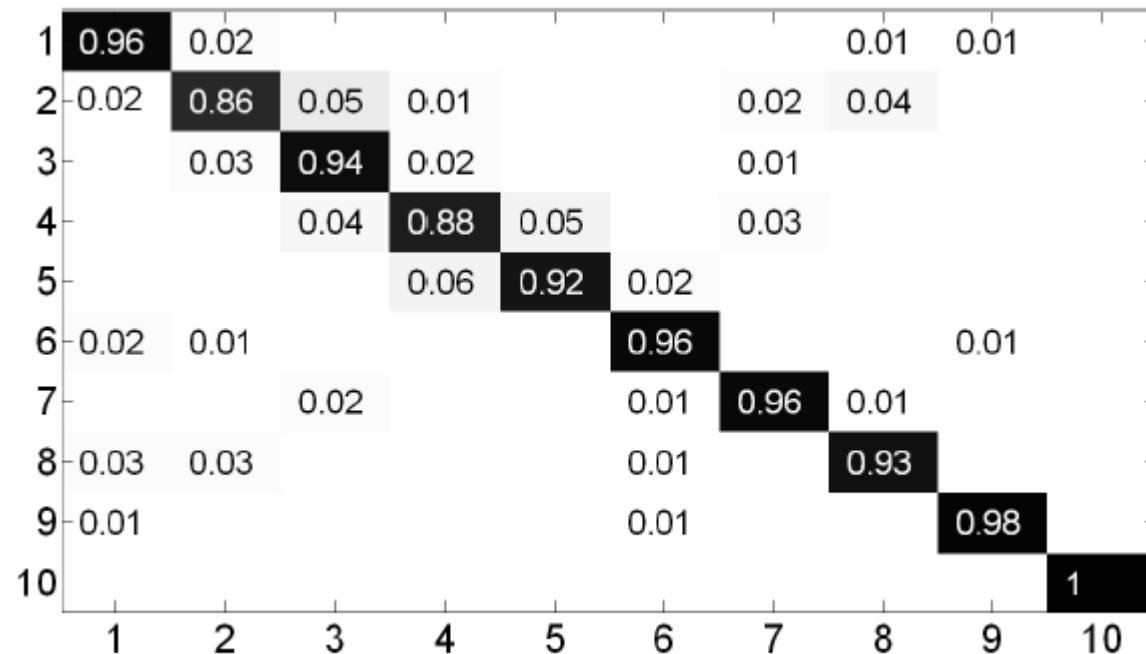


Figure 15: The confusion matrix of Experiment II.



NANYANG  
TECHNOLOGICAL  
UNIVERSITY

Microsoft  
**Research**



# Robust Hand Gesture Recognition with Kinect Sensor

Zhou Ren, Jingjing Meng, Junsong Yuan, Zhengyou Zhang  
School of EEE, NTU, Singapore & Microsoft Research, Scottsdale, USA

