Capture, Analysis, and Applications of 3D Visual Signals

Zhengyou Zhang
Research Manager/Principal Researcher
Microsoft Research
zhang@microsoft.com
http://research.microsoft.com/~zhang/

Microsoft Research

Microsoft Kinect Sensor

http://www.kinecthacks.net/ | portseattle.org | xbox.com | KinectHacks.n...

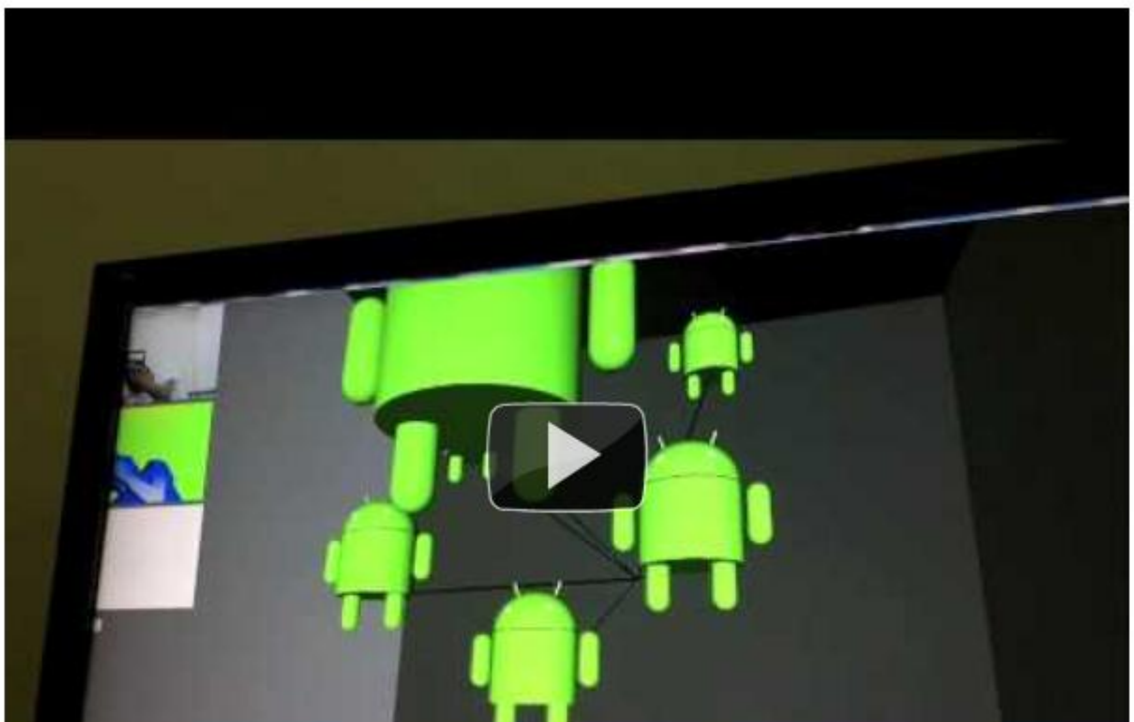
KINECTHACKS

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

Like

Dan Graham Eric Stephan Lawrence

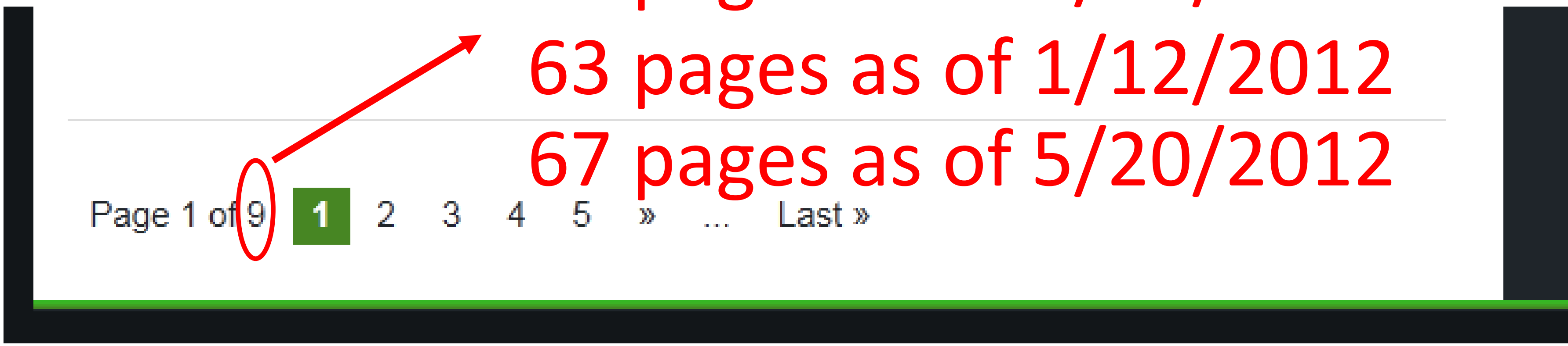
Facebook social plugin

~ 90 Projects as of 12/04/2010

24 pages as of 2/10/2011

63 pages as of 1/12/2012

67 pages as of 5/20/2012



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

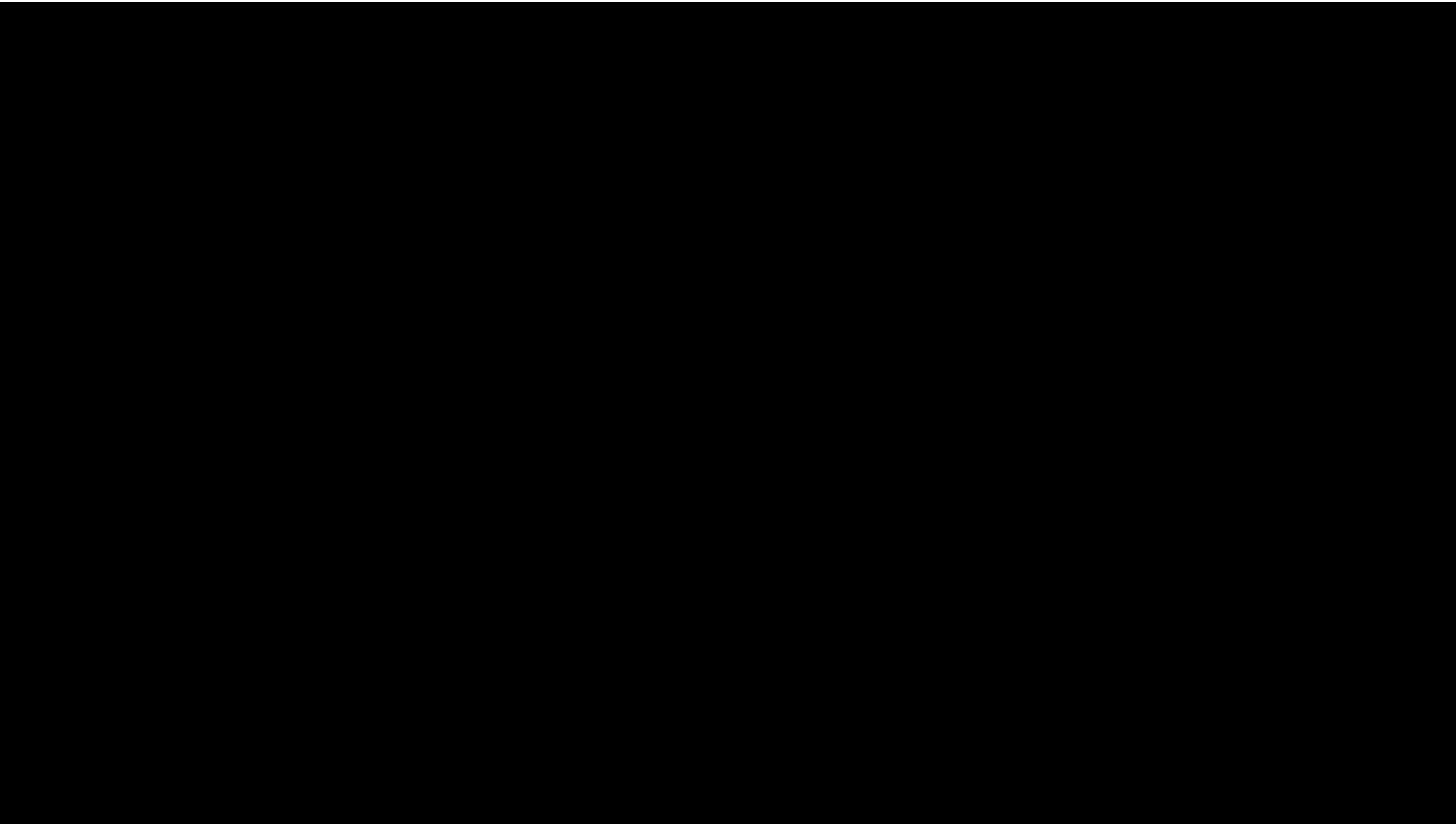
- Quote from KinectHacks.net

http://www.youtube.com/watch?v=7QrnwoO1-8A

3D Video Capture

http://www.youtube.com/watch?v=VMWc2KPFFv4

Music Video



http://www.youtube.com/watch?v=l6QY-eb6NoQ

Navigational Aids for the Visually Impaired

Kinect for Windows SDK

www.microsoft.com/en-us/kinectforwindows

- Access to deep Kinect system information
  - Depth data, **near mode**
  - **Synchronized depth and RGB streams**
  - Audio
  - Direct control of the Kinect sensor
  - System API
  - Skeletal tracking, **sitting** or **standing up**
  - Voice command

Sample App: Shape Game

Human stereo vision

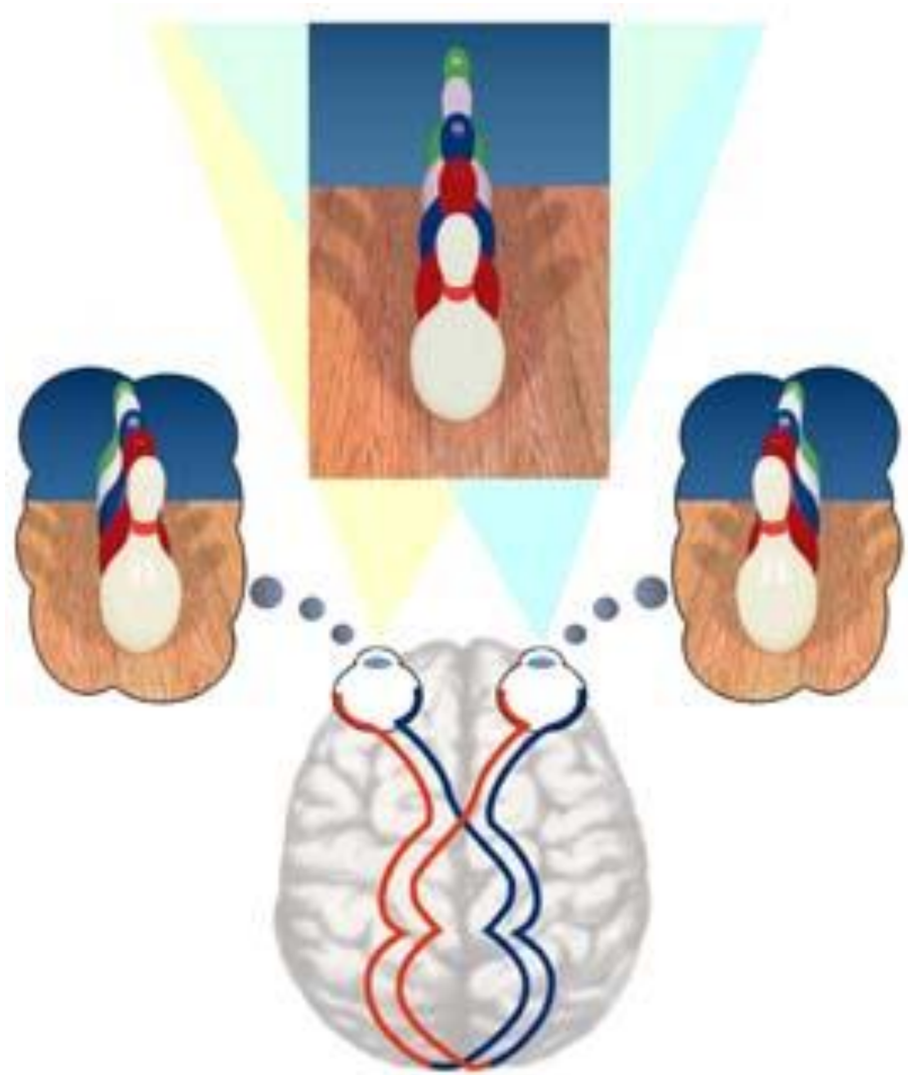
Computer stereo vision

Kinect sensing technology

HOW IT WORKS ?

Human Stereo Vision

Difference
in your two eyes
gives you the ability
to perceive
your surrounding
environment
in **3 Dimensions**



http://www.vision3d.com/stereo.html

Key to Perceive in 3D

- See two different views
- Match similarity between the two views
- Fuse them to reconstruct the scene in 3D

To see 3D,

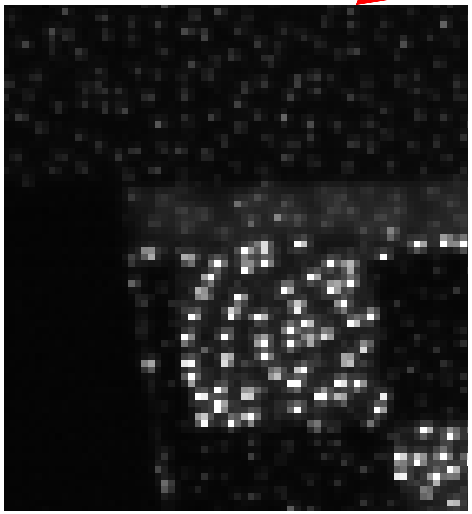
- Your two eyes must work **simultaneously**
- Your **brain** is able to fuse the two views

