Eye Gaze Tracking Using an RGBD Camera: A Comparison with an RGB Solution

Xuehan Xiong (CMU), Qin Cai, Zicheng Liu, Zhengyou Zhang

Microsoft Research, Redmond, WA, USA

zhang@microsoft.com
http://research.microsoft.com/~zhang/

Microsoft Research

The 4th International Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction (PETMEI 2014)

Outline

- Goal and motivation
- Challenges
- Approach
- Results

Goals and motivations

1. Kinect-based eye tracking

2. Comparison between RGBD and RGB alone

Goals and motivations

- Most commercial eye trackers are IR-based
  - Short range
  - Does not work outdoor
- Non-IR based system
  - Outdoor
  - Cheaper
  - Better capability of being integrated
  - Less accurate

Outline

- Motivation
- Challenges
- Approach
- Results

Challenges

- Eye images from IR-based approaches

- Eye images from Kinect

Outline

- Motivation
- Challenges
- Approach
- Results

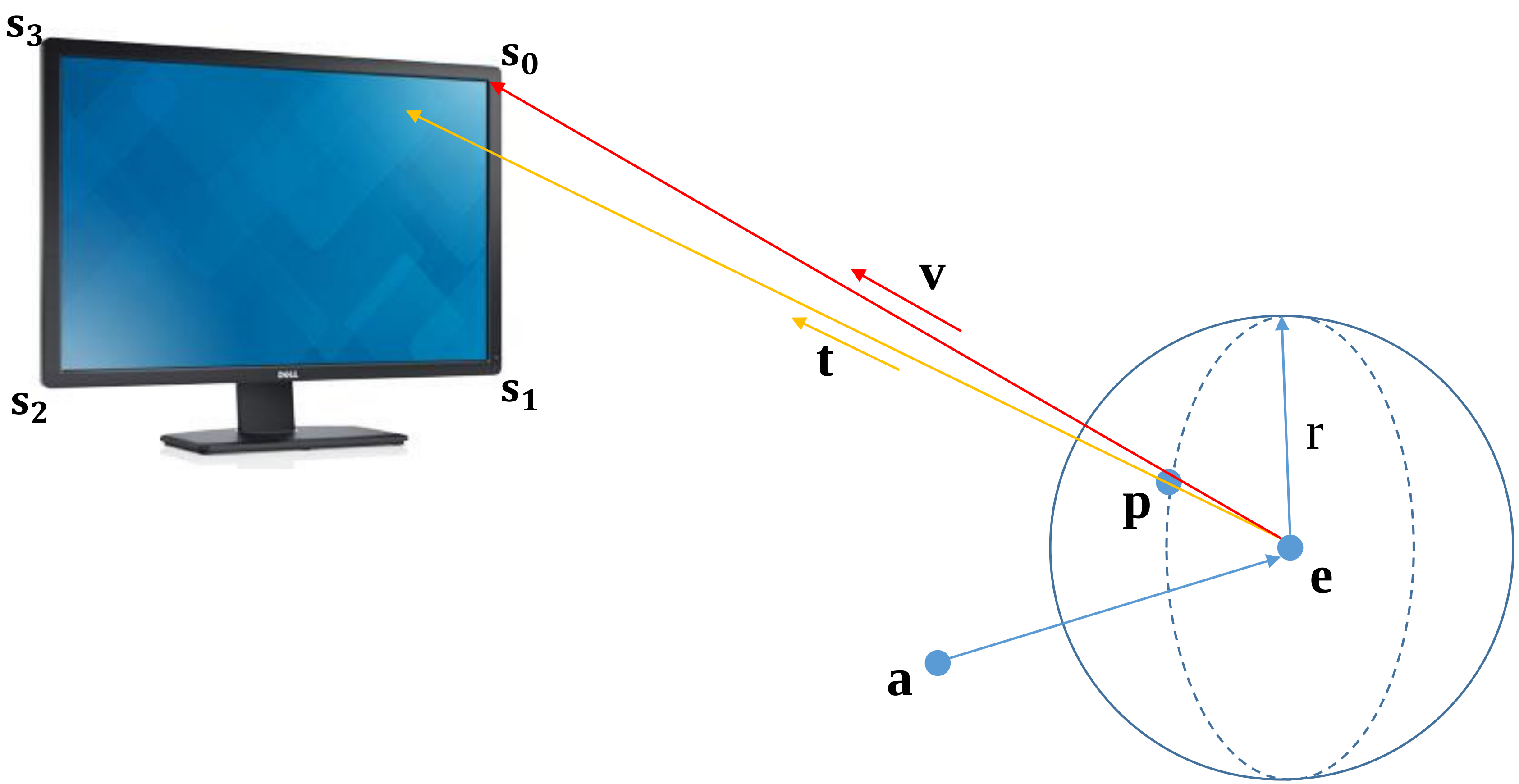
Approach

- What is gaze (in our model)?