*Proc. of ACM Intl. Conf. on Multimedia 2011*

Qin Cai, Cha Zhang, Zhengyou Zhang

# **FACE MODELING**



# Face Model Enhances 3D Scene

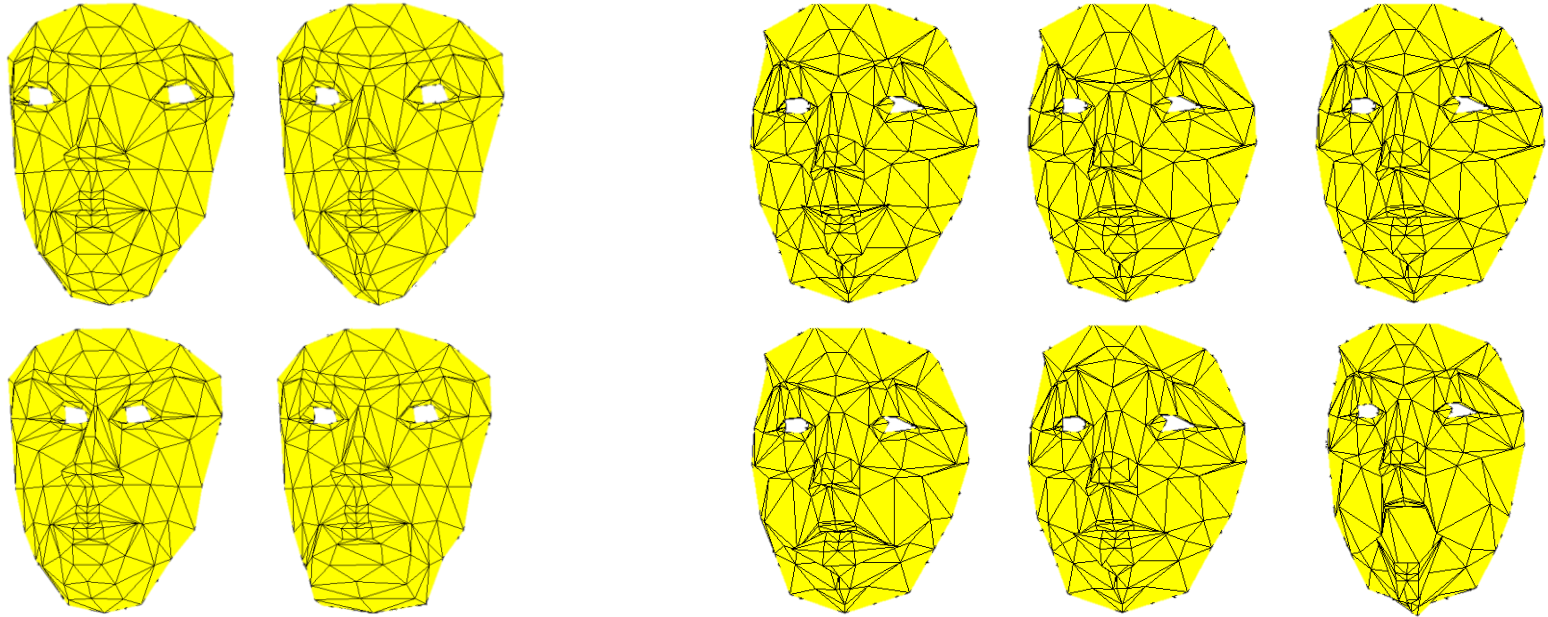
Qin Cai, Cha Zhang, Zhengyou Zhang

# **HEAD POSE & FACIAL EXPRESSION TRACKING**

# Deformable Face Tracking

- Many applications
  - Human computer interaction
  - Performance-driven facial animation
  - Face recognition
- Challenging
  - Limited number of features on the face
  - Dozens of parameters to estimate

# Linear Deformable Model



Static deformations

Action deformations

(**Artist rendered** linear deformable model)

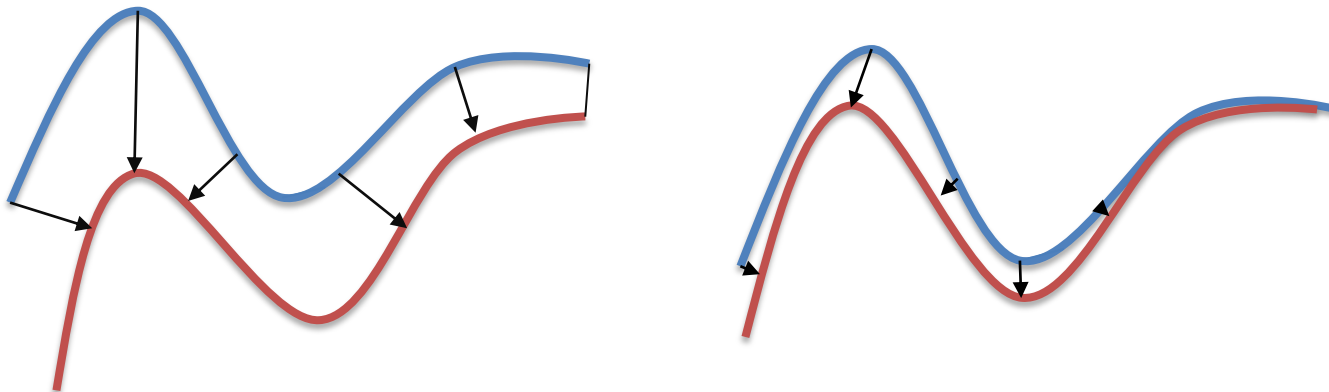
$$\begin{bmatrix} q_1 \\ \vdots \\ q_K \end{bmatrix} = \begin{bmatrix} p_1 \\ \vdots \\ p_K \end{bmatrix} + \mathbf{A} \begin{bmatrix} r_1 \\ \vdots \\ r_K \end{bmatrix} + \mathbf{B} \begin{bmatrix} s_1 \\ \vdots \\ s_K \end{bmatrix}, \text{ where } \mathbf{A} = \begin{bmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_K \end{bmatrix}, \mathbf{B} = \begin{bmatrix} \mathbf{B}_1 \\ \vdots \\ \mathbf{B}_K \end{bmatrix}$$