- At least 12% of people have some problem with their stereo vision

http://www.vision3d.com/whycant.html

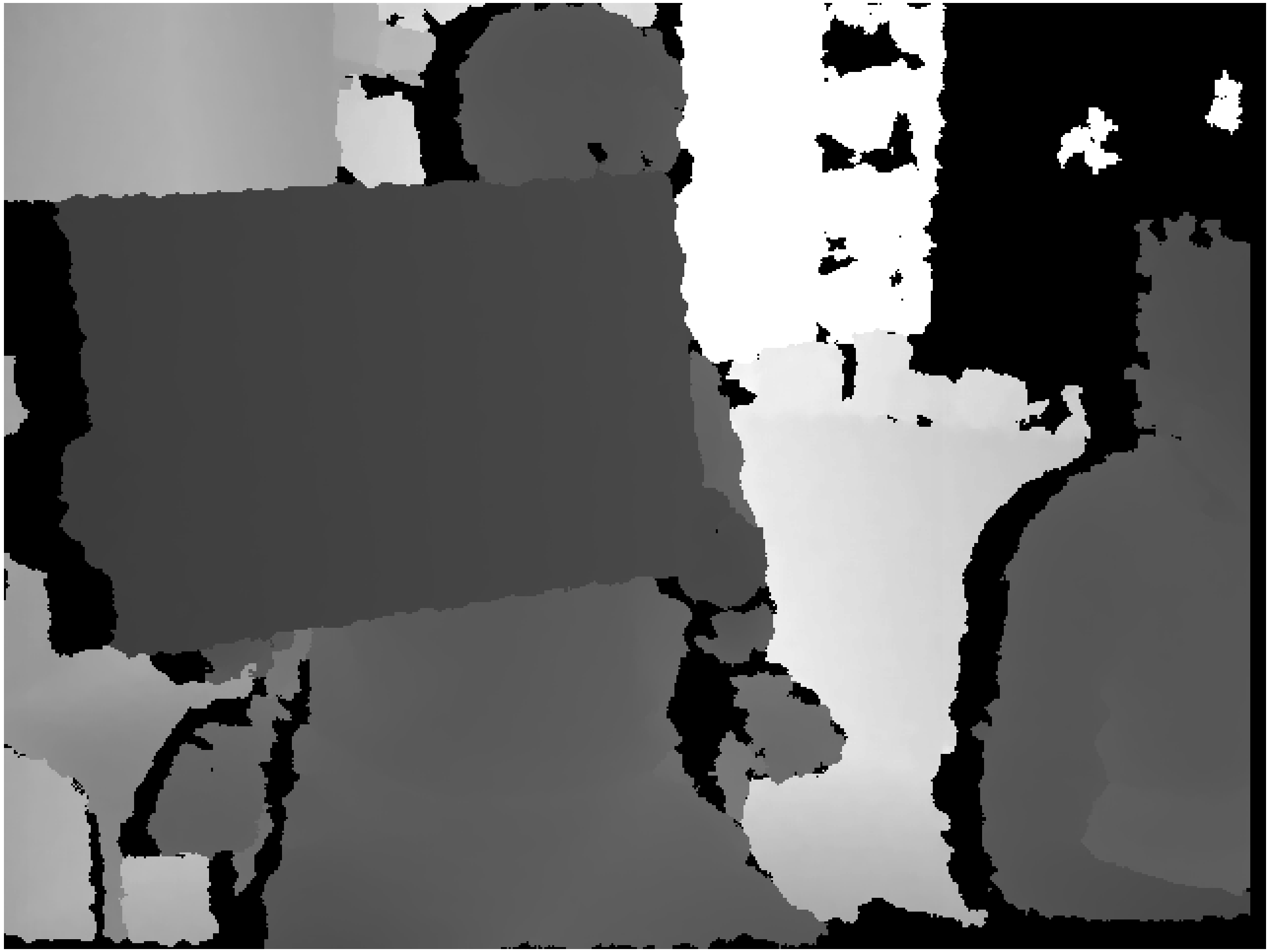
How it works? **Kinect Sensor**

- Modified structured light 3D scanner
  - IR projector
  - IR camera
  - Random pattern

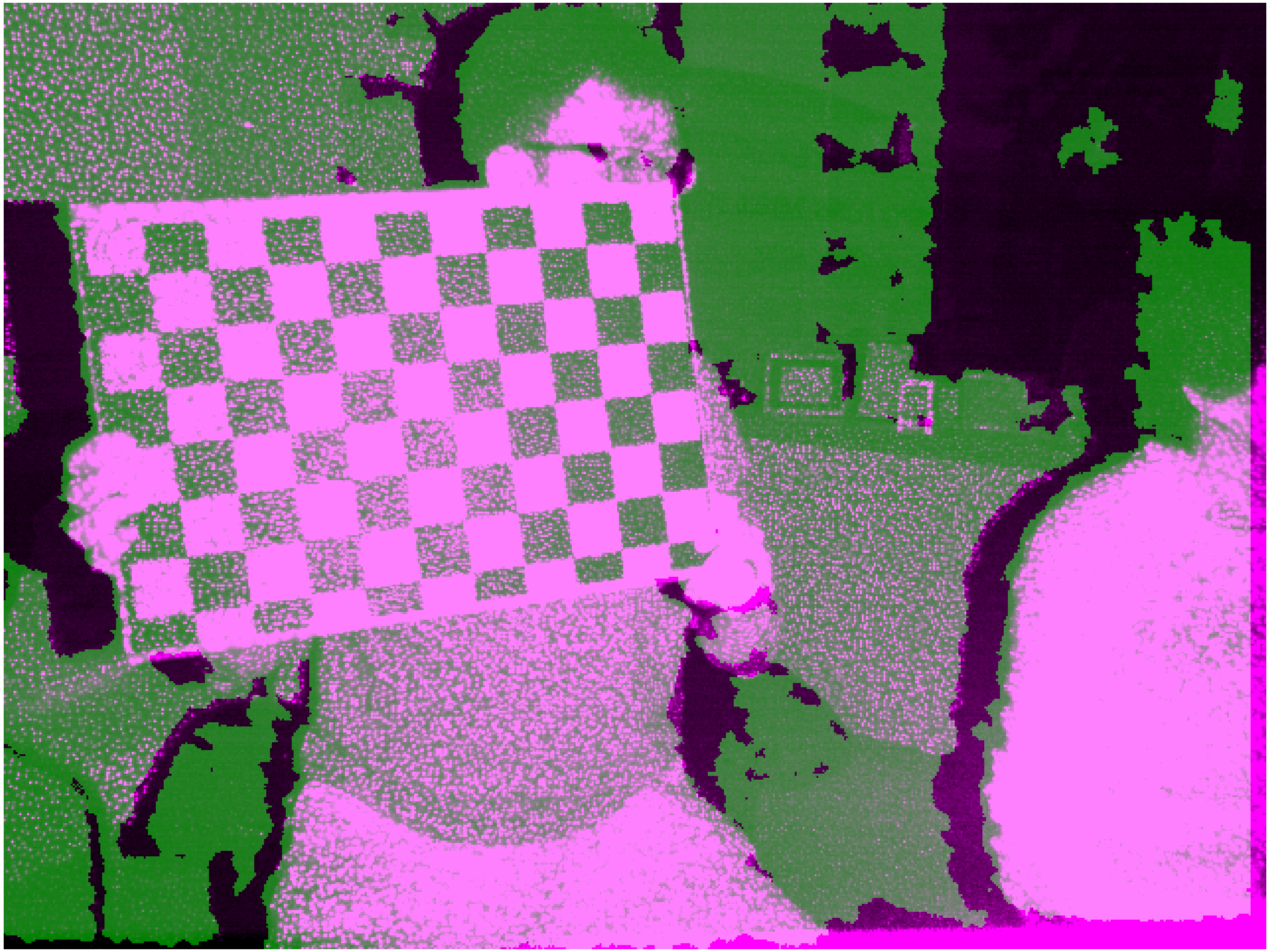


Matching & Depth Map

- Correlation



Overlay of Depth Map on IR Image



Kinect Calibration

- The Kinect calibration card is used to recalibrate your sensor in the event the sensor is not properly tracking your body. The card is included in the Kinect Adventures games.

RGB vs. Depth Sensors

RGB

☒ Only works well lit

☒ Background clutter

☒ Scale unknown

much easier
with depth!

DEPTH

☑ Works in low light

☑ Person 'pops' out from bg

☑ Scale known

☒ Shadows, missing pixels

Challenges

- Noisy data
  - How to characterize the uncertainty?
  - How to deal with the sensor inefficiency (e.g., non-IR-reflective surface, environment with strong ambient IR)?
- Partial data
  - How to fuse multiple views?
  - How to deal with interference between multiple sensors?
  - How to leverage visual sensors?
- Raw data
  - How to infer high-level/semantic information?
- Multimodal data
  - How to collaborate with audio, tactile, inertial sensors to create compelling applications?

Video

HUMAN BODY-LANGUAGE UNDERSTANDING

Human Body Language

- A form of non-verbal communication
  - Body posture
  - Gestures
  - Facial expressions
  - Eye movements (eye gaze)
- Humans send and interpret such signals ***almost entirely subconsciously***

Body Language

Communicator	Mode	Control	Concordance
Send out	Facial expression	Voluntary	In concordance
Receive	Body movement	Involuntary	Discordance (lie)
	Tone of voice		

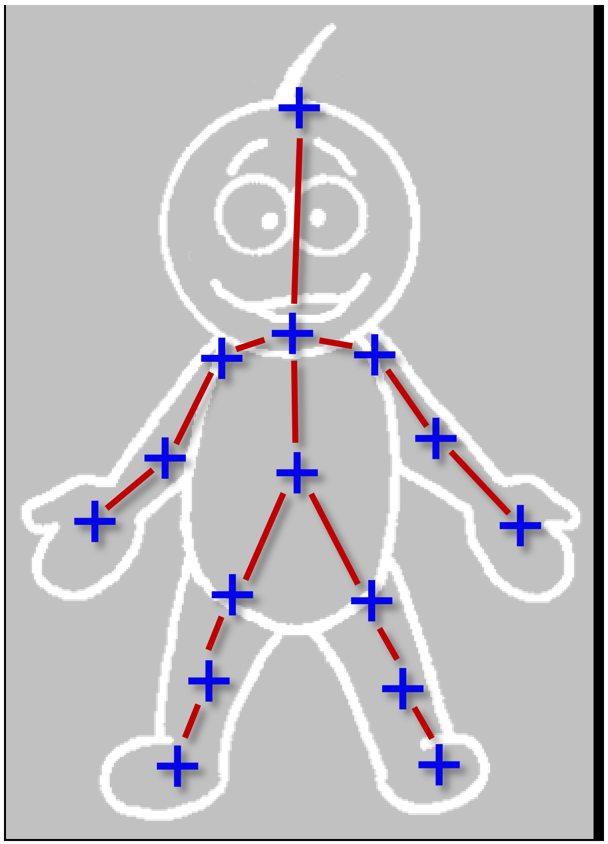
Outline

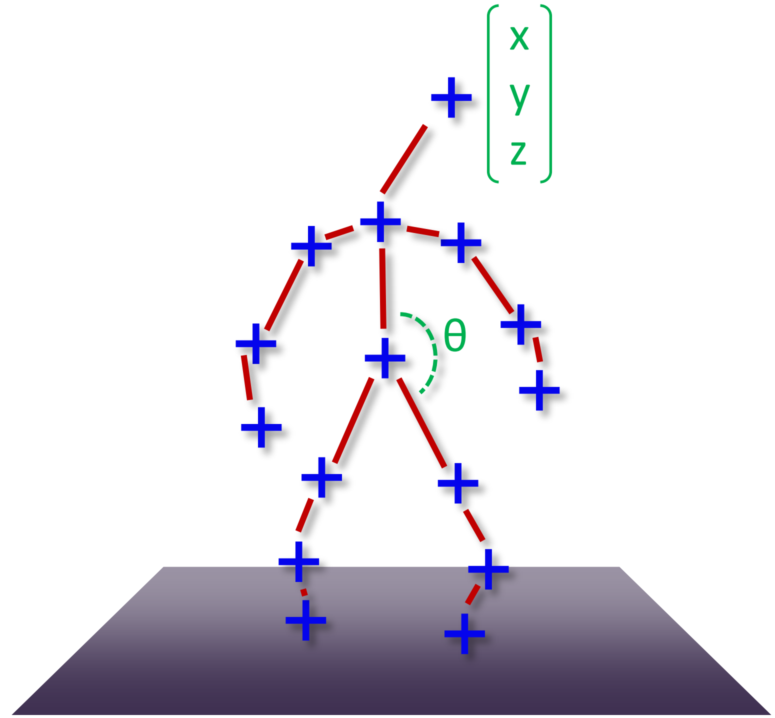
- Skeletal tracking
- Human action recognition
- Hand gesture recognition
- Head pose and facial expression tracking

Jamie Shotton, Andrew Blake, Kinect Team

SKELETAL TRACKING

Human pose estimation

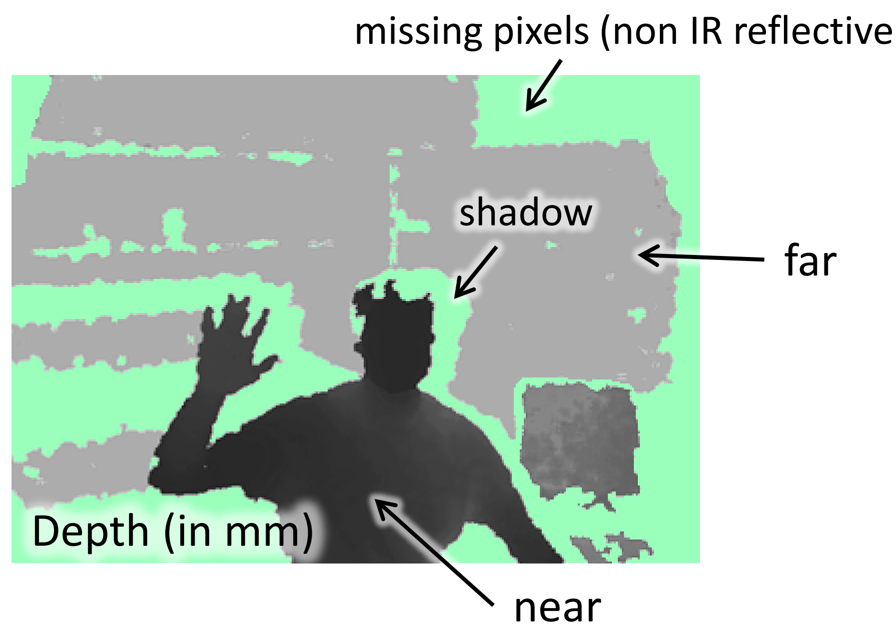




Kinect tracks 20 body joints in real time.