Notation:
p -- pupil
v -- visual axis
t -- optical axis
$\mathbf{R}_{vo}$ -- rotation compensation b/w **v** and **t**
$\mathbf{v} = \mathbf{R}_{vo}\mathbf{t}$

a -- head center
$\overrightarrow{\mathbf{ae}}$ -- offset
$\mathbf{R}_{hp}$ -- head rotation
r – eyeball radius

Eyeball center:
$\mathbf{e} = \mathbf{a} + \mathbf{R}_{hp}\overrightarrow{\mathbf{ae}}$



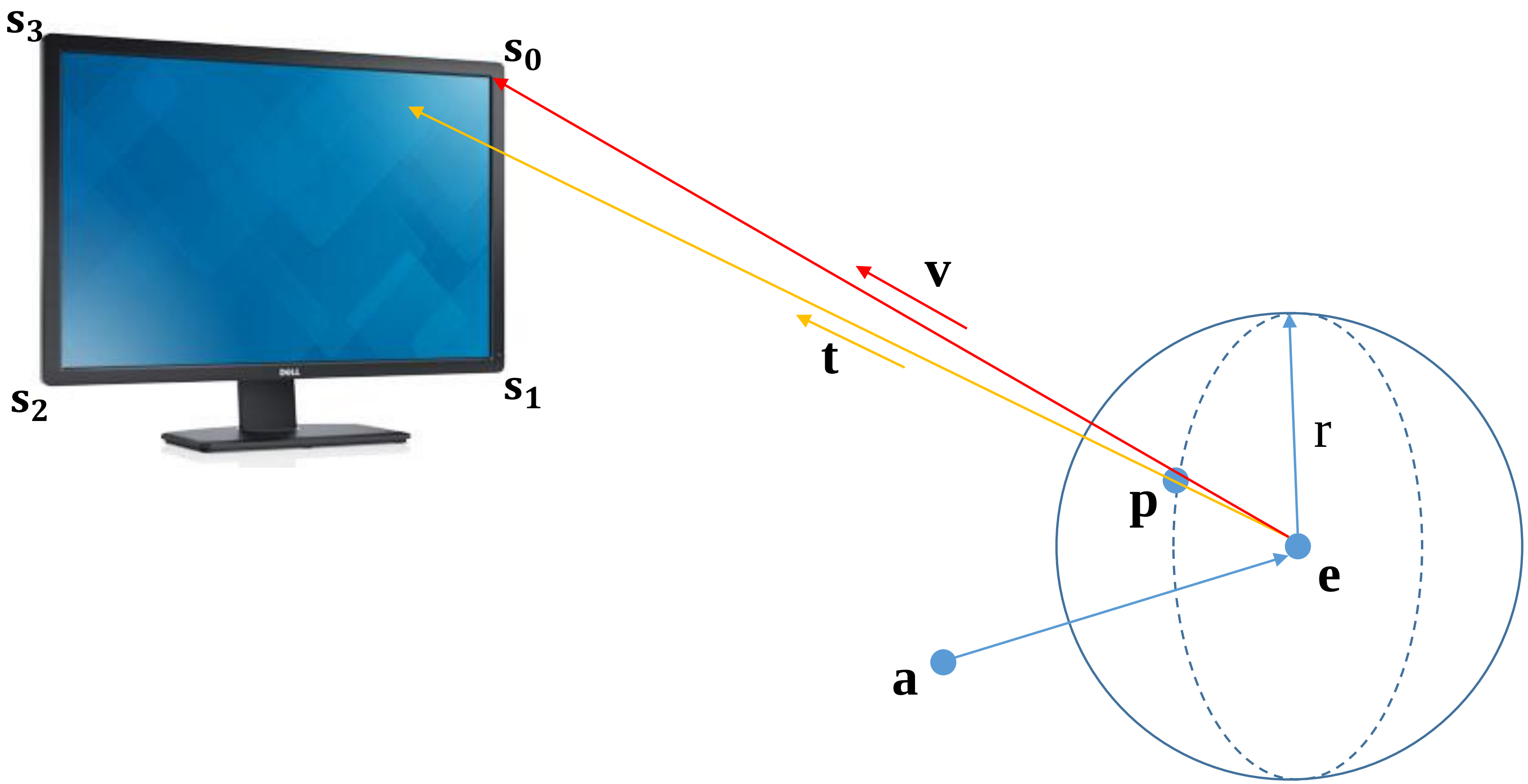
Approach

- What are fixed (in our model)?

Notation:
p -- pupil
v -- visual axis
t -- optical axis
$\mathbf{R}_{vo}$ -- rotation compensation
b/w **v** and **t**
$\mathbf{v} = \mathbf{R}_{vo}\mathbf{t}$

a -- head center
$\overrightarrow{\mathbf{ae}}$ -- offset
$\mathbf{R}_{hp}$ -- head rotation
r – eyeball radius

Eyeball center:
$\mathbf{e} = \mathbf{a} + \mathbf{R}_{hp}\overrightarrow{\mathbf{ae}}$



