The Future of Natural User Interaction

NUI: a great new discipline in the making

Baining Guo
Microsoft Research Asia

Microsoft Research

Microsoft

Kinect from Insiders' Perspective

- An interview with Kinect contributors

Kinect from Insiders' Perspective

- Kinect: A big collaboration project
  - Involving Xbox division, MSR Redmond, MSR Cambridge, MSR Silicon Valley, and MSR Asia

- Kinect :1st **mass market** product, **20 millions** customers → NUI is not science fiction
- NUI is a new engineering discipline in the making
  - NUI is far from where it should be; needs bigger **"foundation & pillars"**
  - "Foundation and pillars"are yet to be **invented**

Core Technologies & Problems

- Kinect is all about connecting you and your avatar – **without a controller**

Kinect, You and Your Avatar

You

Your Avatar

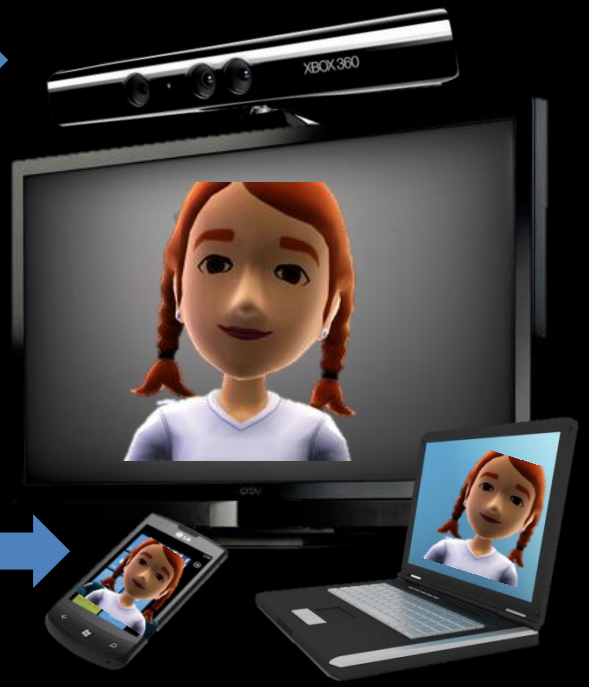
Kinect, You and Your Avatar

Tracking you
- *Your identity*
- *Your facial expression*

Kinect, You and Your Avatar

Tracking you
- *Your identity*
- *Your facial expression*

Controlling your avatar
- *Pose correction & tagging*
- *Face animation*



Core Technologies

- Tracking
  - identity tracking,
  - facial features tracking,
  - head pose tracking
  - …
- Gesture control (“gesture building”)
- Digitization

Tracking

Track you and your movement

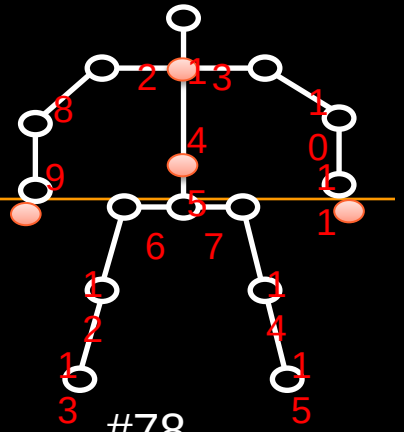
Kinect Identity (identity tracking)

- A core technology of Kinect for **robustly** binding each player with his avatar
- An **essential** part of Kinect & used by all Kinect games as it is a Kinect game certification requirement
- See more info in the article
  - Tommer Leyvand, Casey Meekhof, Yi-Chen Wei, Jian Sun, and Baining Guo, “Kinect Identity: Technology and Experience”, IEEE Computer, April 2011
  - This article made it to the list of the most read articles of IEEE Computing Now

Kinect Identity

- In Kinect, each skeleton ←→ a game character / a player profile
- When skeleton tracking fails & resumes or player leaves & comes back
  - Which skeleton to use? A new player or an existing payer?

Solution?



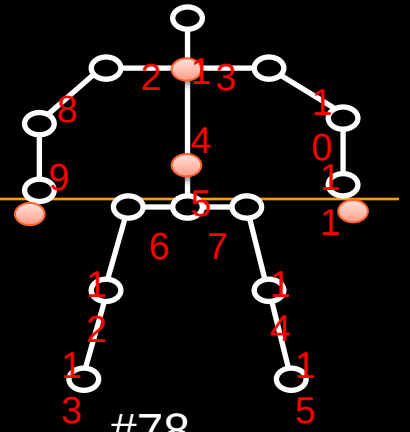
#24 #25 #RGB frame #77 #78





Body dimension

- Body tracking is unstable for recognition



#24 #25 #RGB frame #77 #78

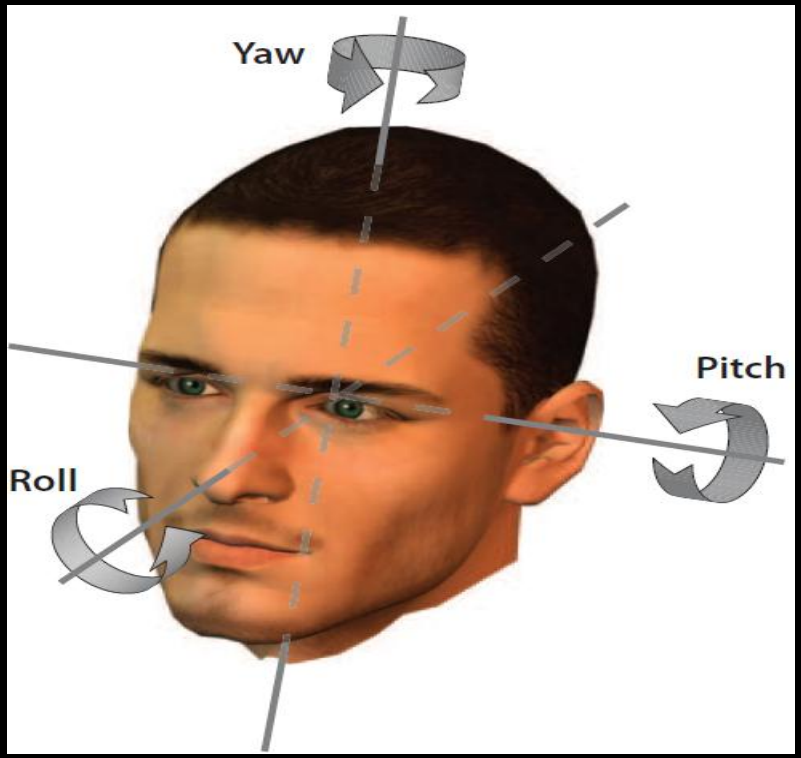
initialized skeletons from individual frames