Depth cameras

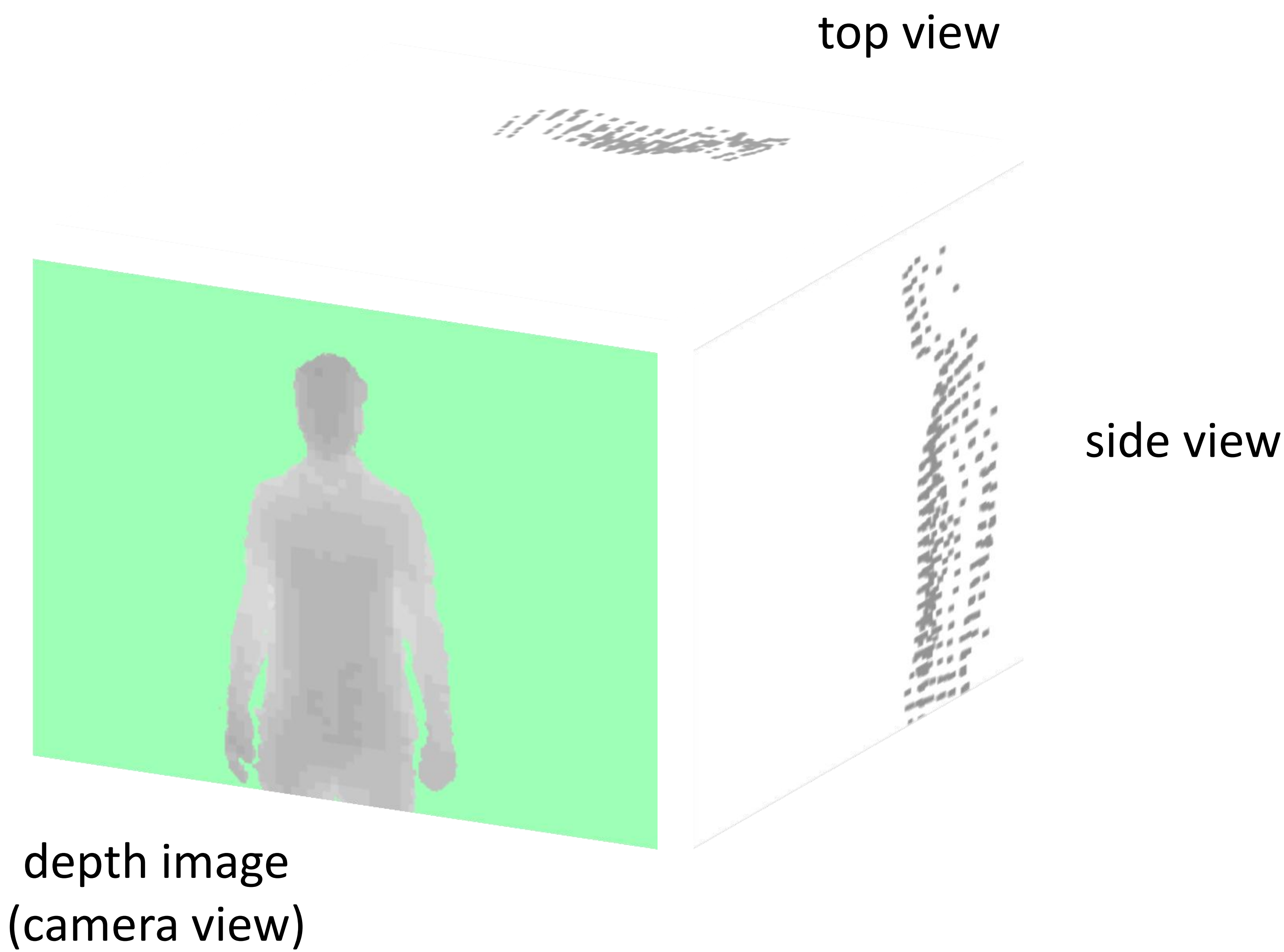
- Technology
  - structured IR light



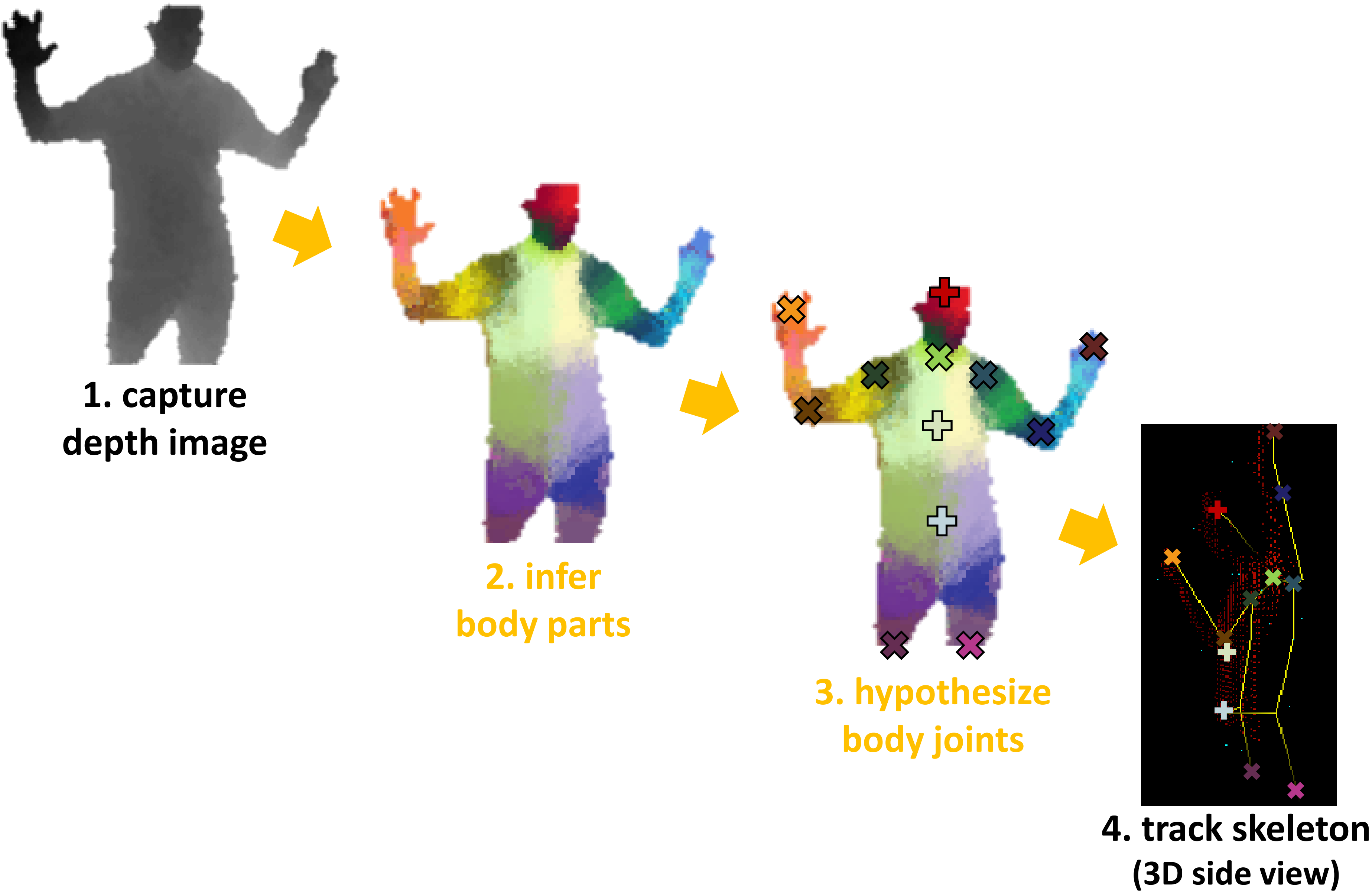
☑ cheap, fast, accurate

☒ missing pixels, shadows

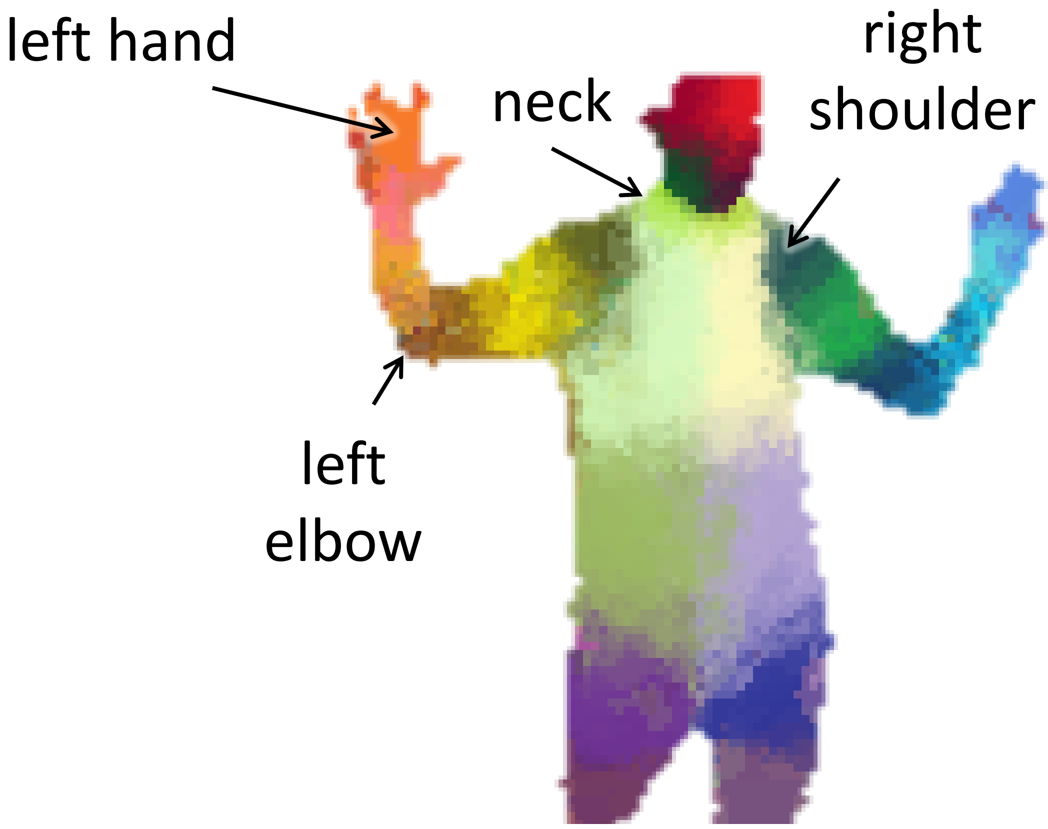
Depth cameras



The Kinect pose estimation pipeline



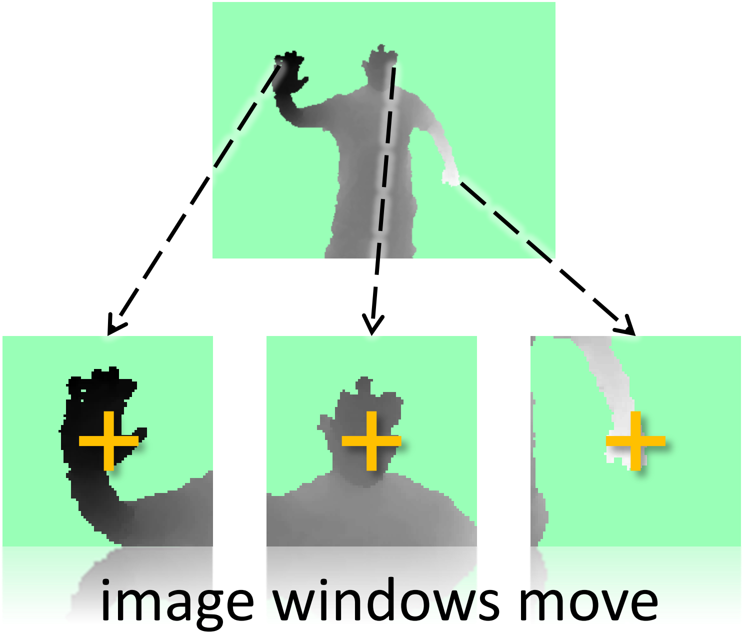
Body part recognition



("left" = player left with camera acting as mirror)

Classifying pixels

- Compute $\mathrm{P}(c_i \mid w_i)$
  - pixels $i = (x, y)$
  - body part c_i
  - image window w_i



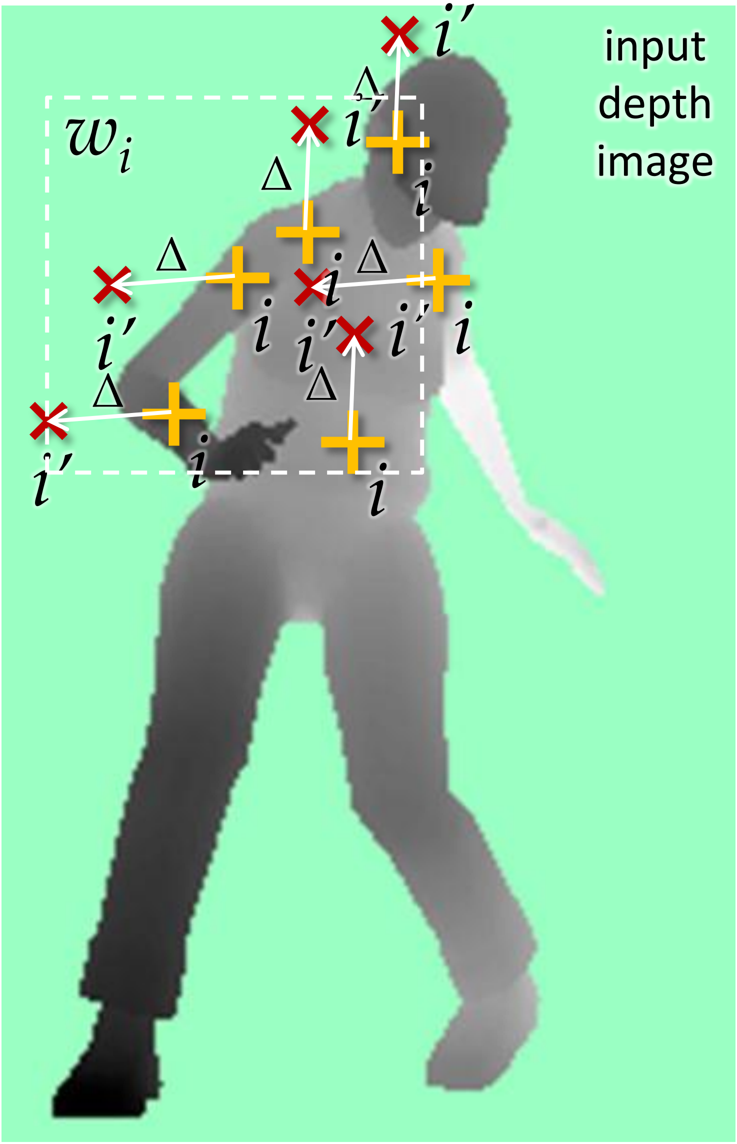
- Discriminative approach

- Learn classifier $\mathrm{P}(c_i \mid w_i)$ from training data

Fast depth image features

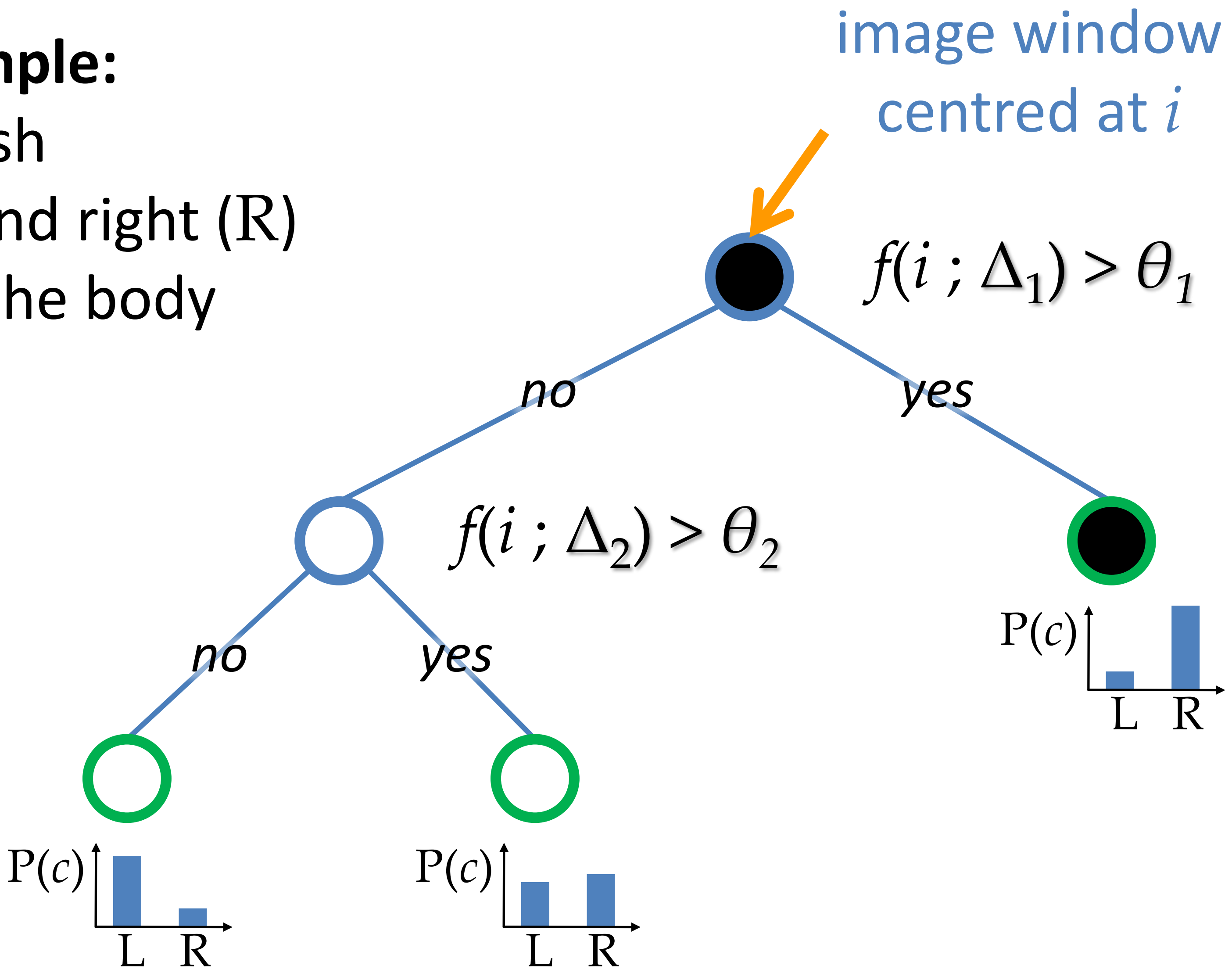
- Depth comparisons:
  - $f(i\,;\Delta) = d(i) - d(i')$
    where $i' = i + \Delta$
- Background pixels
  - d = large constant