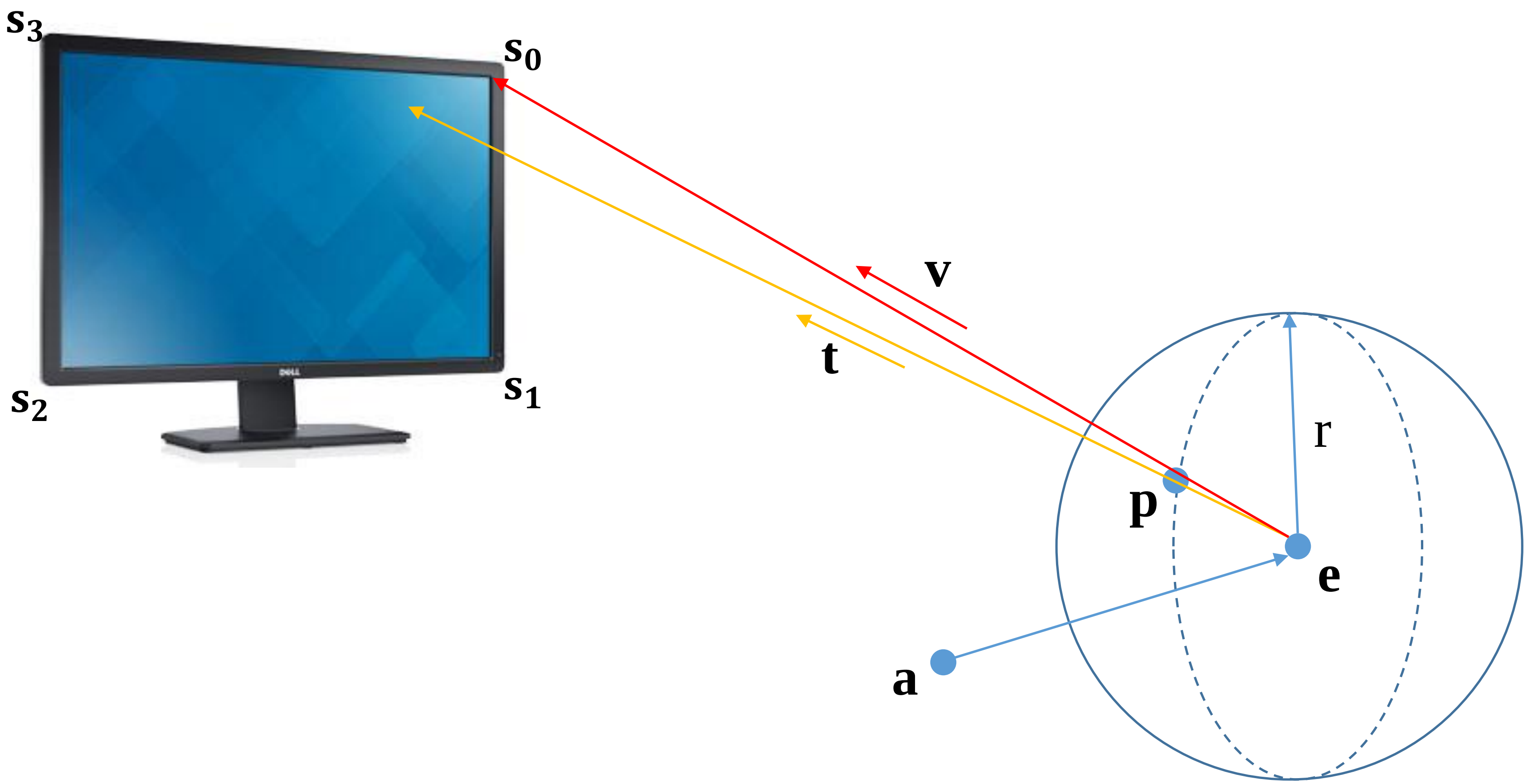
Approach

- What to be measured (in our model)?

Notation:
p -- pupil
v -- visual axis
t -- optical axis
$\mathbf{R_{vo}}$ -- rotation compensation b/w **v** and **t**
$\mathbf{v} = \mathbf{R_{vo}}\mathbf{t}$

a -- head center
$\overrightarrow{\mathbf{ae}}$ -- offset
$\mathbf{R_{hp}}$ -- head rotation
r – eyeball radius

Eyeball center:
$\mathbf{e} = \mathbf{a} + \mathbf{R_{hp}}\overrightarrow{\mathbf{ae}}$