Challenges and our solution

- Challenges: robustness and performance
  - instable skeleton tracking, varying lighting …
  - 2ms/frame buffer for xbox game

- Our solution: fusion of multiple visual signatures
  - facial, clothing, body dimension
  - robust and efficient feature extraction

Tracking Head Pose (head orientation)



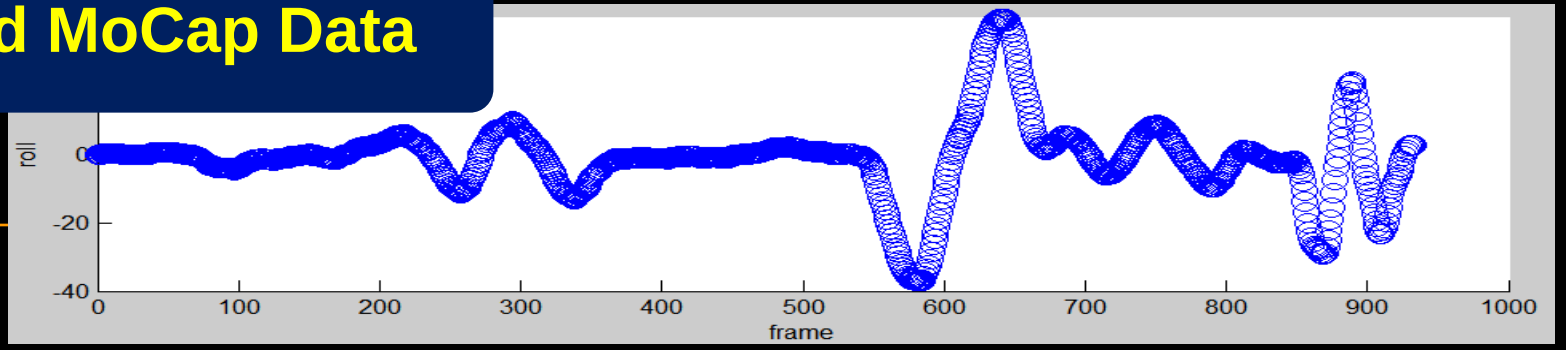
- Roll : (-45, 45)
- Yaw : (-60, 60)
- Pitch : (-60, 60)

Training Kinect to track head pose

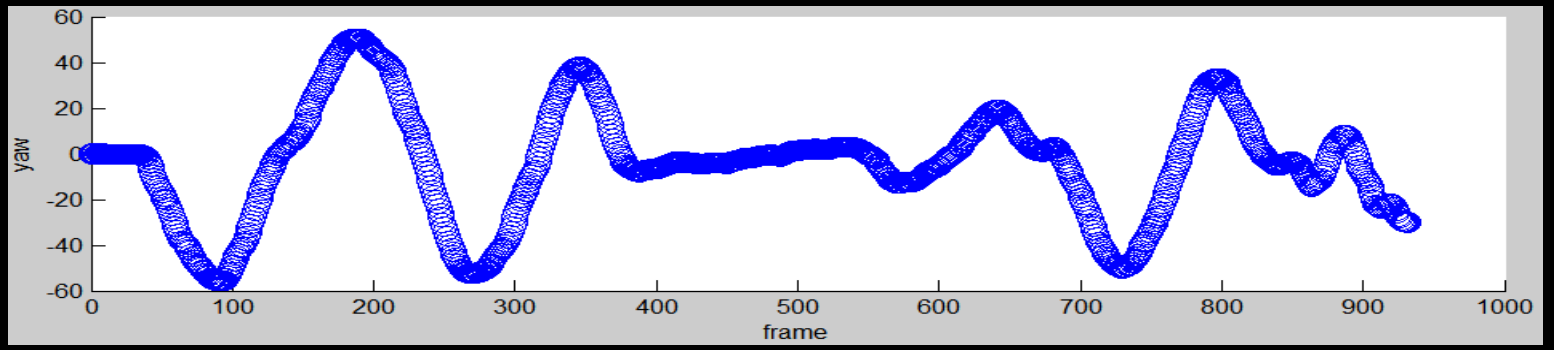
Training data: video and head MoCap data