desired
body parts

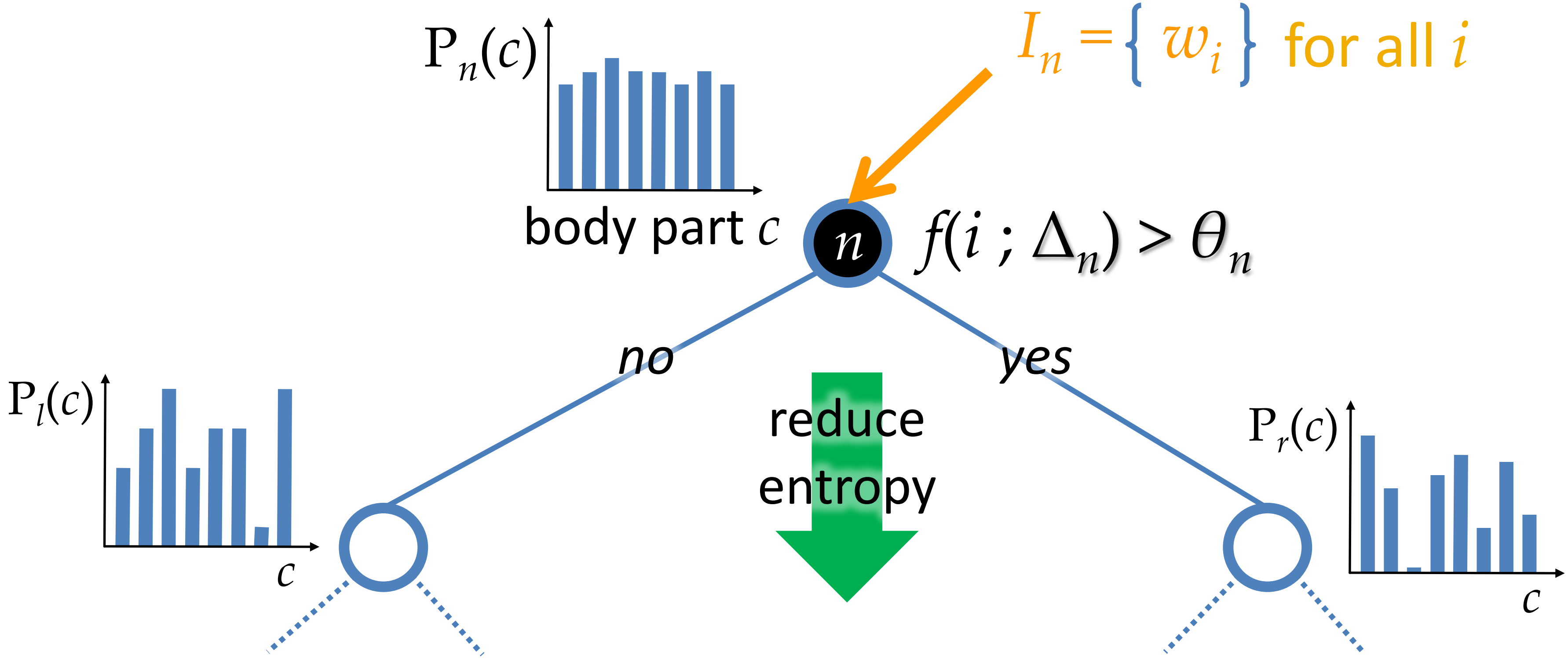


Decision tree classification

Toy example:
distinguish
left (L) and right (R)
sides of the body



Training decision trees [Breiman *et al.* 84]

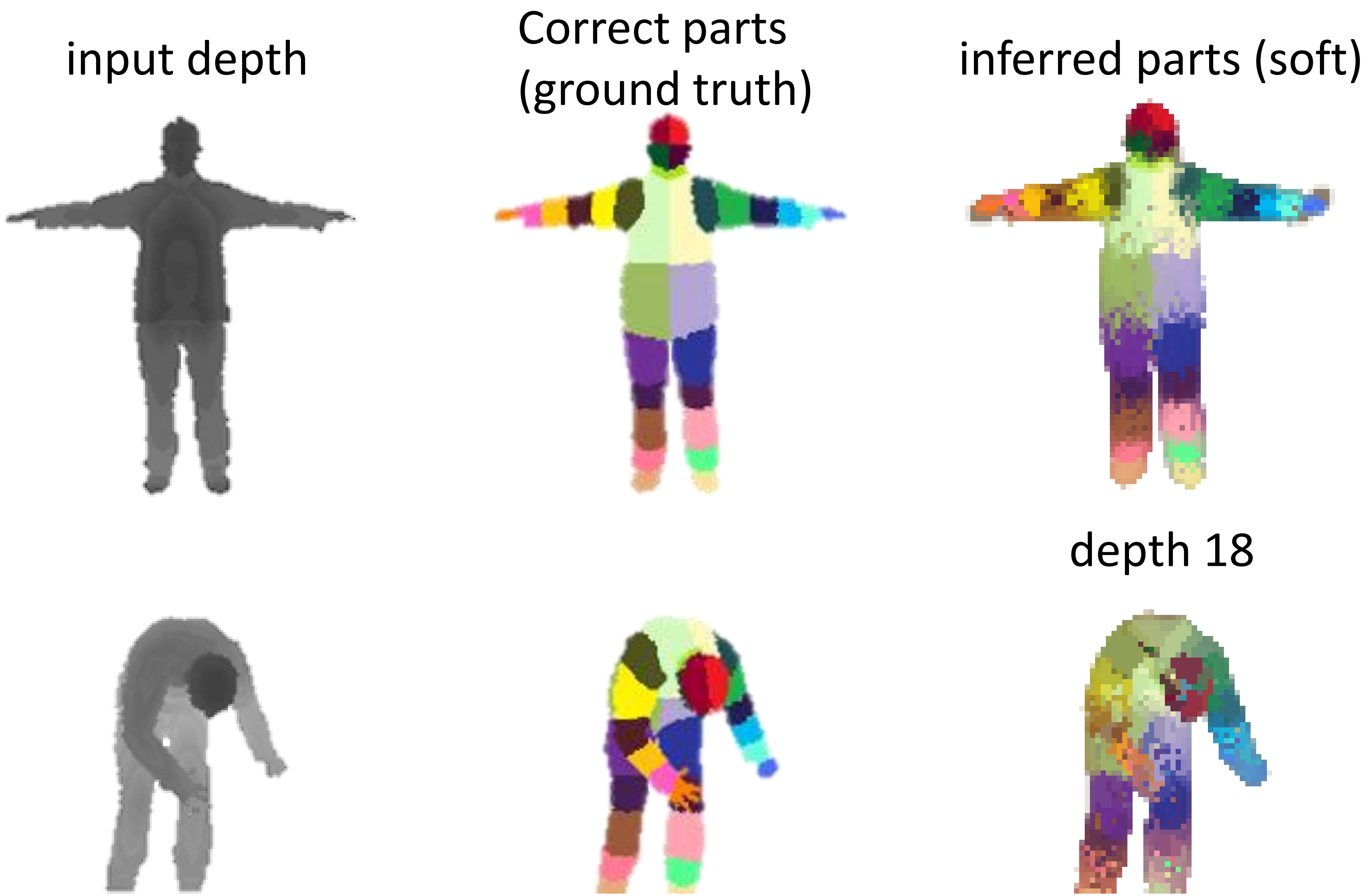


Take (Δ, θ) that maximises information gain:

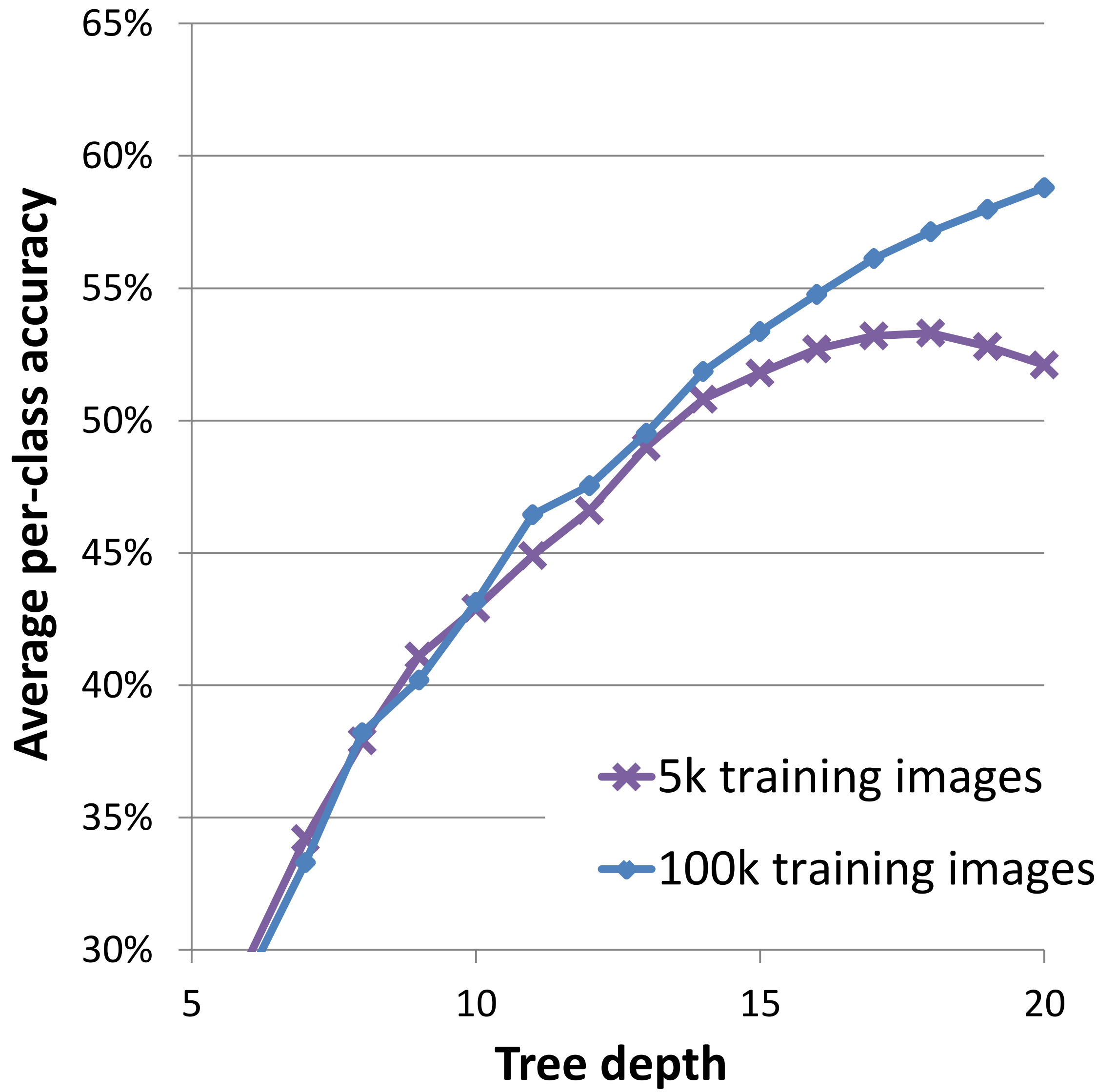
$$\Delta E = -\frac{|I_\mathrm{l}|}{|I_n|}E(I_\mathrm{l}) - \frac{|I_\mathrm{r}|}{|I_n|}E(I_\mathrm{r})$$

Goal: drive entropy at leaf nodes to zero

Depth of trees



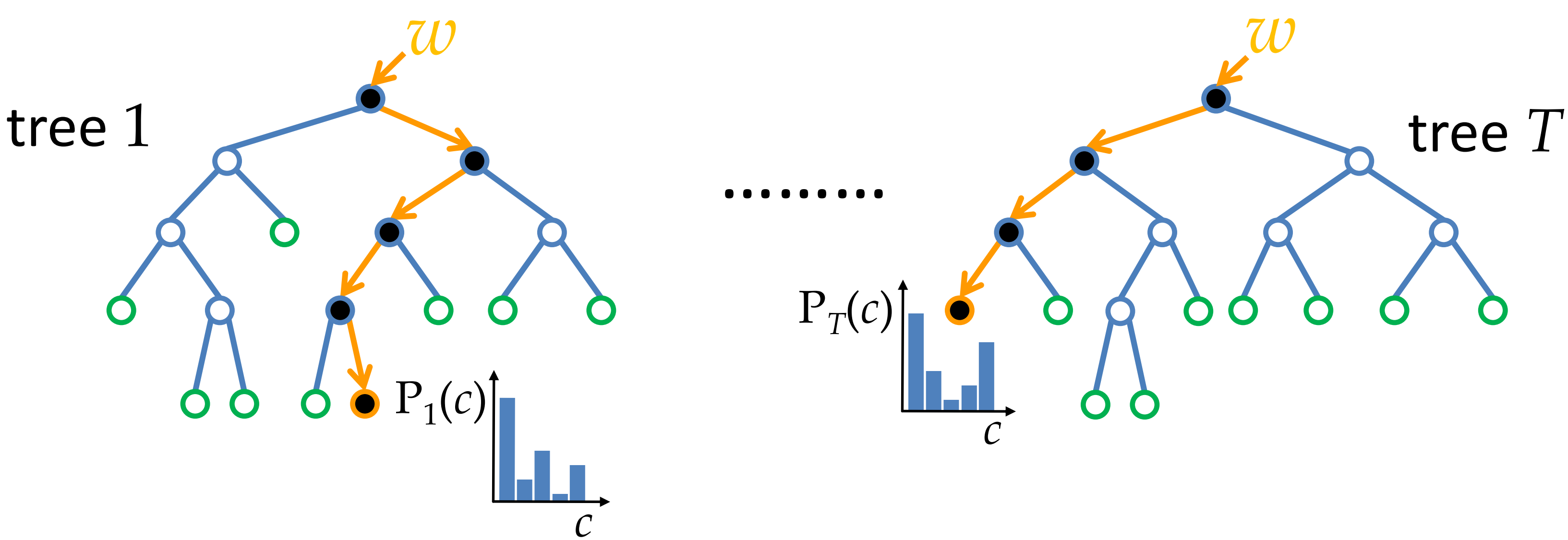
Depth of trees



Decision forests

[Amit & Geman 97]
[Breiman 01]
[Geurts *et al.* 06]