Approach

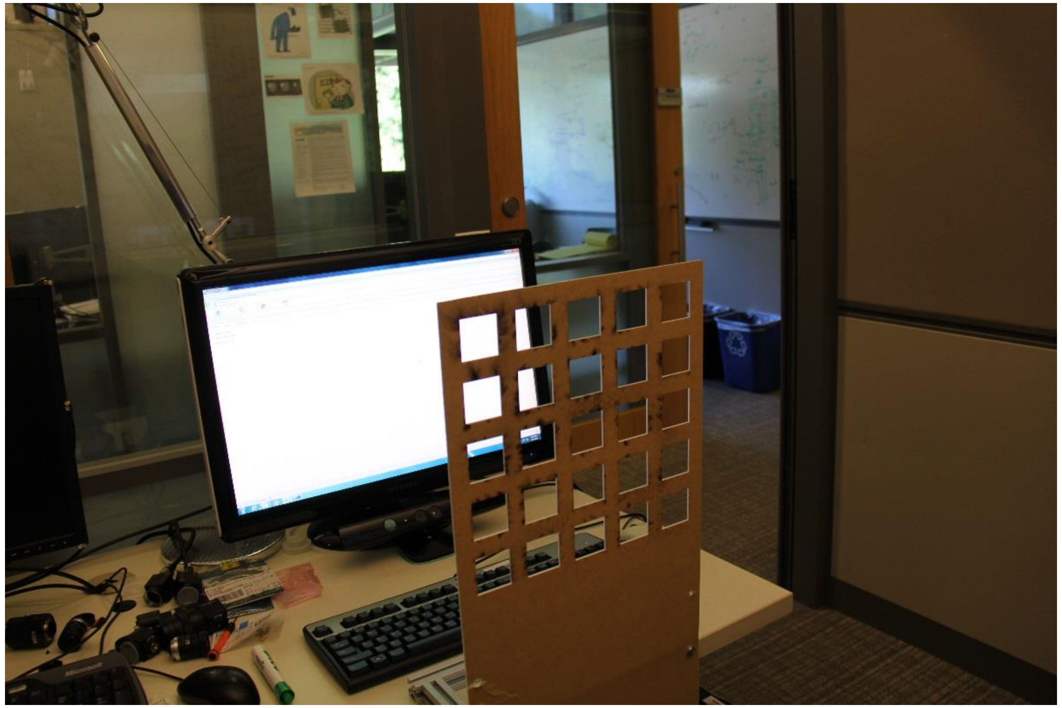
- System calibration
- Head pose
- Head center
- Pupil
- User calibration

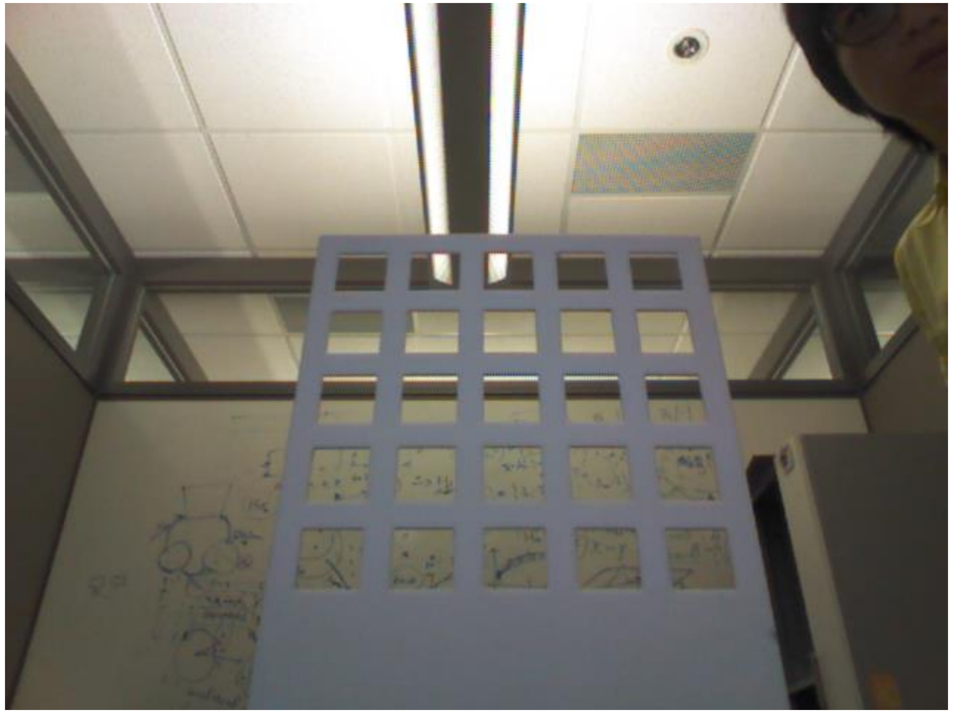
System calibration

- World = color camera
  - Intrinsic parameters, centered at [0,0,0]
- Depth camera
  - Intrinsic and extrinsic parameters
- Monitor screen
  - Screen-camera calibration