Head MoCap Data

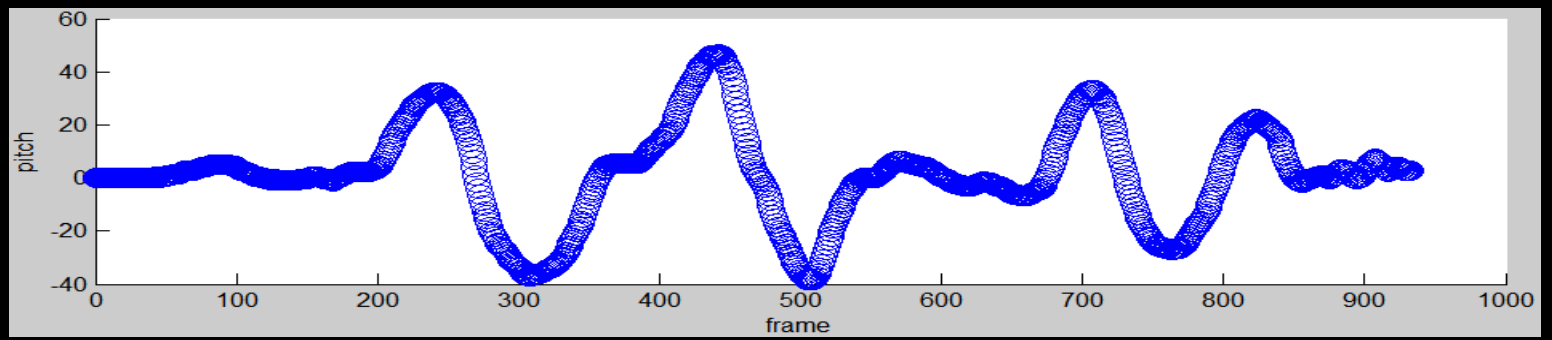
roll

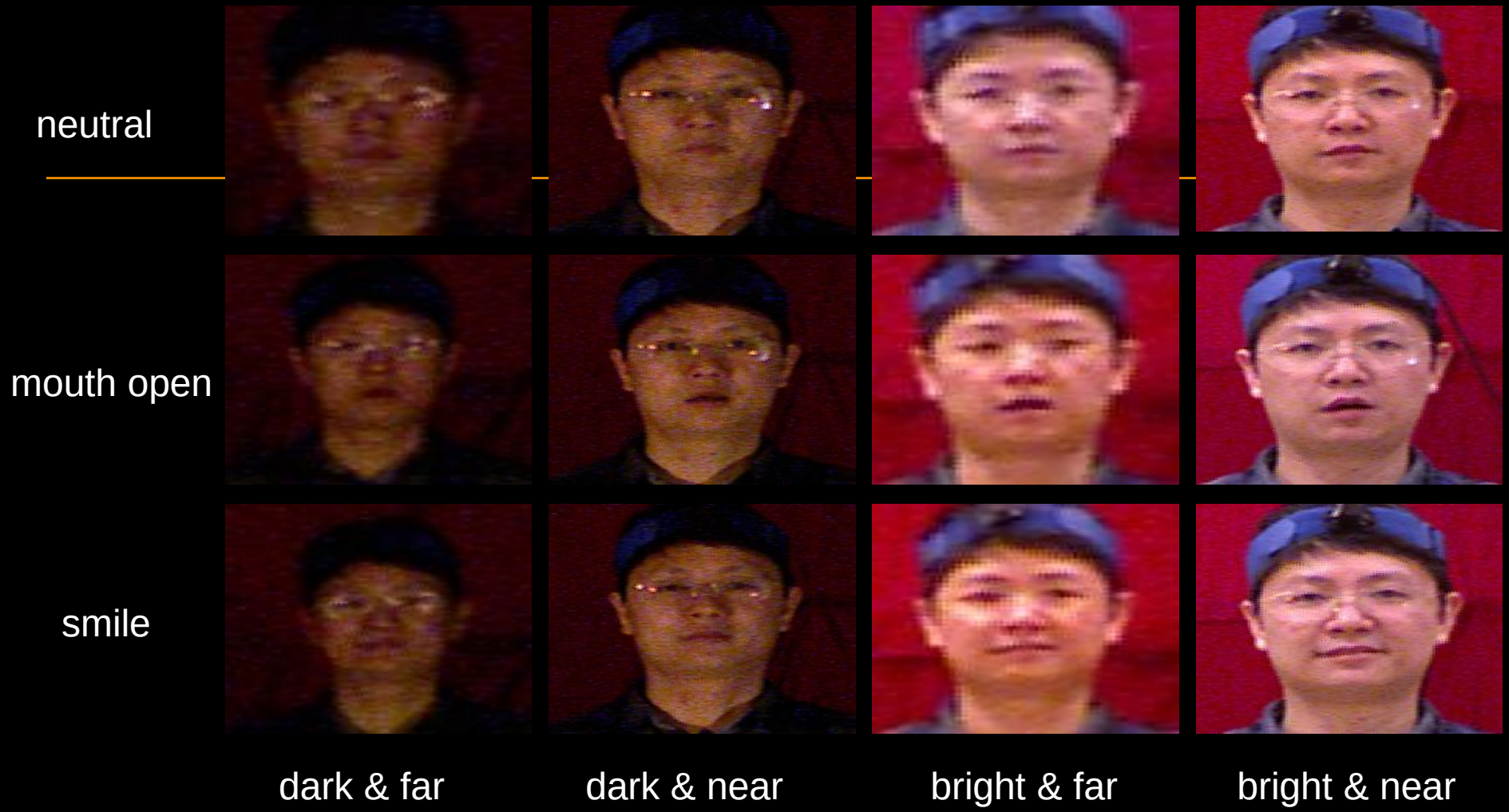


yaw



pitch




neutral
mouth open
smile
dark & far
dark & near
bright & far
bright & near


adils.bmp
bemyers.bmp
bkeogh.bmp
brandf.bmp
cmeekhof.bmp
dankr.bmp
emily.bmp
eyalk.bmp
fangwen.bmp
gdavies.bmp
jasonro.bmp
jenng.bmp
jgup.bmp
jiansun.bmp
jmarien.bmp
johenry-aurora.bmp
johenry-fred.bmp
johenry-lilith.bmp
kareemc.bmp
landond.bmp
lonnym.bmp
martinlo.bmp
matlop.bmp
mattfl.bmp
monicac.bmp
mosesm.bmp
mplagge.bmp
nazeehe.bmp
parhamm.bmp
quais.bmp
robheit.bmp
rredman.bmp
segnan.bmp
stana.bmp
stspeich.bmp
timg.bmp
timkeosa-lisa.bmp
toanh.bmp
v-dejone.bmp
vhouse.bmp

Challenge: fast head pose tracking

- Hard constraint: computing time/frame < **5 ms**
  - **Has to be accurate too**

- A classical regression problem w/ a **big input feature space** (dim = 7100)
- PCA helps but not sufficient (dim = 2500)
  - Accuracy cannot be compromised

From image to pose – and only pose!