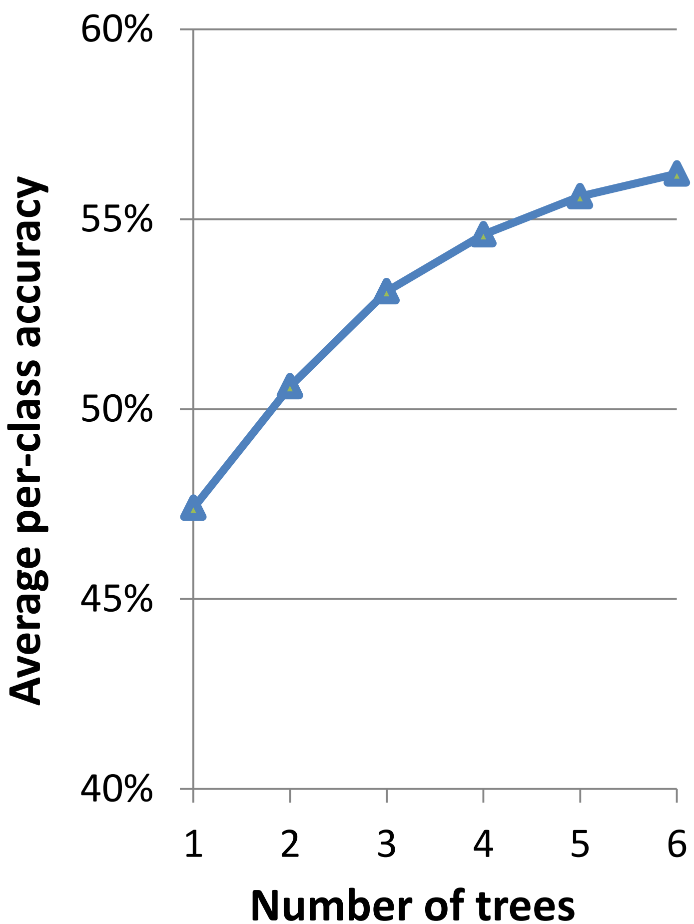
- Single trees tend to over-fit
- Train forest – ensemble of trees:



  - different random subset of images
  - average tree posteriors

$$P(c|w) = \sum_{t=1}^{T} P_t(c|w)$$

Number of trees



input ground truth

inferred body parts (most likely)

1 tree 2 trees 3 trees

Body parts to joint hypotheses

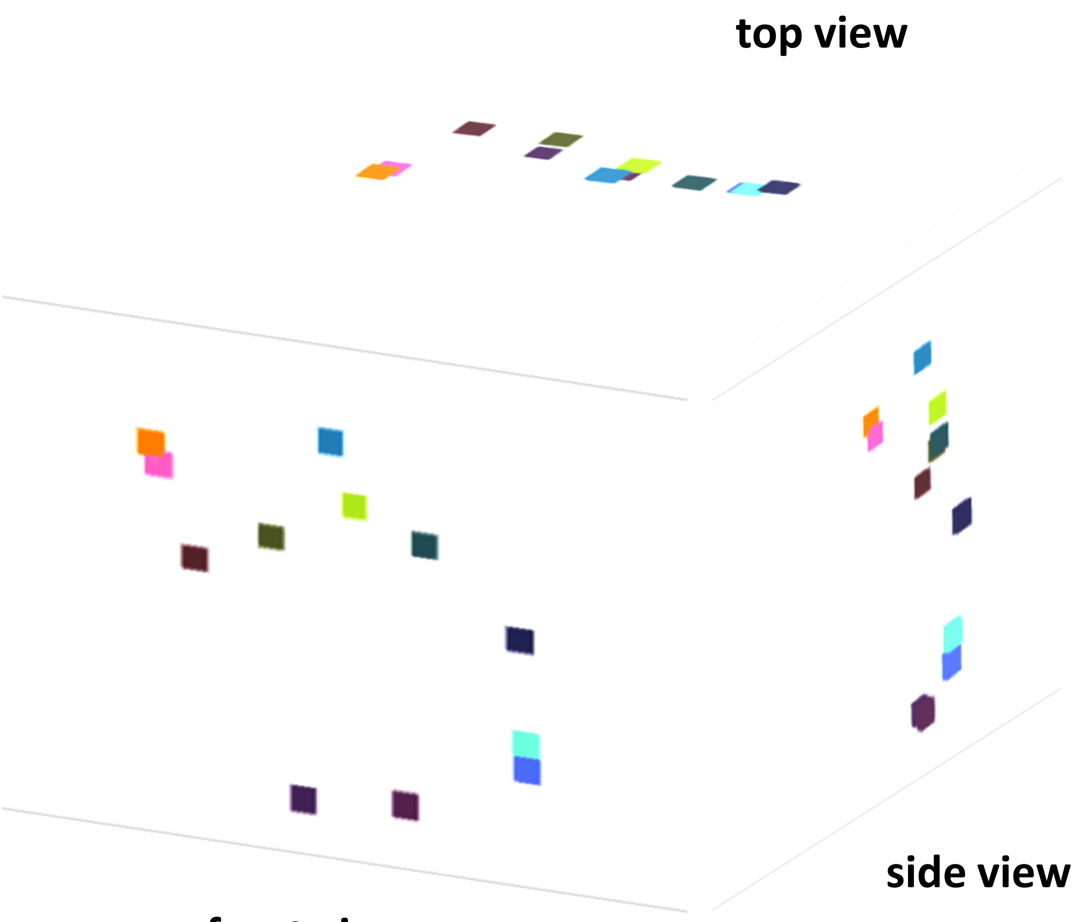
- Depth image & probability mass
- Localize body parts in 3D
  - global centroid of prob. mass
  - local modes of density (mean shift)
- Map body parts to skeletal joints
  - many parts map directly to joints

3D joint hypotheses

NB No tracking or smoothing!

input depth image

inferred body parts & overlaid joint hypotheses

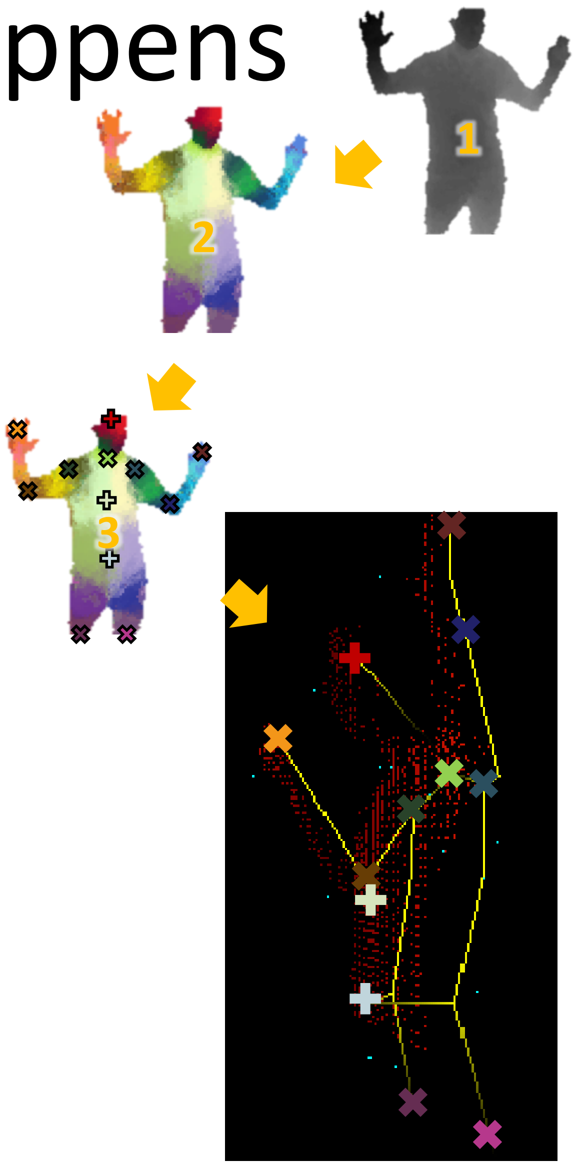


3D joint hypotheses

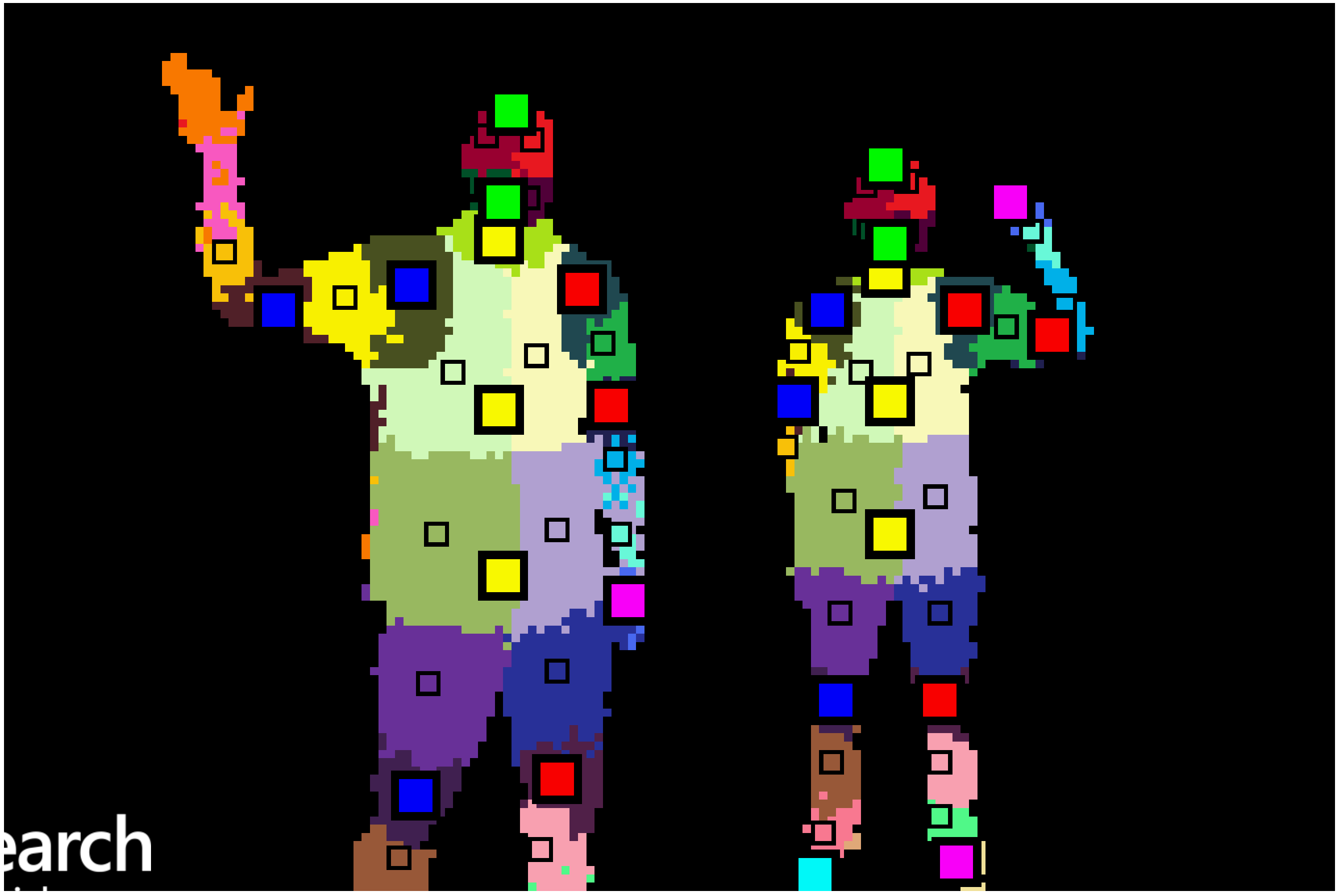
… and then magic happens

- Exploit
  - 3D joint hypotheses
  - kinematic constraints
  - temporal coherence

- Predict
  - full skeleton
  - invisible joints
  - multi-player



4. track skeleton
(3D side view)

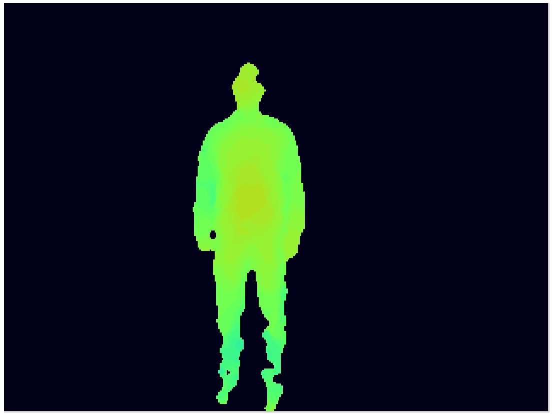
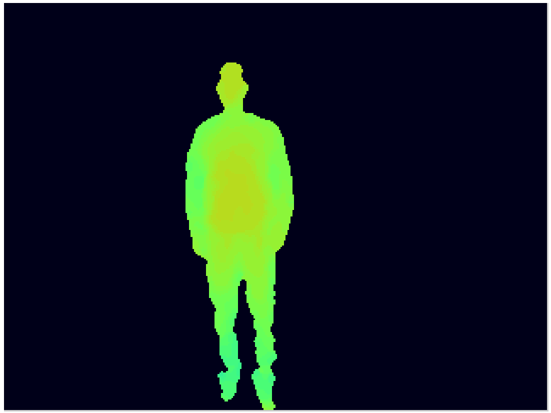
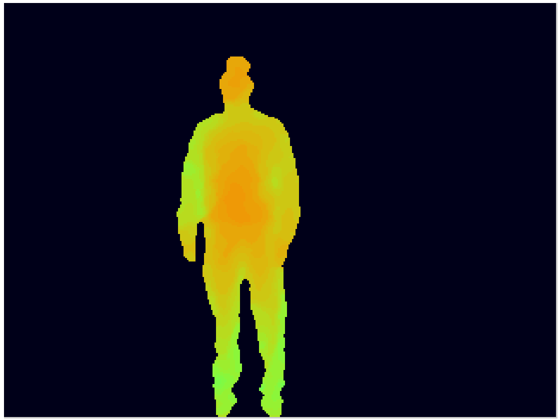