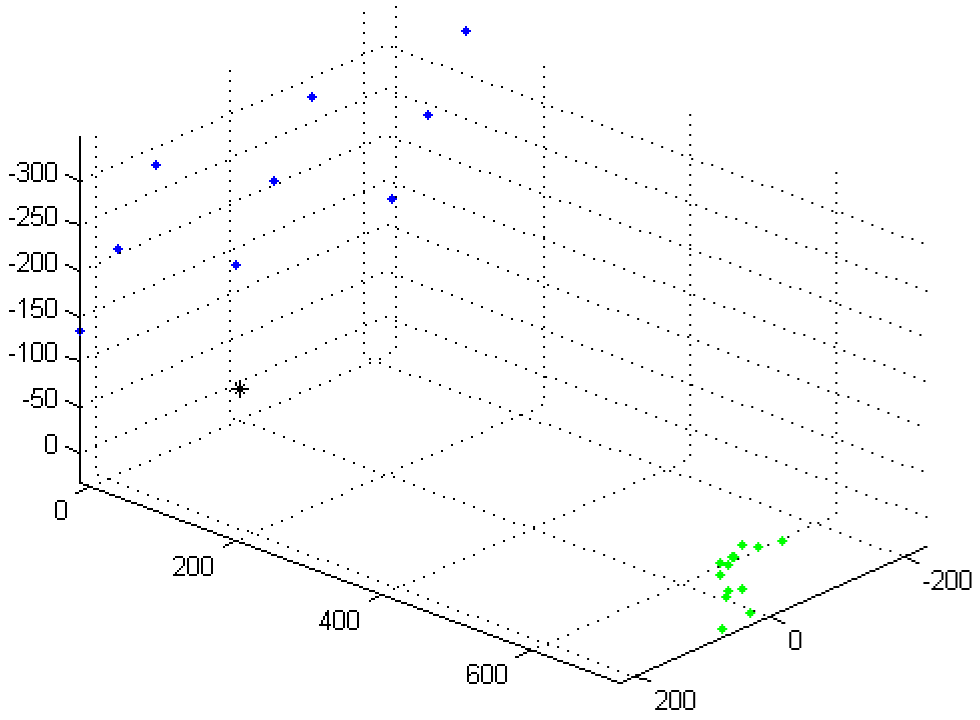
Screen-camera calibration

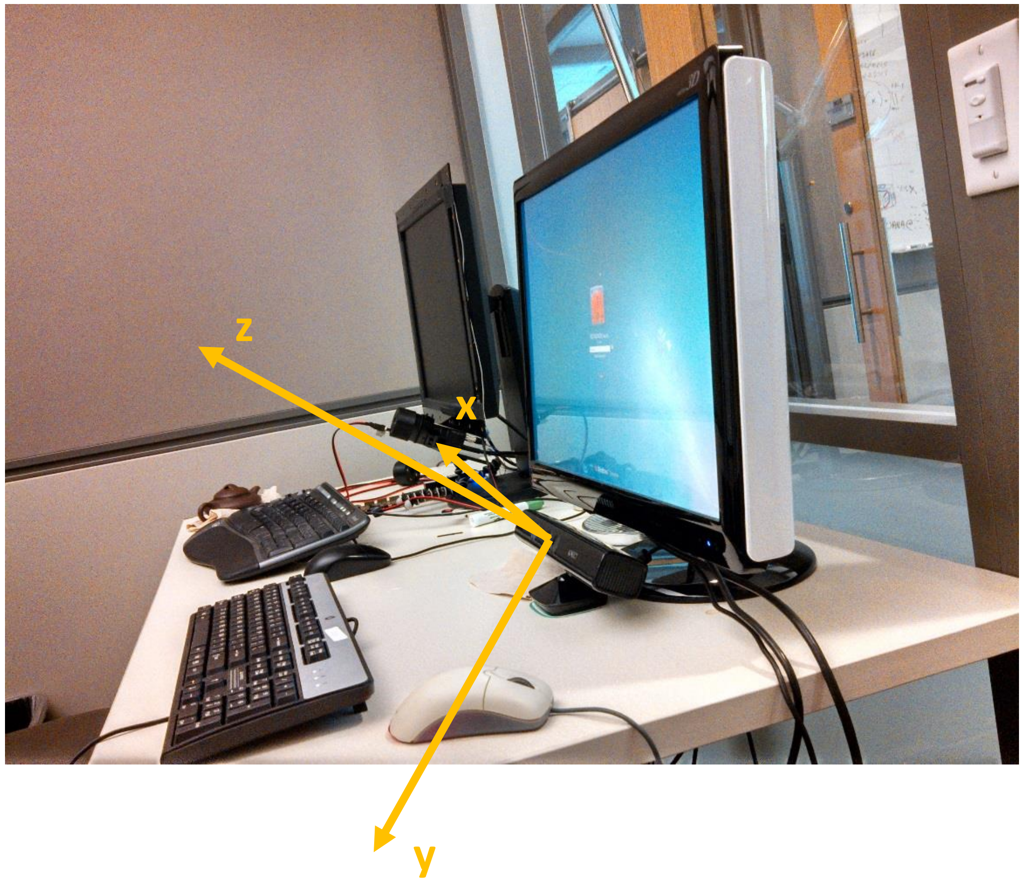
- 4 images capturing screen + pattern
- 1 image from Kinect camera capturing the pattern