- Identity
- Pose
- Lighting
- Expression

Manifold embedding by Multi-class LDA

- Linear Discriminant Analysis (LDA)
  - **Quantize** pose space into discrete pose classes
  - Find optimal subspace projection that maximizes between-class variation and minimizes within-class variation (**track only pose & nothing else**, dim = dozens)

2-class LDA: find subspace projection that maximizes $\frac{|A1-A2|^2}{S1^2+S2^2}$

Manifold embedding by Multi-class LDA

- **Linear Discriminant Analysis (LDA)**
  - **find optimal subspace projection that maximizes between-class variation and minimizes within-class variation**

$$cov_{between} = \frac{1}{C}\sum_{i=1}^{C}(u_i - u)(u_i - u)^T \quad cov_{within} = \frac{1}{C}\sum_{i=1}^{C} cov_i$$

dim eigen-space of **pose space**

u_i: mean of samples in class i

u: mean of all samples

- **compact but keeps essential pose info**

$$S = \frac{w^T cov_{between} w}{w^T cov_{within} w}$$

Head Pose Tracking

- Shipped in Feb 2012 in Kinect SDK, NUI API
  - Available to all Kinect developers worldwide

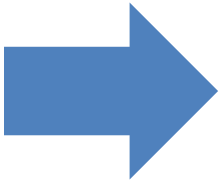
- Kinect as a “**publication venue**”

Facial Feature Tracking

- Avatar Kinect: a Kinect service for people to **chat & interact** via their Xbox Live avatars in virtual chat rooms
  - Chat & interact: A new **digital lifestyle**
- Shipped to all Kinect users in July 2011 with Kinect Fun Labs
- Collaboration w/ MSRA & MSR Redmond (Zhengyou Zhang)

Avatar Kinect (video)

Research Issues



Estimated shape

Research Issues

- Requirements: accurate, robust, efficient

Existing Best Approach – AAM (Active Appearance Model)

Reconstruction Error

Cost Function

Model generated Input Face

Greedy Optimization:

Drawbacks of AAM-based Approaches

- Cost Function
  - Bad generalization on unseen person
- High computation cost
  - Local minimum; sensitive to initialization
- Parametric model
  - Not adaptive in the iterative optimization

The standard theoretical framework for the past 20 year!

AAM vs Explicit Shape Regression

- Cost Function
  - Bad generalization on unseen person
- High computation cost
  - Local minimum; sensitive to initialization
- Parametric model
  - Not adaptive in the iterative optimization

AAM-based approaches

- Cost Function
  - A repressor learned from a very large training data
- Super fast (2-10ms)
  - Two level cascade and multiple initialization
- Non-parametric model
  - Adaptive coarse-to-fine shape constraint

Explicit shape regression (CVPR ’12 oral)

Gesture Control

Control how your avatar moves

Understanding the player's movement

From raw data to high-level knowledge

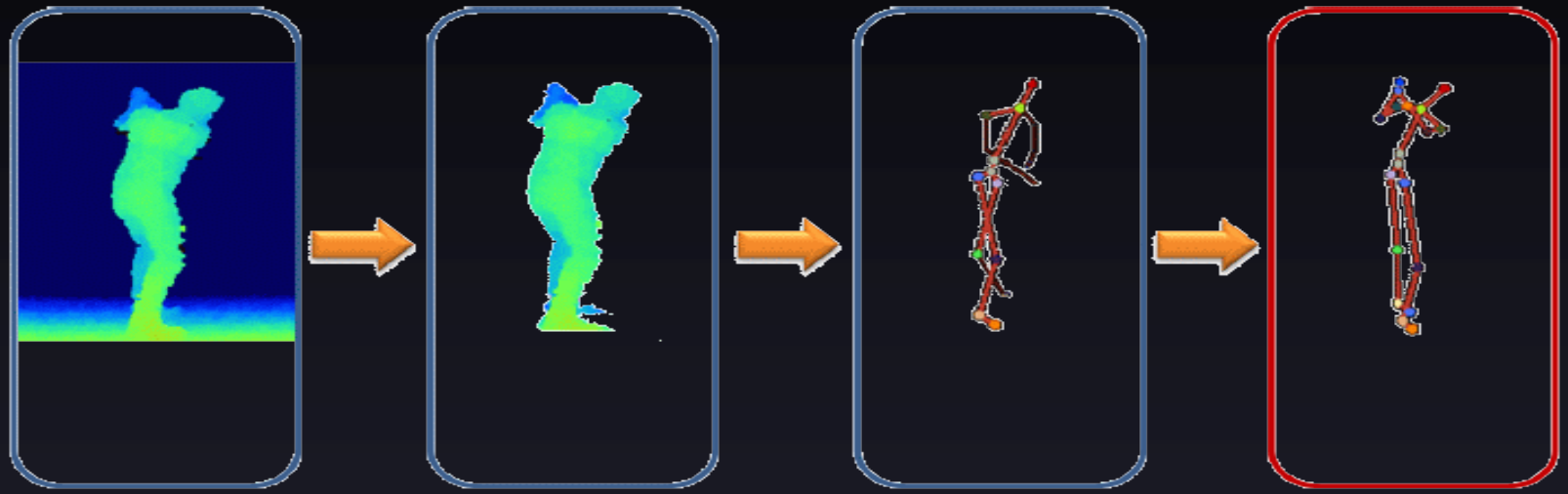
Depth map → skeleton → gesture

- Depth map: low-level raw data
- Skeleton: intermediate representation
- Gesture: high-level knowledge for controlling avatars

Can we define gesture using the skeleton?