# Maximum Likelihood DMF

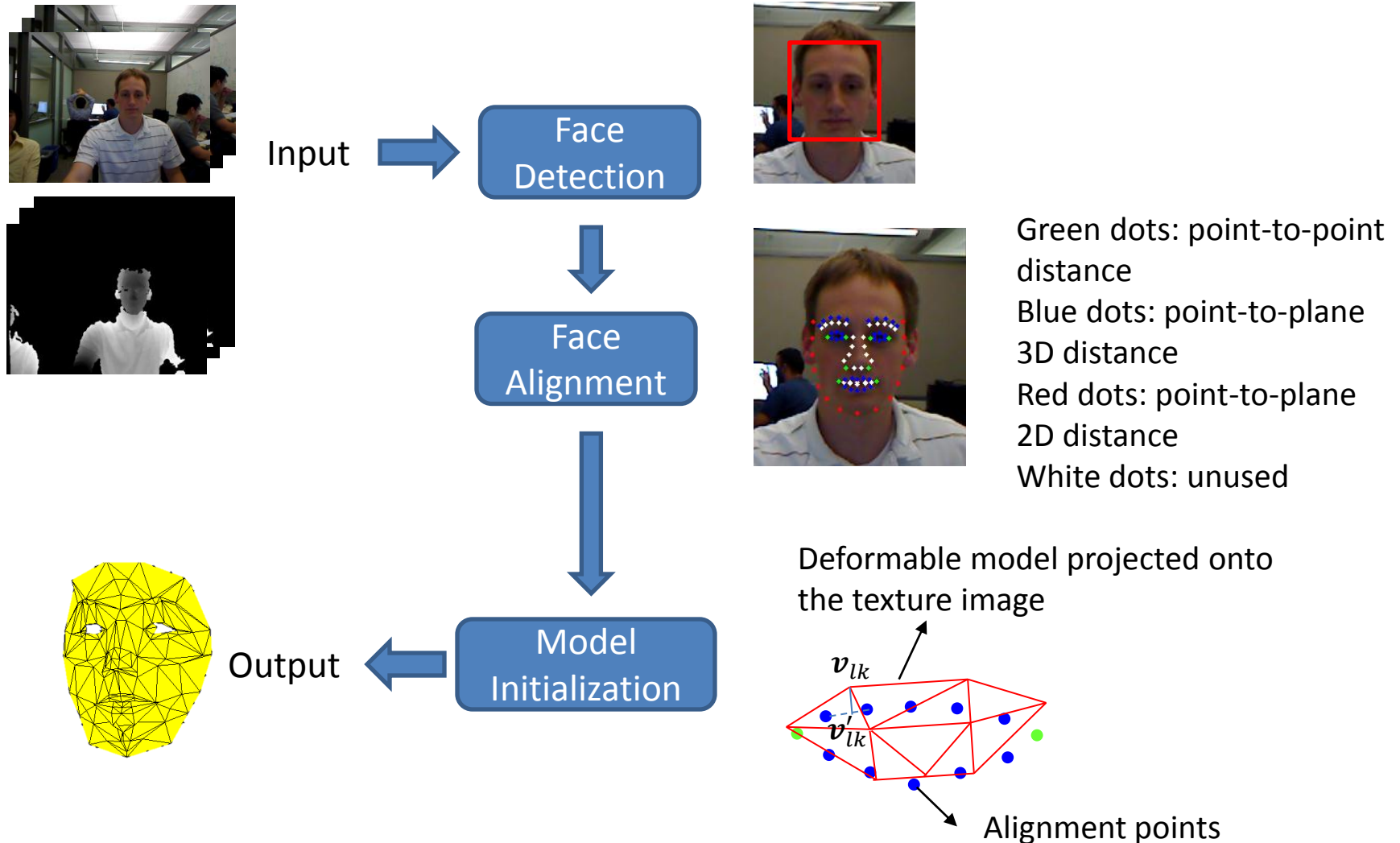
- Formulation,  $(\mathbf{q}_k, \mathbf{g}_k)$  correspondence pair:

$$\mathbf{R}(\mathbf{p}_k + \mathbf{A}_k \mathbf{r} + \mathbf{B}_k \mathbf{s}) + \mathbf{t} = \mathbf{g}_k + \mathbf{x}_k$$
$$\mathbf{x}_k \sim N(\mathbf{0}, \Sigma_{\mathbf{x}_k})$$