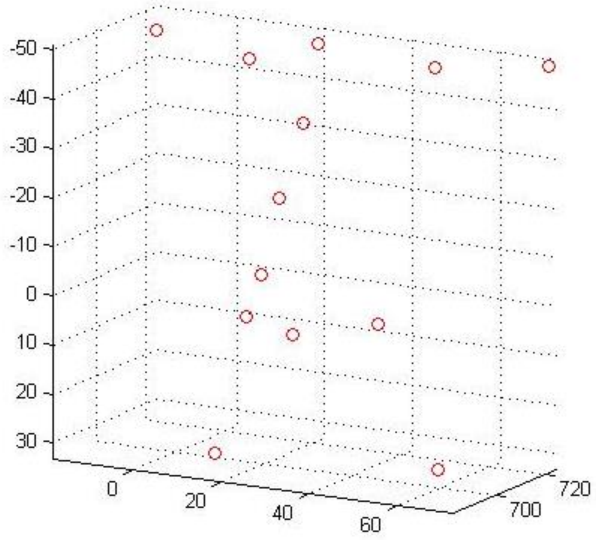
Calibration results





Head pose estimation

- Build a person-specific 3D face model

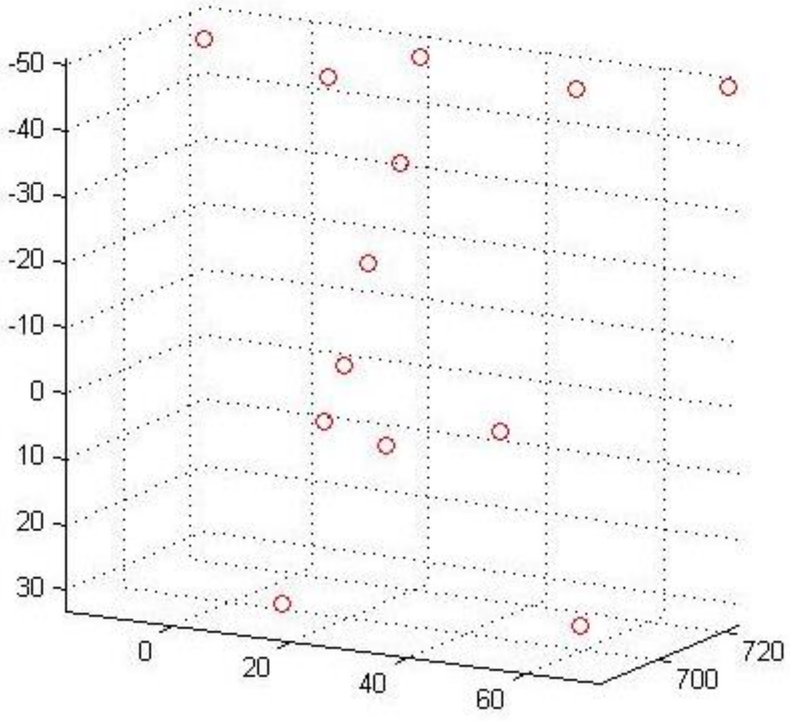


Average over 10 frames

Rigid points

Head pose estimation

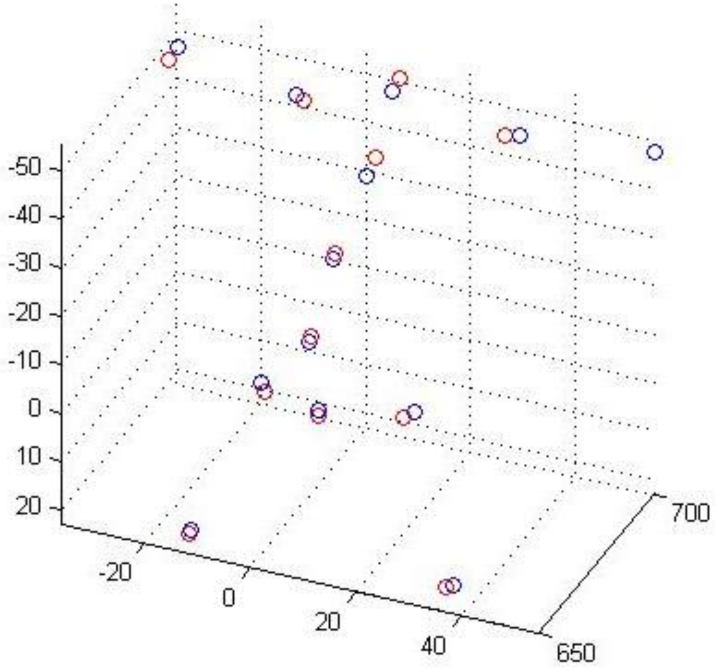
- For each frame t



Reference model

R,T

Procrustes

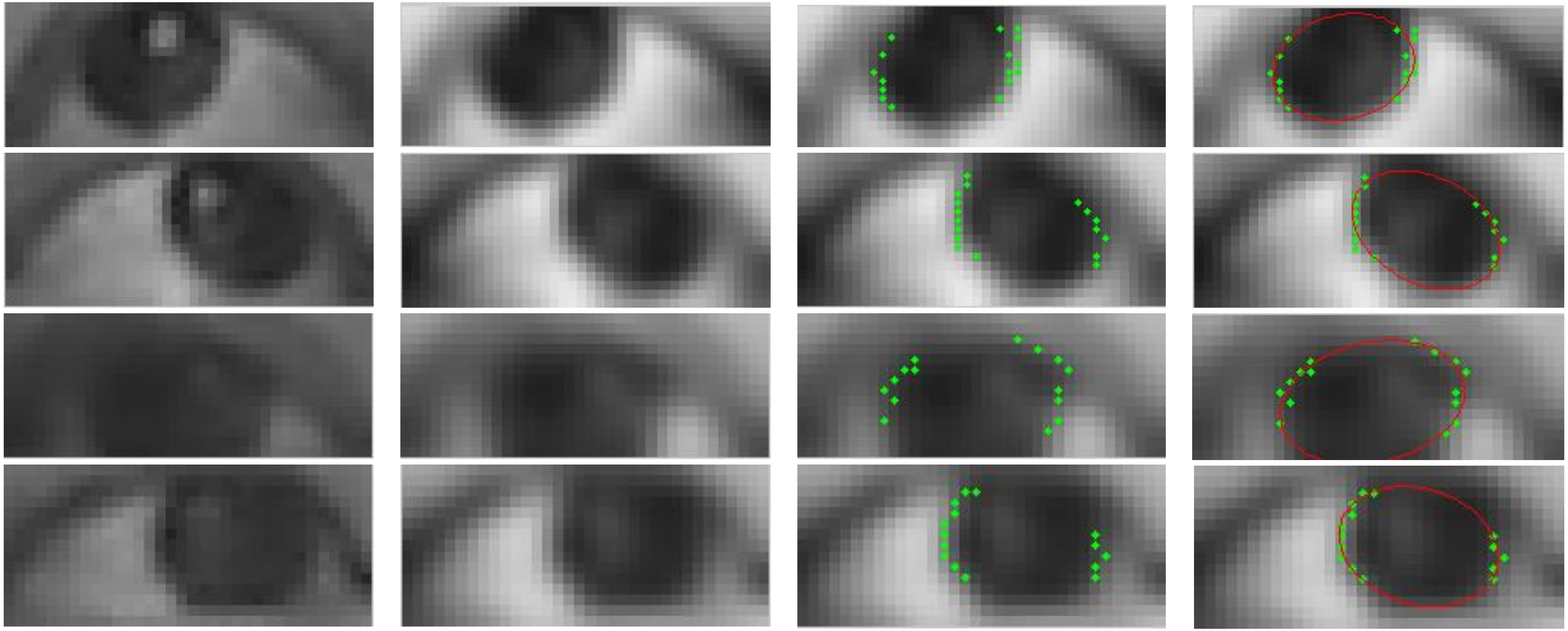


Red – Noisy