Wanqing Li, Zhengyou Zhang, Zicheng Liu

HUMAN ACTION RECOGNITION

The Problem

- Recognize actions from sequences of depth maps

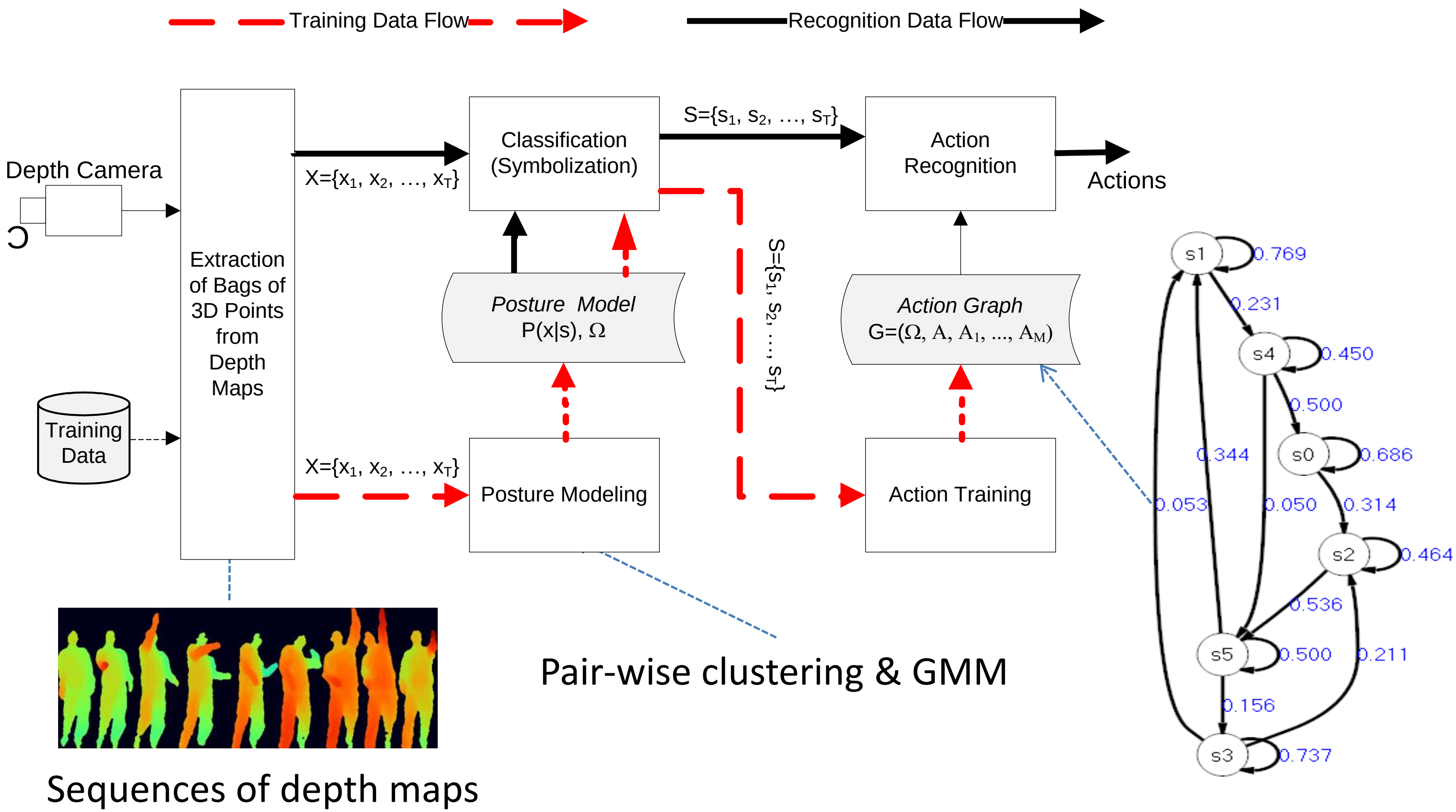




- Issues to address
  - large amount of data
  - Coarse and noisy depth measurement

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

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

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

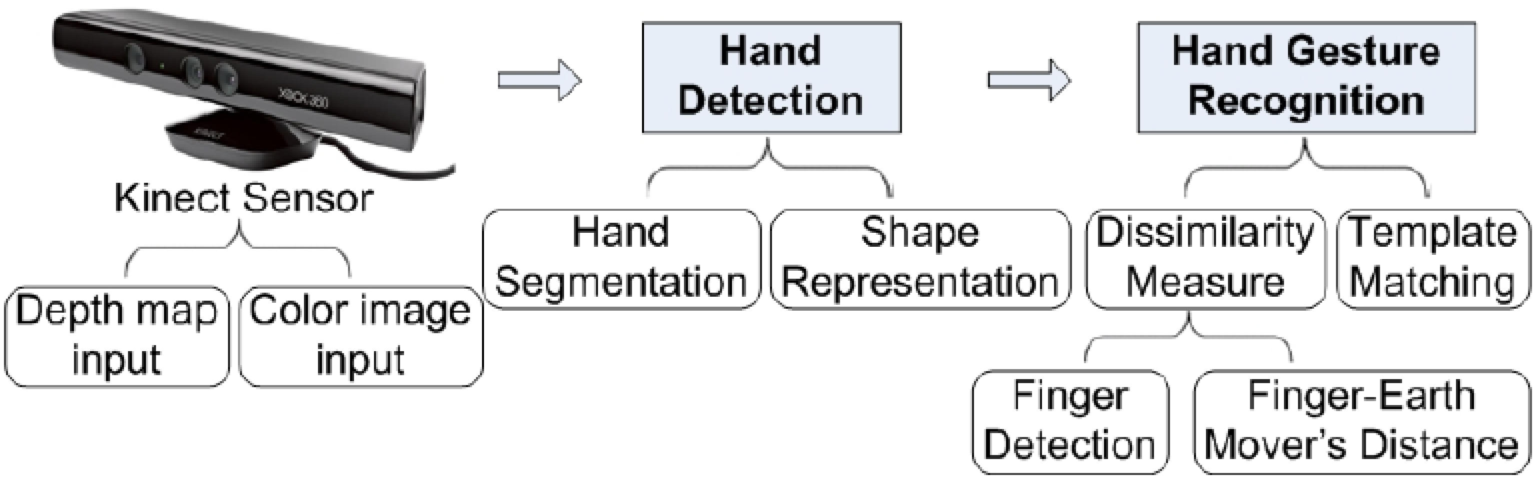


Figure 2: The framework of our real-life hand gesture recognition system.

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

Hand Segmentation & Representation

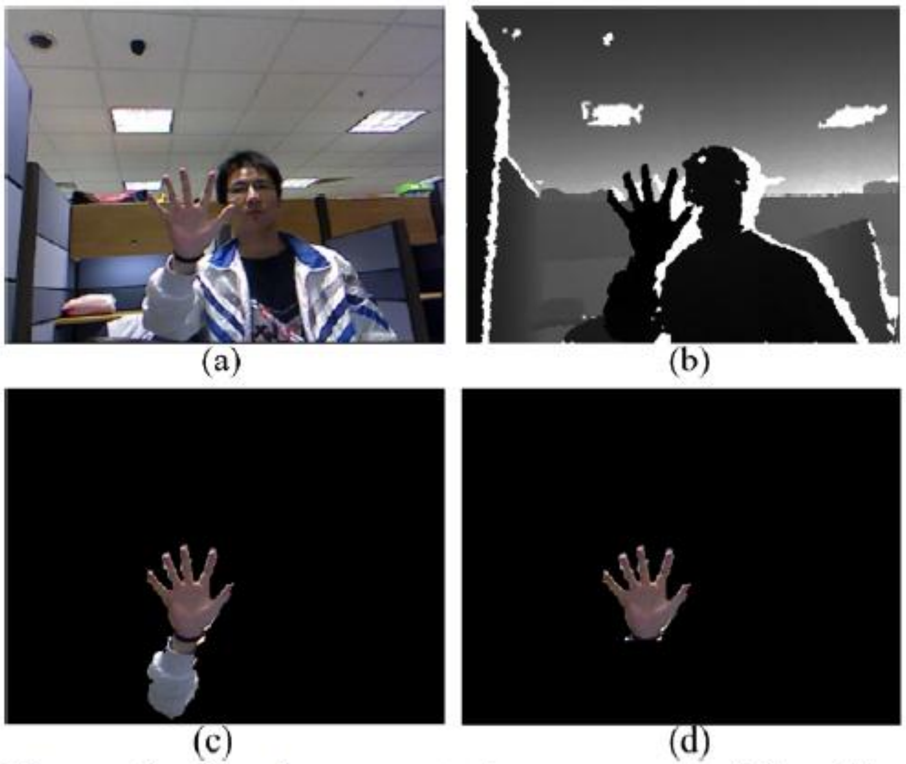


Figure 3: Hand segmentation process. (a). The RGB color image captured by Kinect Sensor; (b). The depth map captured by Kinect Sensor; (c). The area segmented using depth information; (d). The hand shape segmented using RGB information.

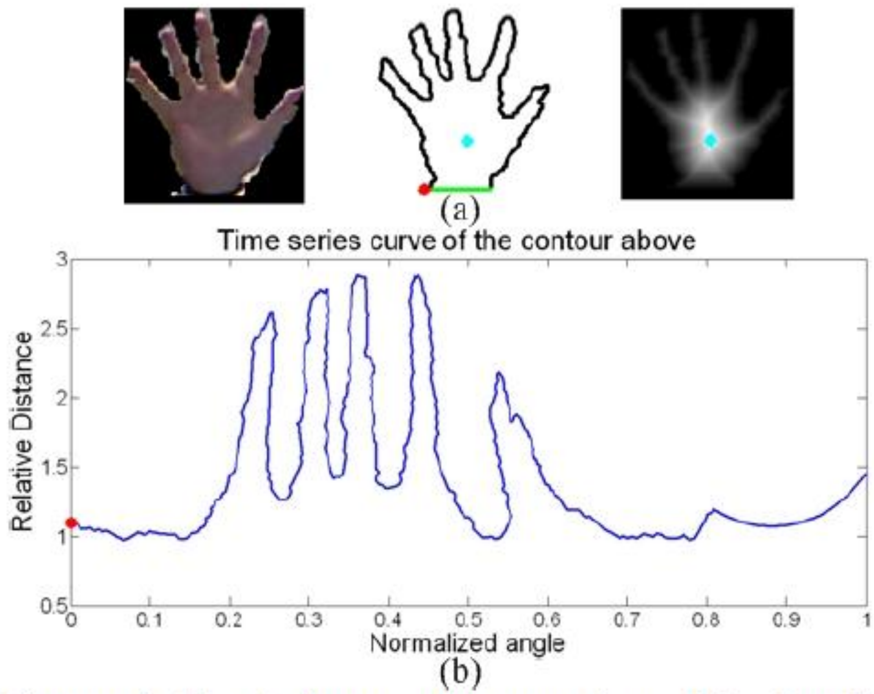


Figure 4: Hand shape representation. (a). On the contour of the segmented hand, the green line is the detection of the black belt; the red point is the initial point; the cyan point is the center point detected by Distance Transform; (b). The time-series curve of the shape above.

Finger Detection via shape decomposition

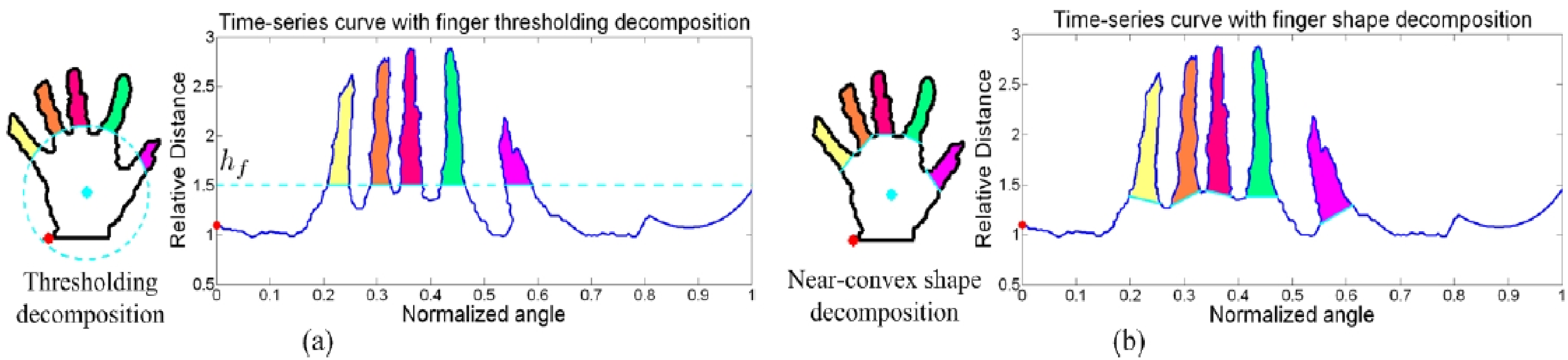


Figure 6: Illustration of the two proposed finger detection methods: (a). Thresholding decomposition uses a height threshold h_f in the time-series curve to detect fingers, which means to decompose the shape with a circle, thus information is inevitably lost; (b). Near-convex decomposition decomposes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

$$\min \quad \alpha \parallel \mathbf{x} \parallel_0 + (1-\alpha)\mathbf{w}^\top \mathbf{x},$$

$$s.t. \quad \mathbf{A}\mathbf{x} \geq 1, \quad \mathbf{x}^\top \mathbf{B}\mathbf{x} = 0, \quad \mathbf{x} \in \{0,1\}^{\overline{n}}$$

Distance Metric: Finger-Earth Mover's Distance

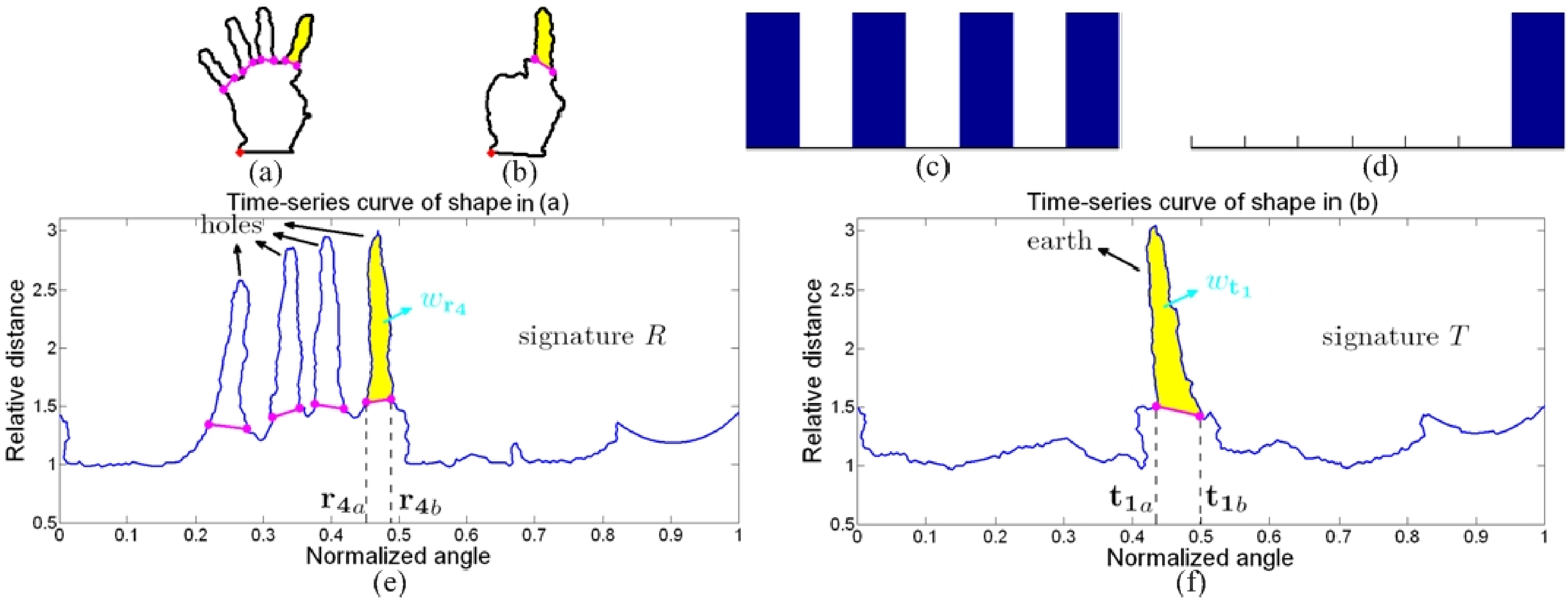


Figure 5: The motivation of using Finger-Earth Mover's Distance. (a) and (b) are two different hand shapes, whose time-series curves are shown in (e) and (f), respectively. Their major difference is the fingers. (c) and (d) are two signatures that partially match, their EMD cost is 0, however they are 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:

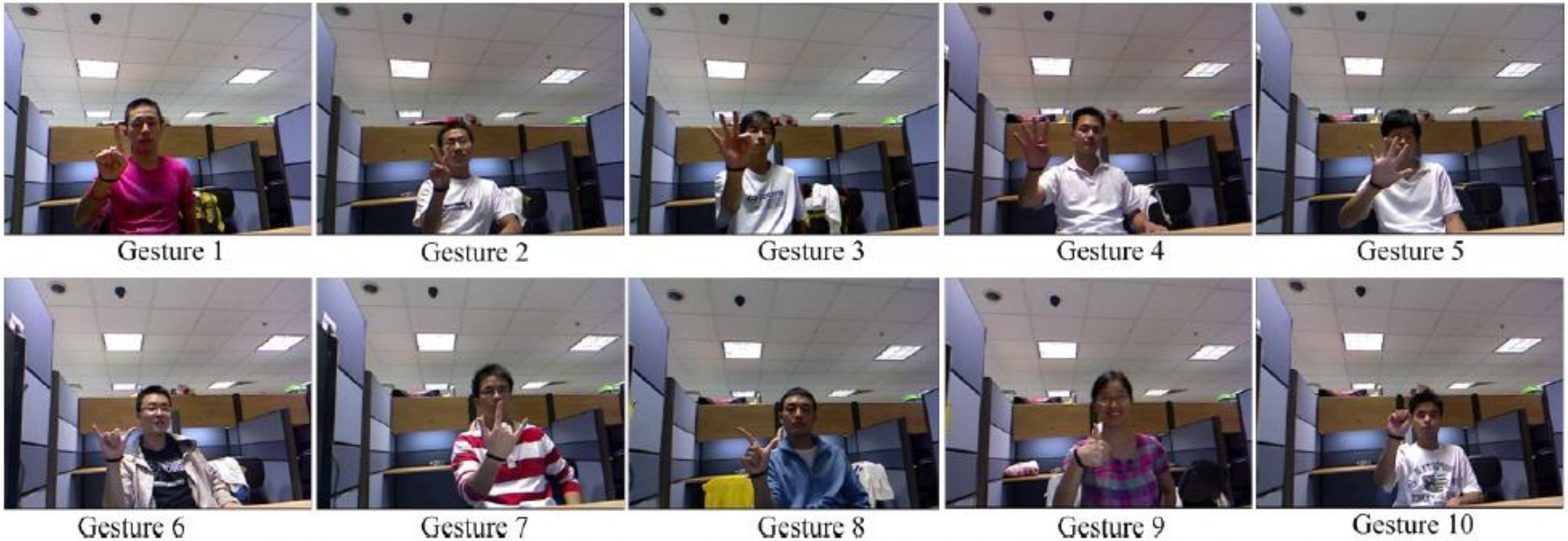


Figure 8: The color image examples for the 10 gestures in our dataset.

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.

	1	2	3	4	5	6	7	8	9	10
1	0.96	0.02						0.01	0.01	
2	0.02	0.86	0.05	0.01			0.02	0.04		
3		0.03	0.94	0.02			0.01			
4			0.04	0.88	0.05		0.03			
5				0.06	0.92	0.02				
6	0.02	0.01				0.96			0.01	
7			0.02			0.01	0.96	0.01		
8	0.03	0.03				0.01		0.93		
9	0.01					0.01			0.98	
10										1

Figure 15: The confusion matrix of Experiment II.

NANYANG TECHNOLOGICAL UNIVERSITY

Microsoft Research

SCOTTSDALE ARIZONA ACM MULTIMEDIA 2011 USA

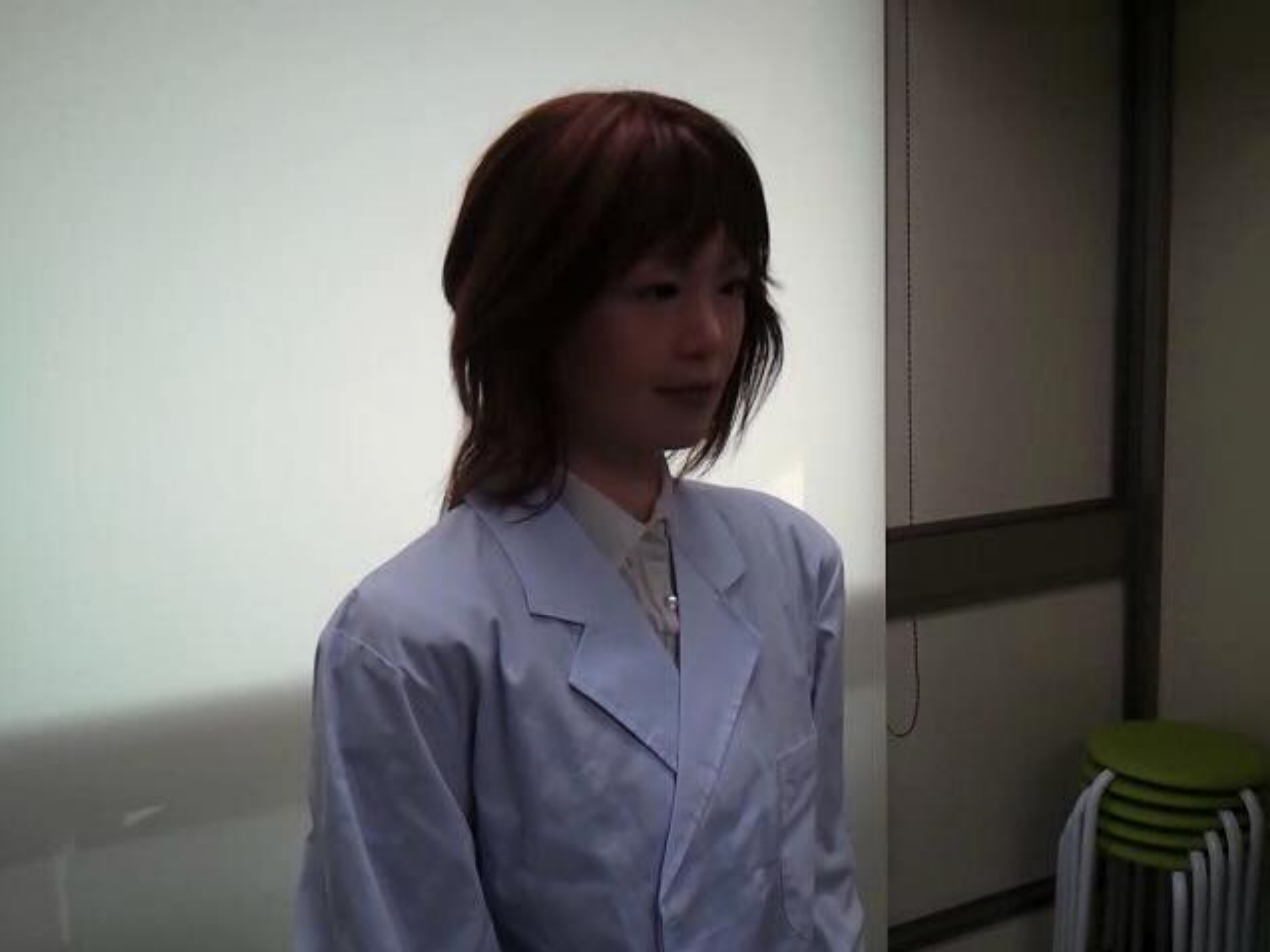
Robust Hand Gesture Recognition with Kinect Sensor

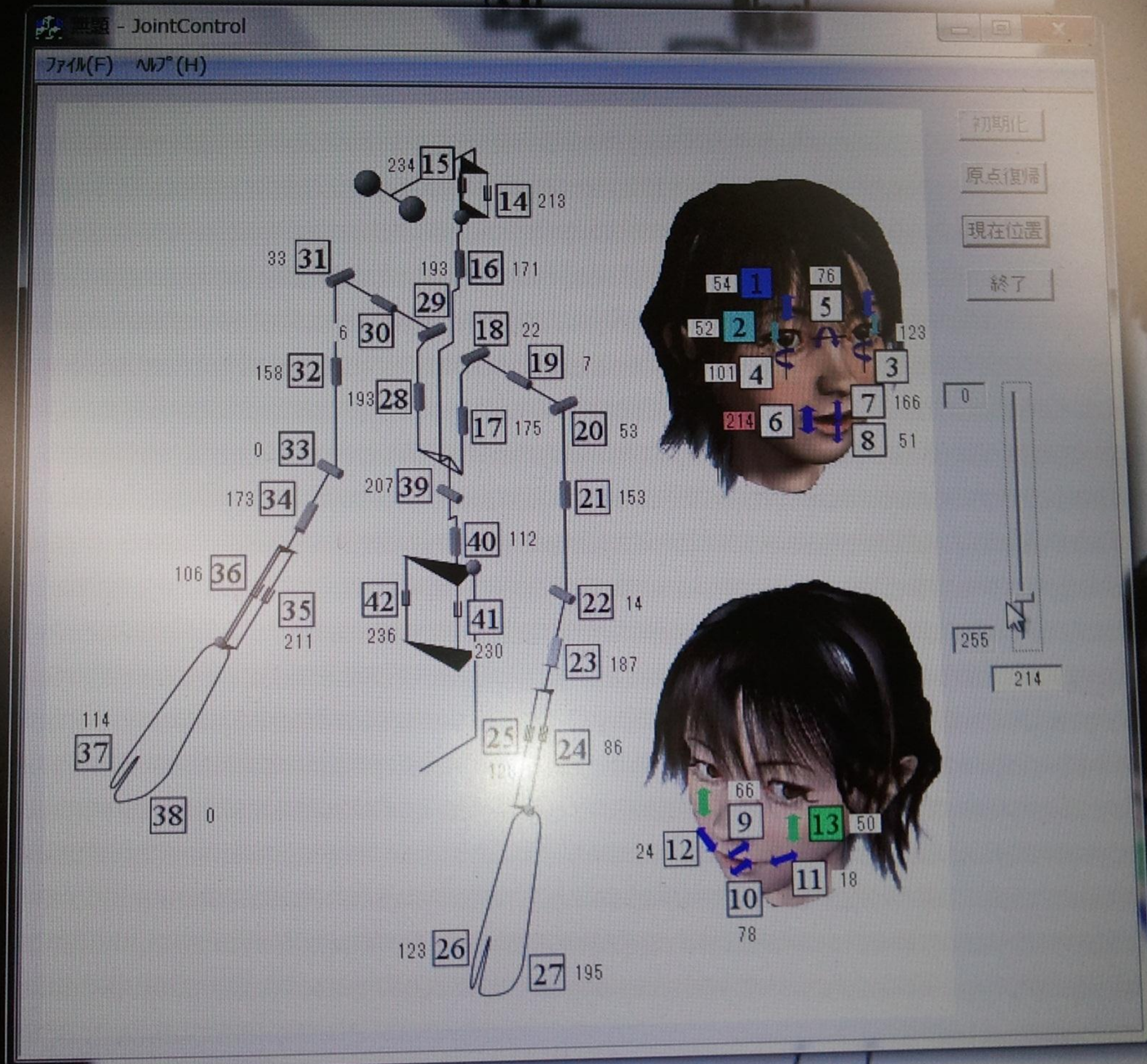
Zhou Ren, Jingjing Meng, Junsong Yuan, Zhengyou Zhang
School of EEE, NTU, Singapore & Microsoft Research, [illegible], USA