- Iterative closest point
  - Assume closest points correspond
  - Compute transformation
  - Iterate until convergence



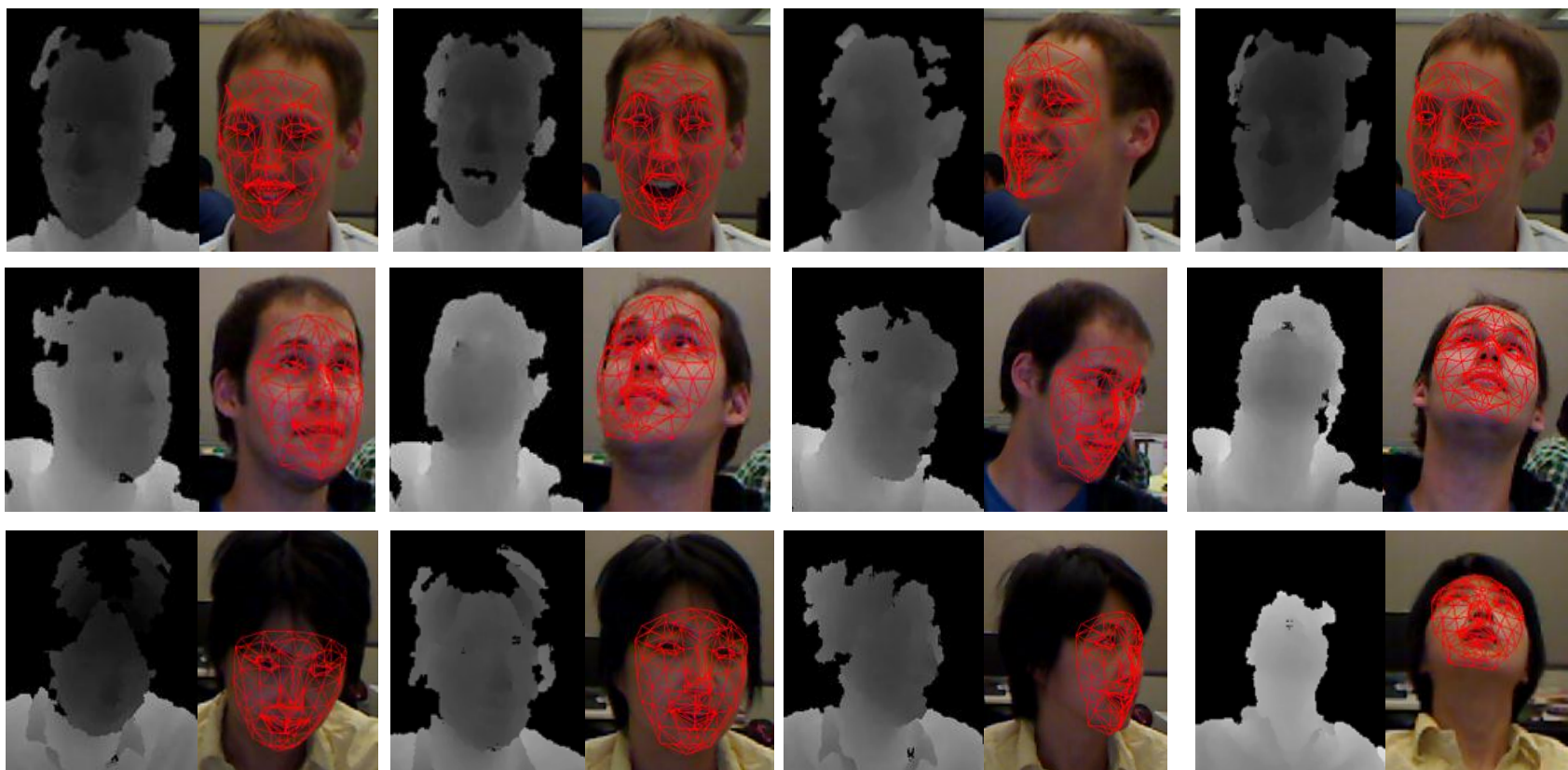
# Model Initialization



# Face Tracking

- Tracking
  - Shape deformations fixed
  - Based on feature point correspondence
  - Solve for action deformation, rotation and translation
  - Regularization
    - $l_2$  norm constraining the difference between neighboring frames' action deformations
    - $l_1$  norm constraining the number of non-zero action deformation parameters