- skeleton is not reliable at critical moments (arms crossing, legs crossing, body turning sideways)

Pose Correction & Tagging



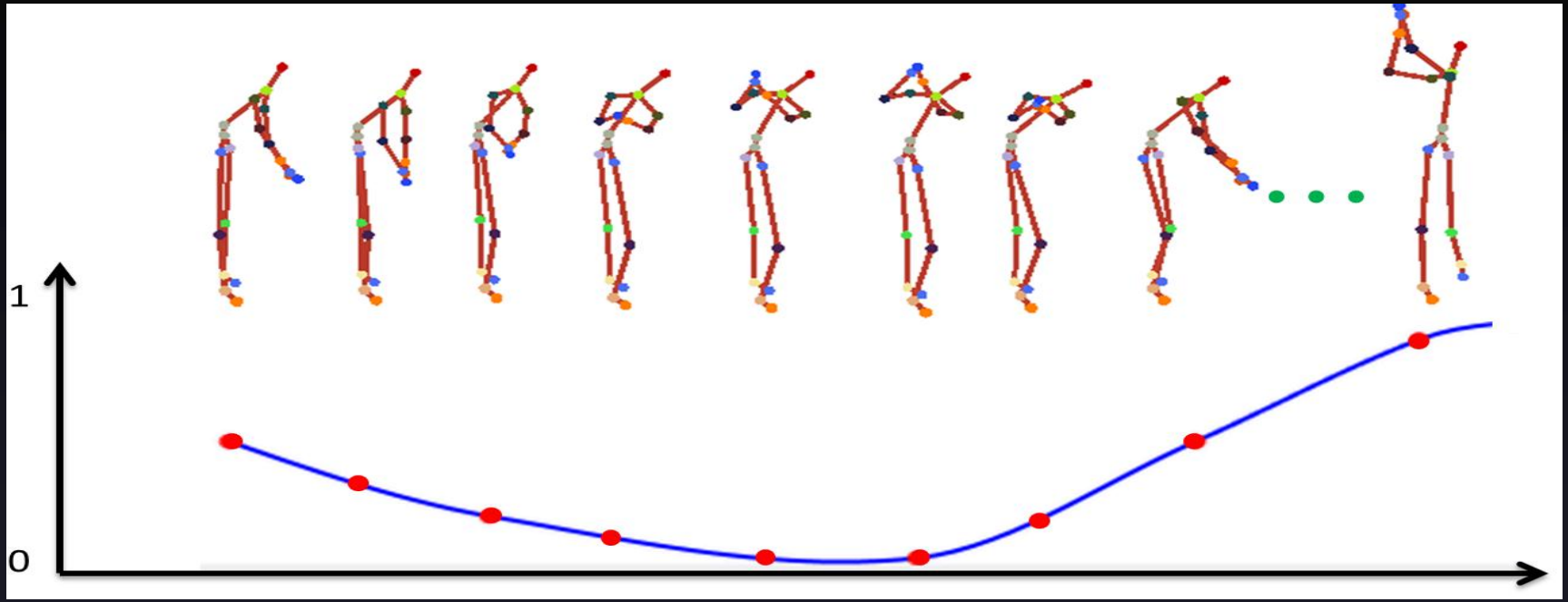
Depth Image

Background Removal

Skeleton Extraction

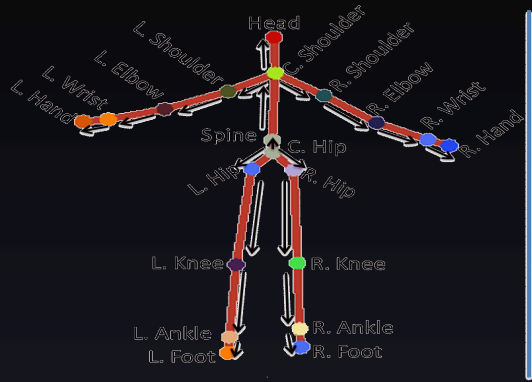
Skeleton Correction

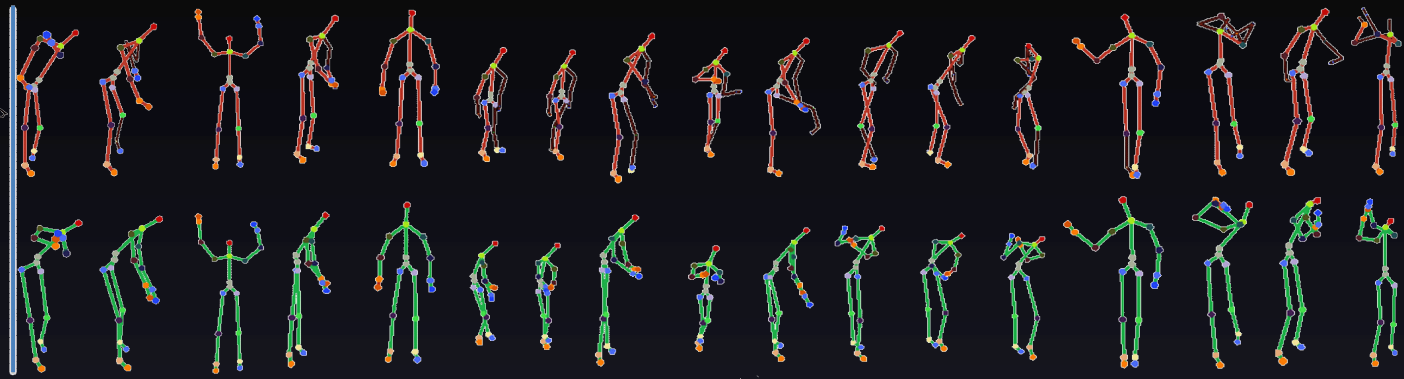
Skeleton/Pose Tagging



- The corrected skeletons are tagged w/ numerical values
- The numerical values are used to drive the avatar

Context-Based Pose Correction & Tagging





- Context == the activity that the user is doing
- For the given context, gather ground truth data
  - manually labeled skeleton with tags
- From the ground truth data, train a random forests regressor for automatic pose correction & tagging (in this context!)