Proc. of ACM Intl. Conf. on Multimedia 2011

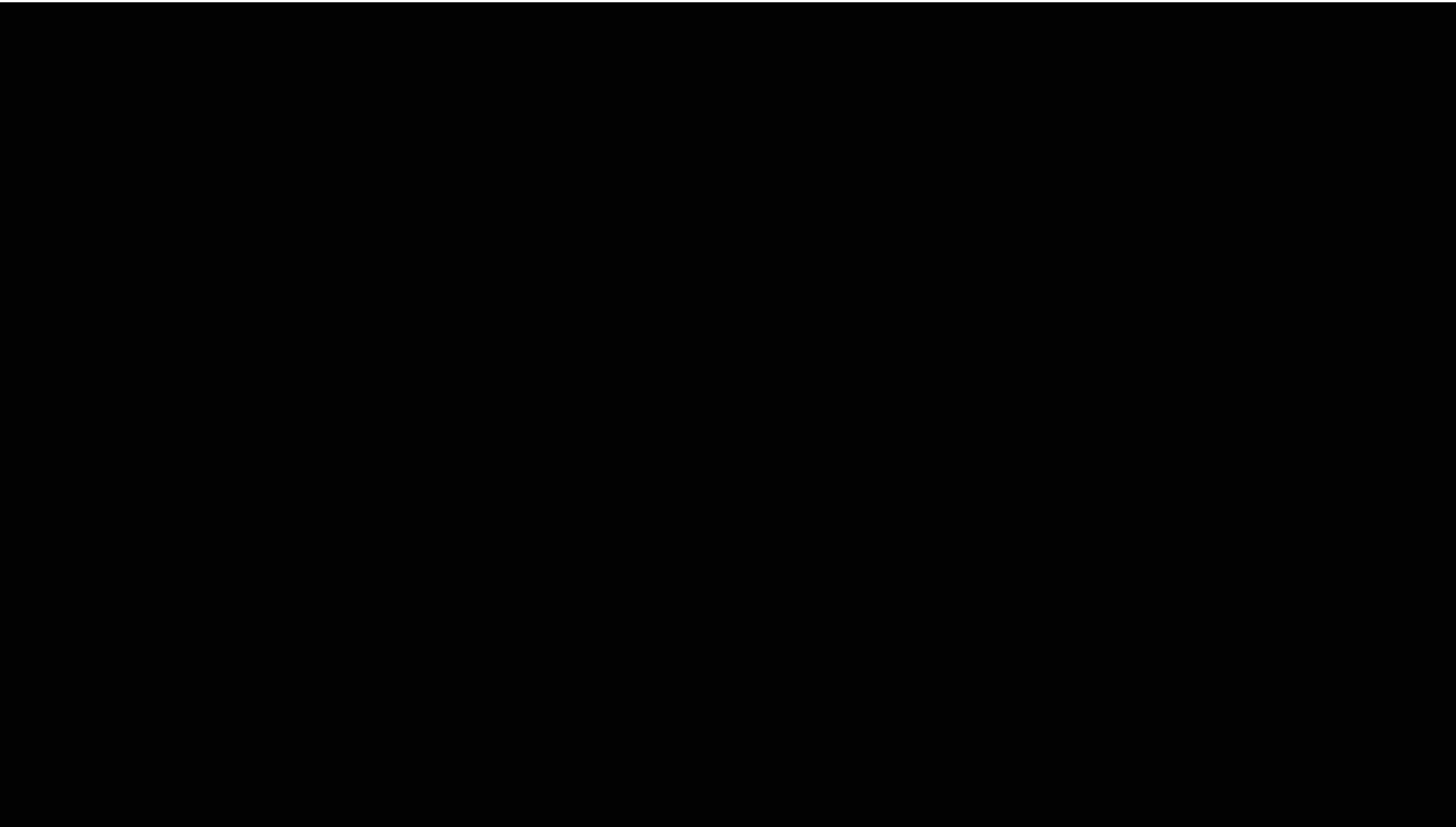
Qin Cai, Cha Zhang, Zhengyou Zhang

HEAD POSE & FACIAL EXPRESSION TRACKING




無題 - JointControl
ファイル(F)
ヘルプ(H)
初期化
原点復帰
現在位置
終了
0
255
214

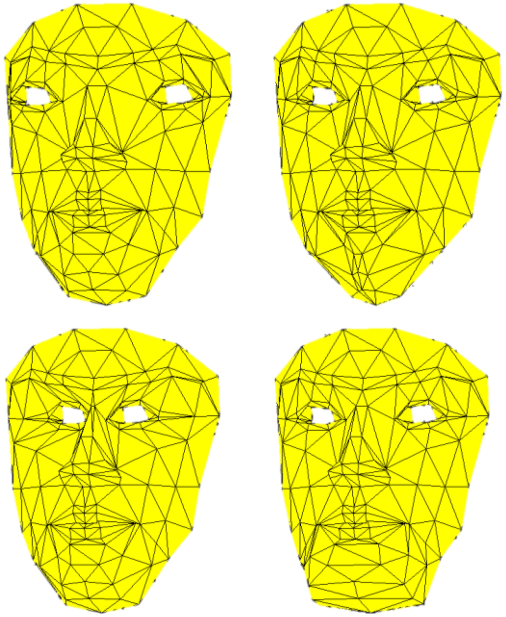
Geminoid Summit



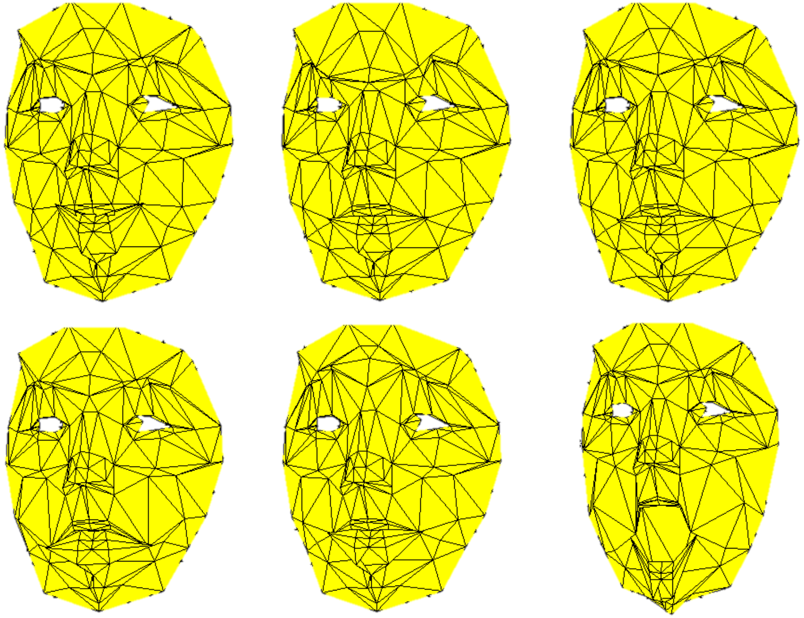
Deformable Face Tracking

- 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)

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

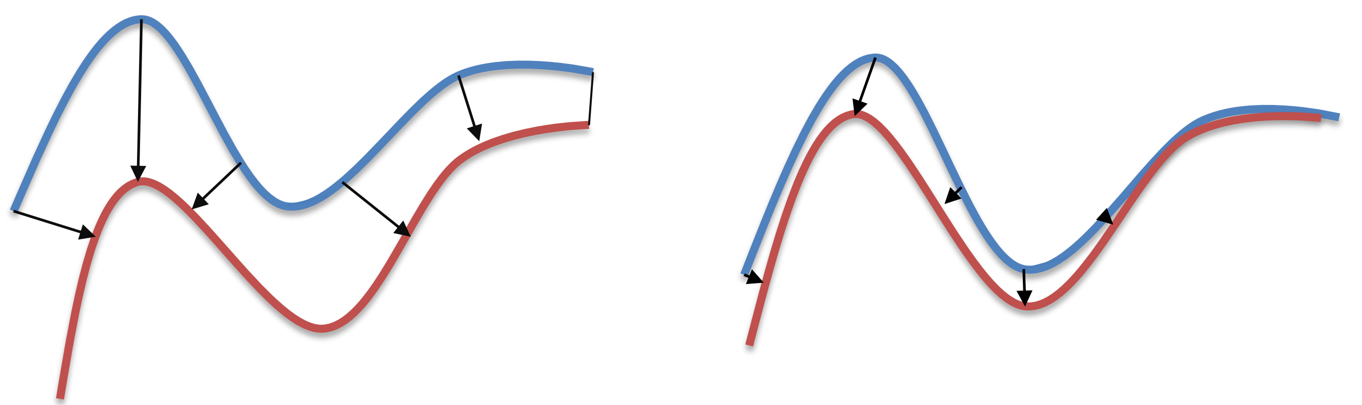
Maximum Likelihood DMF

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

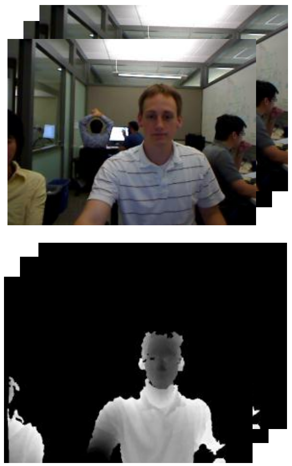
$$\boldsymbol{R}(\boldsymbol{p}_k + \boldsymbol{A}_k\boldsymbol{r} + \boldsymbol{B}_k\boldsymbol{s}) + \boldsymbol{t} = \boldsymbol{g}_k + \boldsymbol{x}_k$$

$$\boldsymbol{x}_k \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{x}_k})$$

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

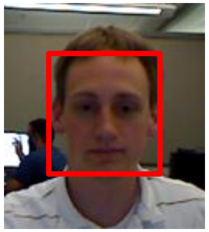


Model Initialization

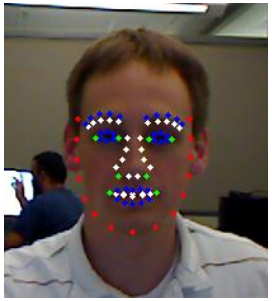


Input

Face Detection



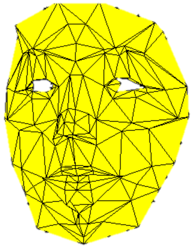
Face Alignment



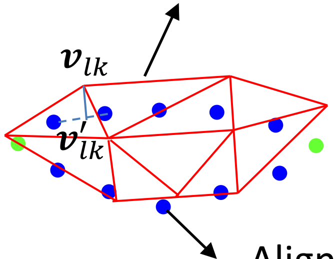
Green dots: point-to-point distance
Blue dots: point-to-plane 3D distance
Red dots: point-to-plane 2D distance
White dots: unused

Model Initialization

Output



Deformable model projected onto the texture image

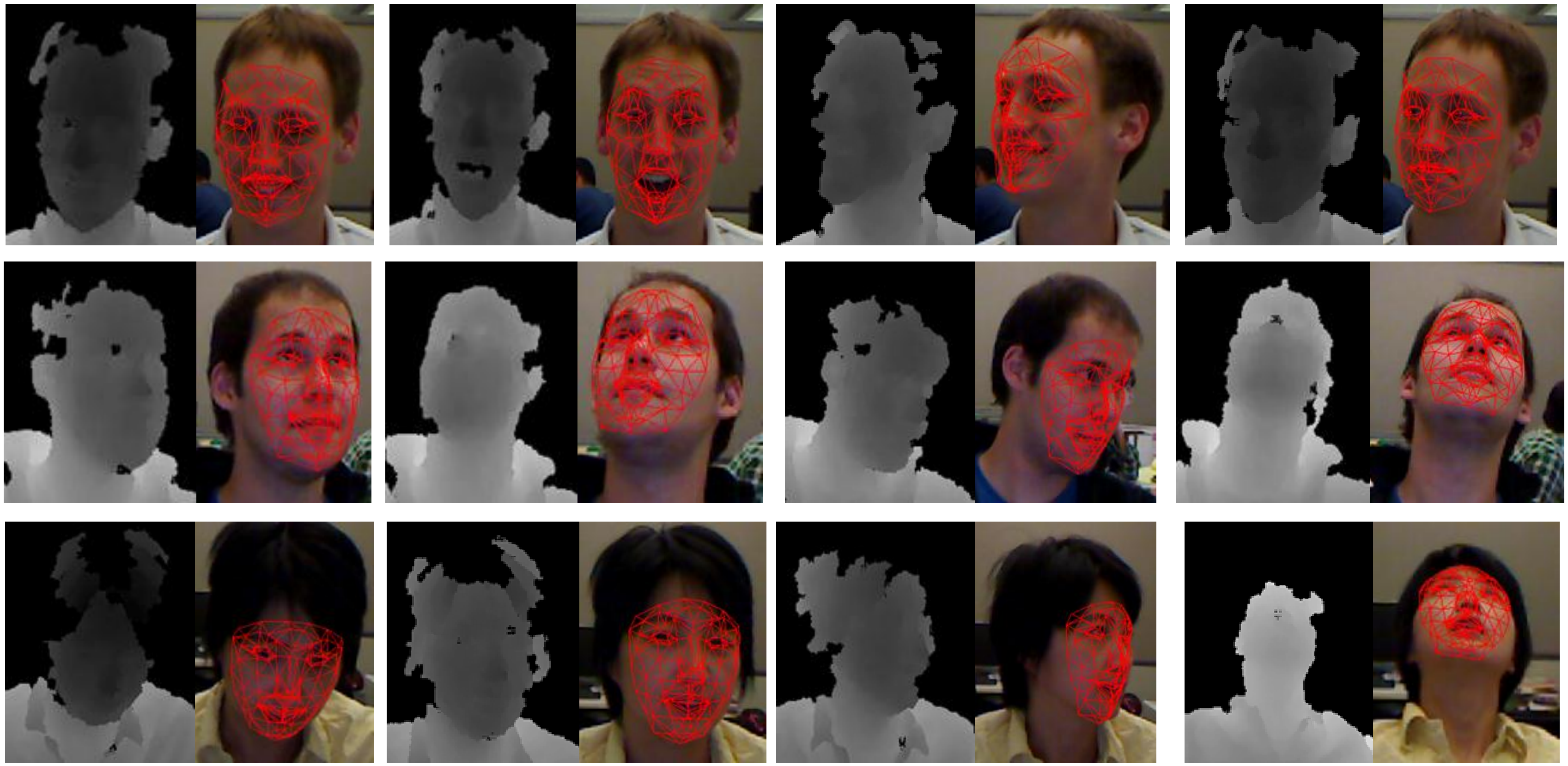


Alignment points

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

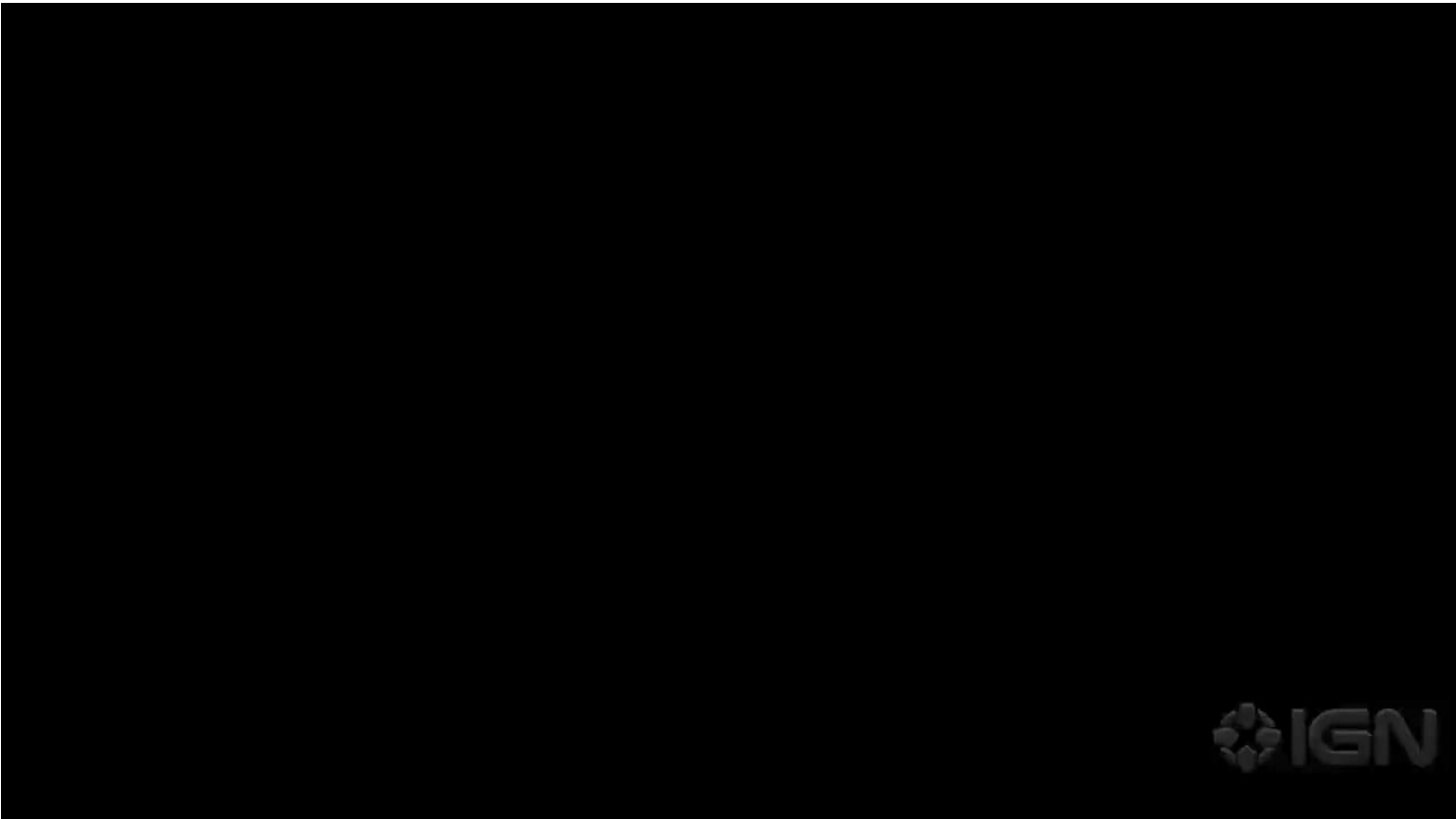
Median tracking error in pixels

	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	3.98**L**	3.91	3.91	3.92**L**	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

Avatar Kinect

Avatar Kinect



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

References

- J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, “Real-Time Human Pose Recognition in Parts from a Single Depth Image”, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (CVPR*)*, pages 1297-1304, 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.
- Z. Zhang, ``Microsoft Kinect Sensor and Its Effect’’, *IEEE MultiMedia*, Vol.19, No.2, pages 4-10, 2012.

Acknowledgment

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