# Tracking Results: [Video](#)



Top to bottom: Seq #1 (810 frames), Seq #2 (681 frames), Seq #3 (300 frames)



# Qualitative Results

Median tracking error in pixels

	ID+ $l_2$	ID+ $l_1$	ID+ $l_2+l_1$	NM+ $l_2$	NM+ $l_1$	NM+ $l_2+l_1$
Seq #1	3.56	2.88	2.78	2.85	2.69	2.66
Seq #2	4.48	3.78	3.71	4.30	3.64	3.55
Seq #3	3.98L	3.91	3.91	3.92L	3.91	3.50

ID: use identity covariance matrix for sensor noise

NM: use the proposed noise modeling scheme

$l_2$ : quadratic constraint between successive frames

$l_1$ : sparse constraint on the action transforms

**L**: lost tracking in the middle and never recover

Phil Chou, Niru Chandrasekaran, Qin Cai, Cha Zhang, Zhengyou Zhang

# **TELE-IMMERSION**

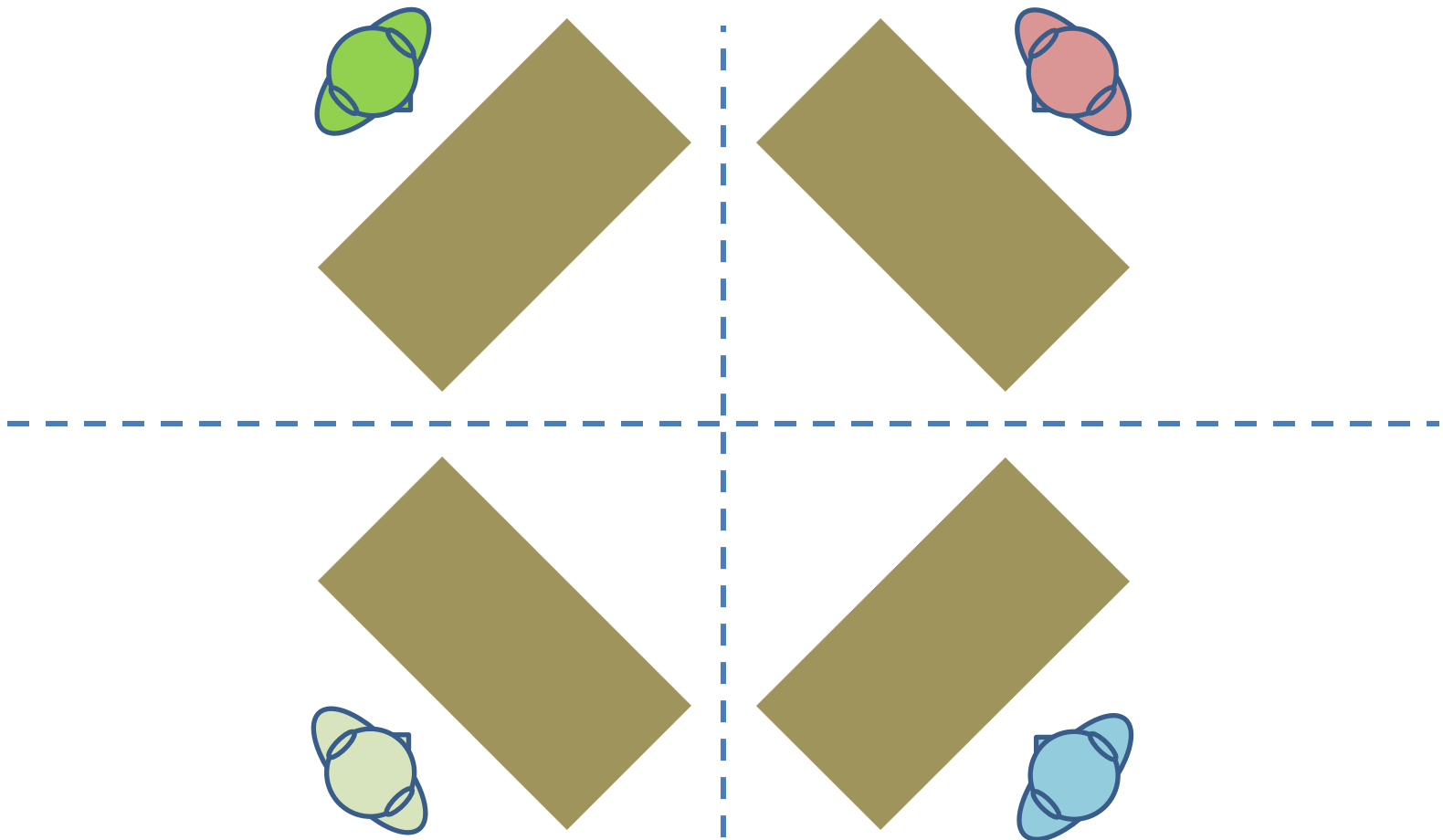
# Future of Human Communication

- Economy globalization
  - Geographically distributed work team
  - Travel cost
  - Work-life balance
- 
- Hardware cost dropped
  - Bandwidth increased
  - Multimedia technology advanced

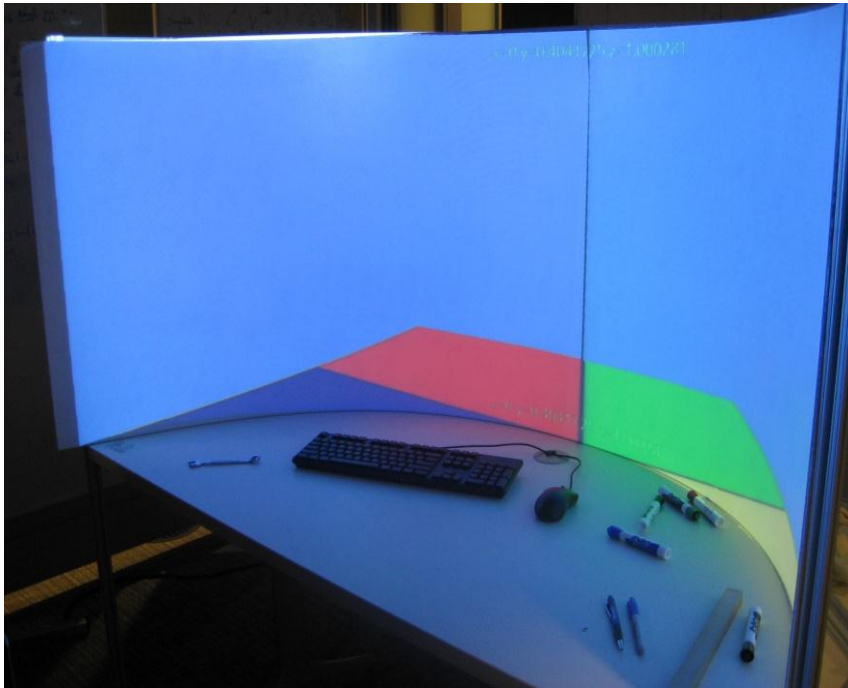
# Tele-immersion

- Geographically distributed participants feel like they are in the same room
- Tele-immersion experience
  - Life size
  - Mutual gaze
  - 3D
  - Motion parallax
  - Spatial audio