The Gesture Component of Kinect SDK

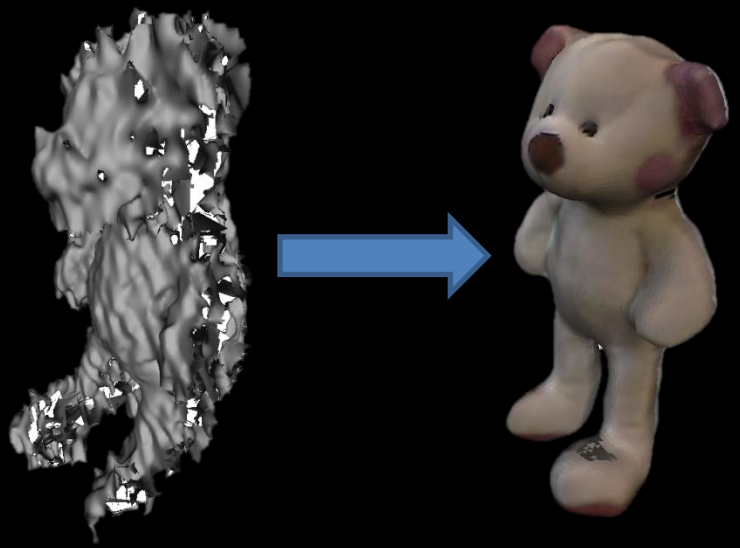
- Shipped in Feb 2012 Kinect SDK, NUI API
  - Official name: Kinect gesture builder
  - Available to all Kinect developers world wide
- More details in
  - "Exemplar-Based Human Action Pose Correction and Tagging", W. Shen, K. Deng, X. Bai, T. Leyvand, B. Guo & Z. Tu, IEEE Computer Vision and Pattern Recognition, 2012

Digitization

Bring physical objects into cyber space

Object Digitization

- Simple inputs
  - Front and back snapshots of objects
- Good 3D reconstruction results
  - From noisy input to smooth outputs
- Fast
  - Using both CPU and GPU



Object Digitization

- Shipped with the Kinect Fun Labs in July 2011
  - The object capture lab
  - Available to all Xbox Live members on Kinect

Major Research Issues

- Dealing with noisy Kinect data: **Poisson geometry processing**
  - Mesh editing with Poisson-based gradient field manipulation, Y Yu, K Zhou, D Xu, X Shi, H Bao, B Guo, HY Shum, ACM Siggraph 2004
  - Laplacian surface editing, O Sorkine, D Cohen-Or, Y Lipman, M Alexa, C Rössl, HP Seidel, Eurographics SGP, 2004
  - Poisson surface reconstruction, M Kazhdan, M Bolitho, H Hoppe, Eurographics SGP, 2006

"it preserves surface details and produces visually pleasing results by distributing errors globally through least-squares minimization",

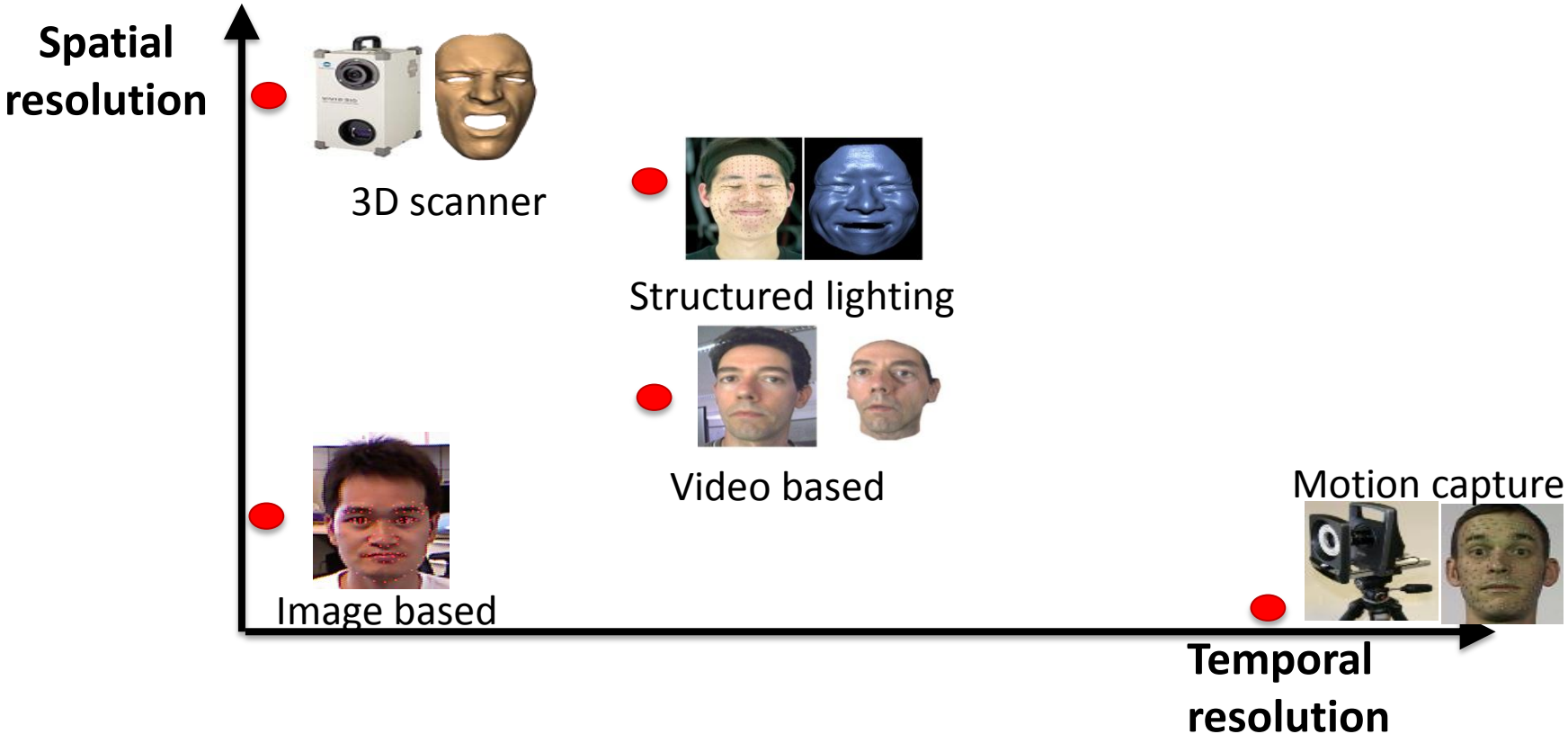
-- K Zhou, J Huang, J Snyder, X Liu, H Bao, B Guo & H Shum (2005)