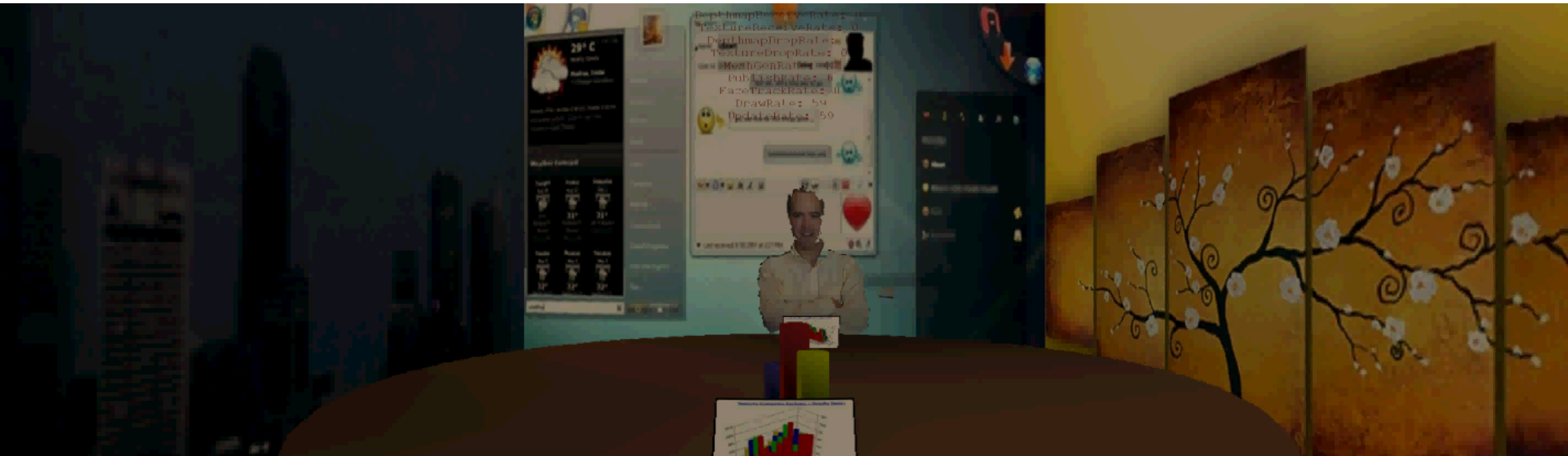
# Virtual (“Matrix”) approach to fully distributed meetings



# Tele-Immersion Booth



# Tele-Immersion Booth video



# References

- Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake, "Real-Time Human Pose Recognition in Parts from a Single Depth Image", in *Proc. CVPR*, June 2011.
- W. Li, Z. Zhang, and Z. Liu, "Expandable Data-Driven Graphical Modeling of Human Actions Based on Salient Postures", *IEEE Transaction on Circuits and Systems for Video Technology*, Vol.18 No.11, pages 1499-1510, 2008.
- W. Li, Z. Zhang, and Z. Liu, "Action Recognition Based on A Bag of 3D Points", in *Proc. IEEE International Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB)*, pages 9-14, San Francisco, CA, USA, June 18, 2010.
- Z. Ren, J. Yuan, and Z. Zhang, "Robust Hand Gesture Recognition Based on Finger-Earth Mover's Distance with a Commodity Depth Camera", in *Proc. ACM International Conference on Multimedia (ACM MM)*, Scottsdale, Arizona, USA, Nov. 28--Dec. 1, 2011. To Appear.
- Q. Cai, A. Sankaranarayanan, Q. Zhang, Z. Zhang, and Z. Liu, "Real Time Head Pose Tracking from Multiple Cameras with a Generic Model", in *Proc. IEEE Workshop on Analysis and Modeling of Faces and Gestures*, pages 25-32, June 2010.
- Q. Cai, D. Gallup, C. Zhang, and Z. Zhang, "3D Deformable Face Tracking with a Commodity Depth Camera", in *Proc. European Conference on Computer Vision (ECCV)*, Vol. III, pages 229--242, Crete, Greece, Sep. 2010.



# Acknowledgment

- Qin Cai
- Phil Chou
- Anoop Gupta
- Zicheng Liu
- Cha Zhang
- Niru Chandrasekaran
- Jamie Shotton
- Wanqing Li
- Zhou Ren
- Junsong Yuan

**CHALLENGES**

# Challenges (1)

- Model human body language
  - Facial expression
  - Head gesture
  - Hand gesture
  - Body gesture
  - Motion dynamics
  - Behaviors
  - Human-human interaction
  - ...

# Challenges (2)

- Improve sensor quality
  - Short range vs. Long range
  - Day vs. Night
  - Indoor vs. Outdoor
  - Different surface materials
- Model sensor imprecision
- Fuse multiple sensors

# Challenges (3)

- Develop efficient and robust algorithms
  - Deal with various challenging situations
  - Process a large amount of data
  - Handle inter-/intra- person variations
  - Collect and label large-scale training/test datasets
  - ...
- Understand societal implications
  - E.g. Privacy