(See "Large deformation using volumetric graph Laplacian", Siggraph 2005)

Major Research Issues

- Making it fast: **data-parallel octree**
  - **Can we build geometry octrees on the GPU?**
  - Data-parallel octrees for surface reconstruction, K Zhou, M Gong, X Huang, B Guo, IEEE TVCG 2011

Face Digitization & Animation ("creating my avatar that looks me & moves like me")



High-Fidelity Facial Animation

- Traditional motion capture
  - Realistic motion details, but lacks geometry details
- Our approach: FaceMocap+
  - Realistic motion details just like motion capture
  - Plus geometry details as in laser scans
  - Paper published in Siggraph'11

Results with Texture

Results without Texture

Results with Texture

Results without Texture

Face Digitization

- Generate personalized 3D avatar based on Kinect input
  - on-going research w/ MSRA + MSR Redmond (Zicheng Liu)

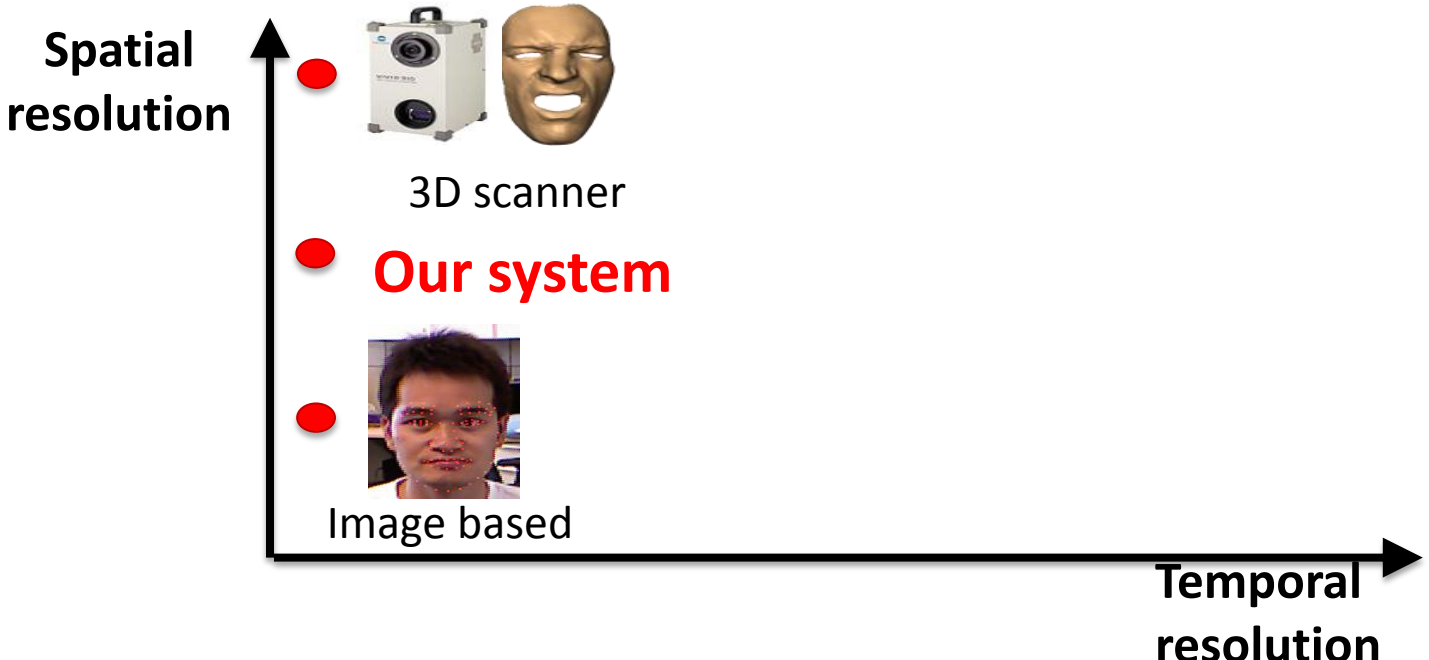


Image-based Approach

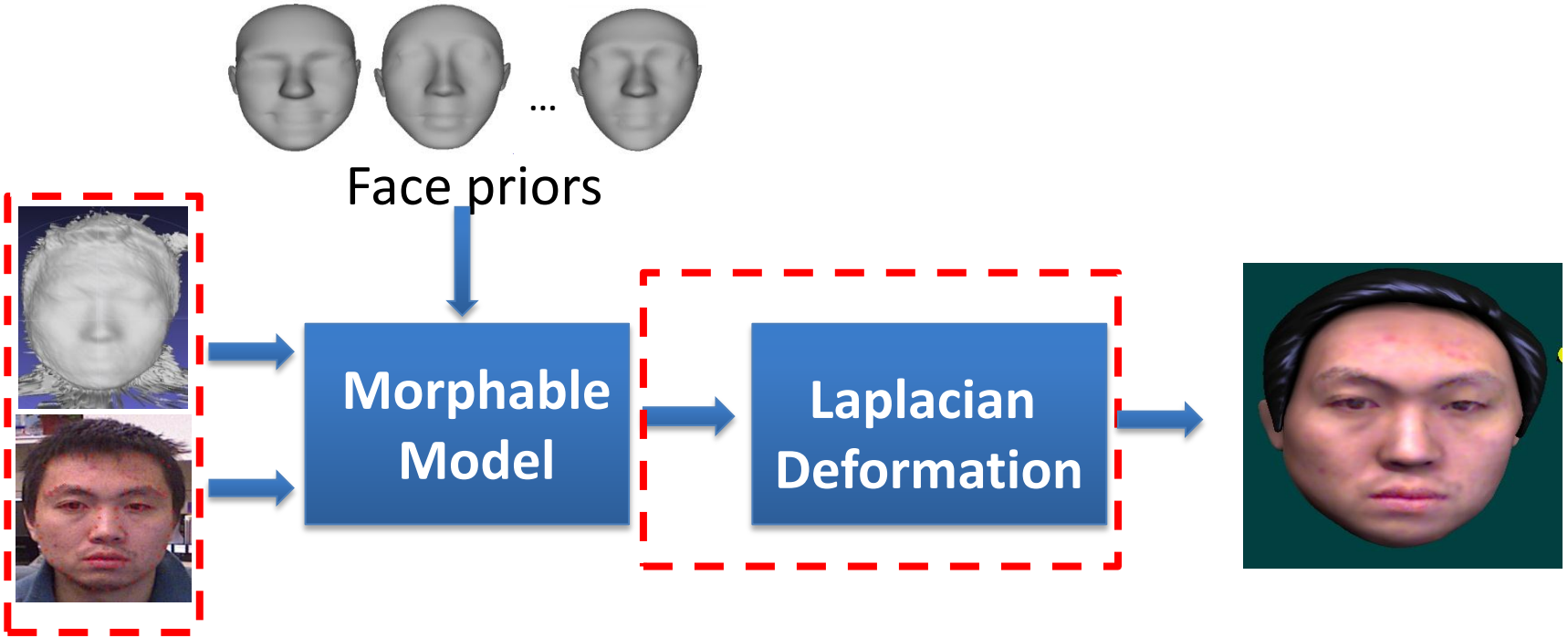
Face priors

Image

2D feature points

Morphable Model

Kinect-based 3D Face Modeling



Comparison

Photo

Image-based

Our approach

What about hair?

Single-View Hair Modeling for Portrait Manipulation

Submitted to ACM SIGGRAPH 2012
Online Submission ID: 0424

Our Vision

Our Vision

- Is to do vision

Our Vision

- Is to do vision, graphics & multimedia

Our Vision

- Is to do vision, graphics & multimedia

- Make technologies disappear

Future of Kinect: turning Sci-Fi into reality

Beyond Kinect: Cloud + NUI

- Old era of “PC + GUI”
  - PC has processing powers
  - GUI allows easy access to the processing power
- New era of “Cloud + NUI”
  - The cloud has knowledge & data
  - NUI allows easy access to the knowledge & data

Where are we today?

Kinect is the“Kitty Hawk”of NUI

Emerging Research Themes

- Tracking
  - identity tracking,
  - facial features tracking,
  - head pose tracking
  - …
- Gesture control
- Digitization

To Recap …

- Kinect from insiders'perspective
- Core technologies of Kinect
- Emerging research themes in NUI
- A great new engineering discipline -- huge opportunities, huge challenges

Thank You!

Microsoft Research

Microsoft

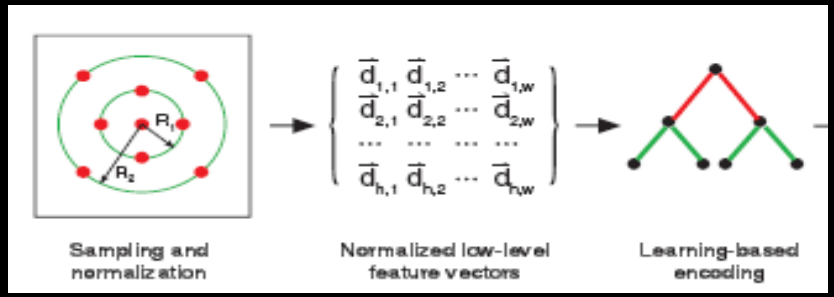
Facial signature

- Face detection
- Face alignment
- Signature extraction

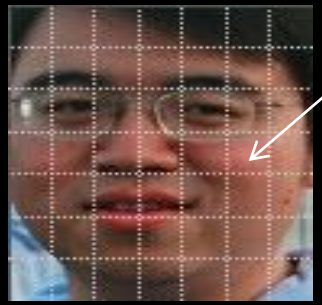
- <5 ms

Facial signature

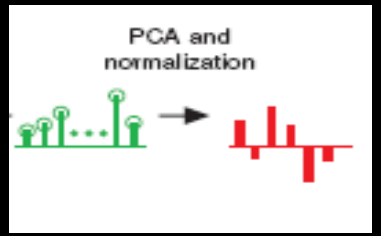
- Signature is learned from data



Learning-based local descriptor



Concatenated patch histogram



PCA dimension reduction

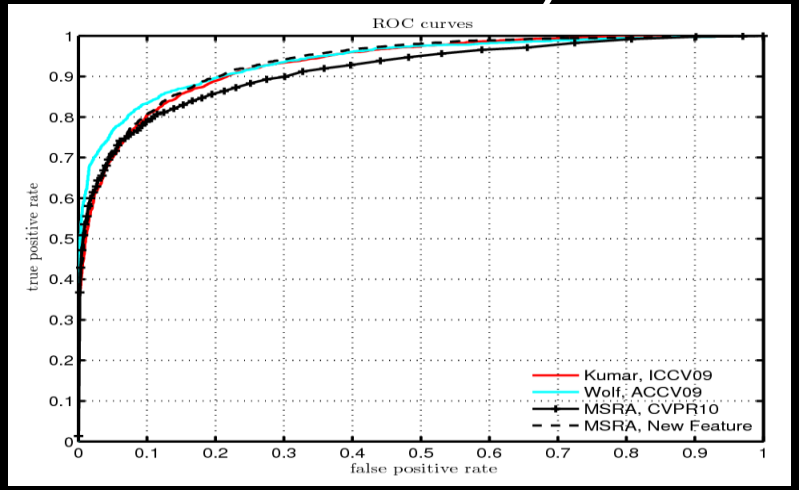
Zhimin Cao, Qi Yin, Xiaoou Tang, and Jian Sun. Face Recognition with Learning based Descriptor. CVPR 2010.

Facial signature
state-of-the-art in face recognition

- Top #1 in LFW face recognition benchmark
- Microsoft FaceLibrary (http://toolbox/Facelib/)

Implementation is adapted for Kinect

  - Kinect camera
    - lighting, resolution…



Algorithm: training

1. LBP feature extraction (dim = 7139)
2. PCA (dim = 2500)
3. Multi-class LDA (dim = dozens)
4. Clustering to find a small number of exemplars (cluster centers)