Blue – De-noised

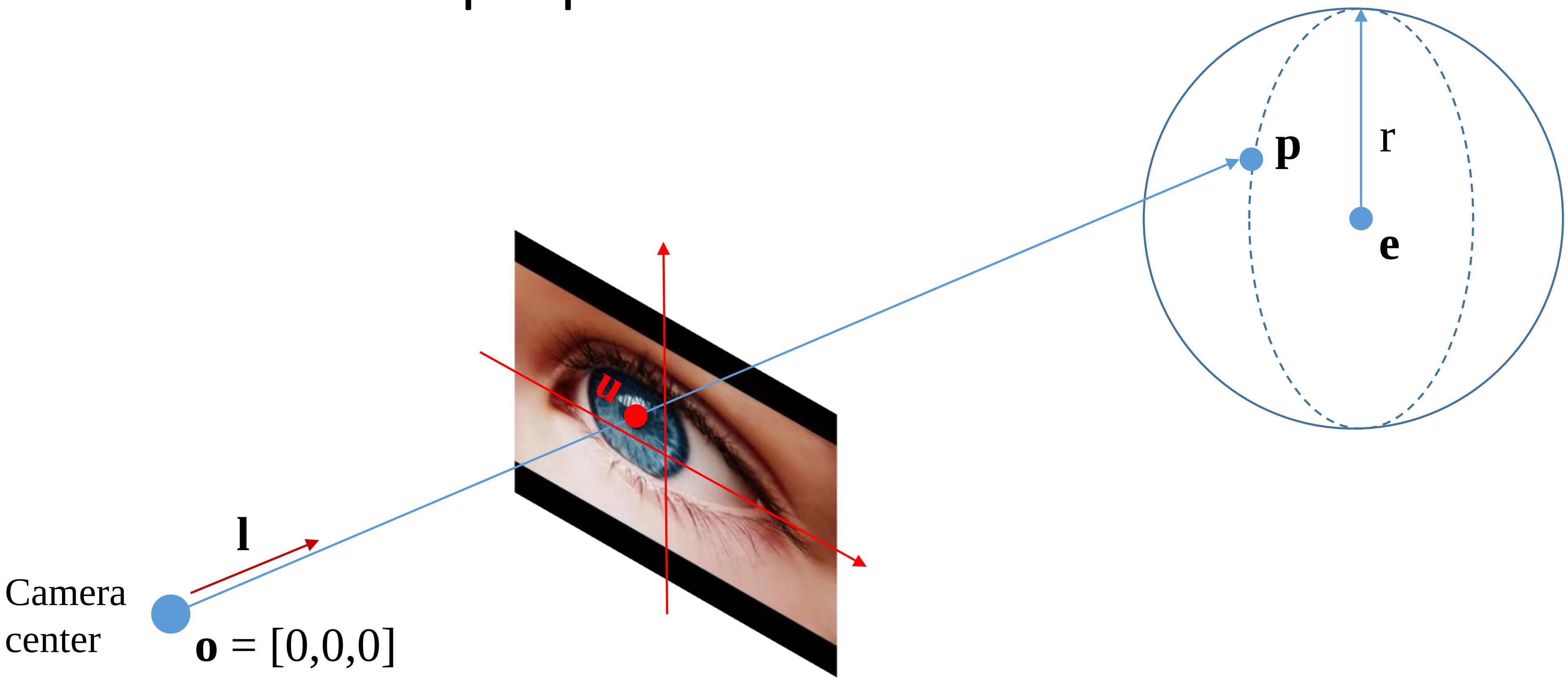
Head center

- The average of 13 landmarks

2D Iris detection



3D pupil estimation



$\mathbf{u} = [u, v, f\ ]^{\mathrm{T}}$ from camera intrinsic parameters

$$\mathbf{l} = \frac{\mathbf{u}}{\|\mathbf{u}\|}$$

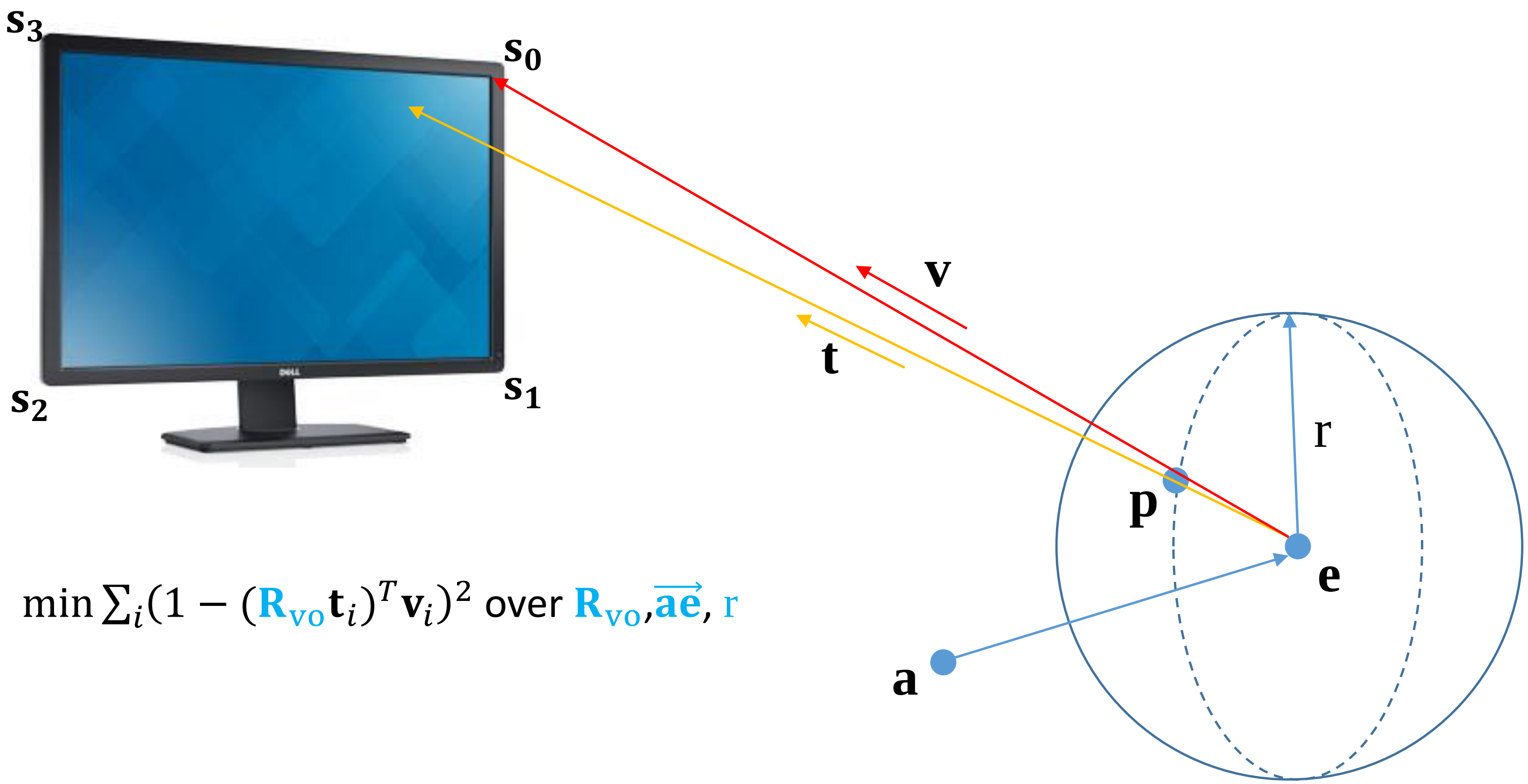
User calibration

- What are fixed (in our model)?

Notation:
p -- pupil
v -- visual axis
t -- optical axis
$\mathbf{R}_{\mathrm{vo}}$ -- rotation compensation
b/w **v** and **t**
$\mathbf{v} = \mathbf{R}_{\mathrm{vo}}\mathbf{t}$

a -- head center
$\overrightarrow{\mathbf{ae}}$ -- offset
$\mathbf{R}_{\mathrm{hp}}$ -- head rotation
r – eyeball radius

Eyeball center:
$\mathbf{e} = \mathbf{a} + \mathbf{R}_{\mathrm{hp}}\overrightarrow{\mathbf{ae}}$



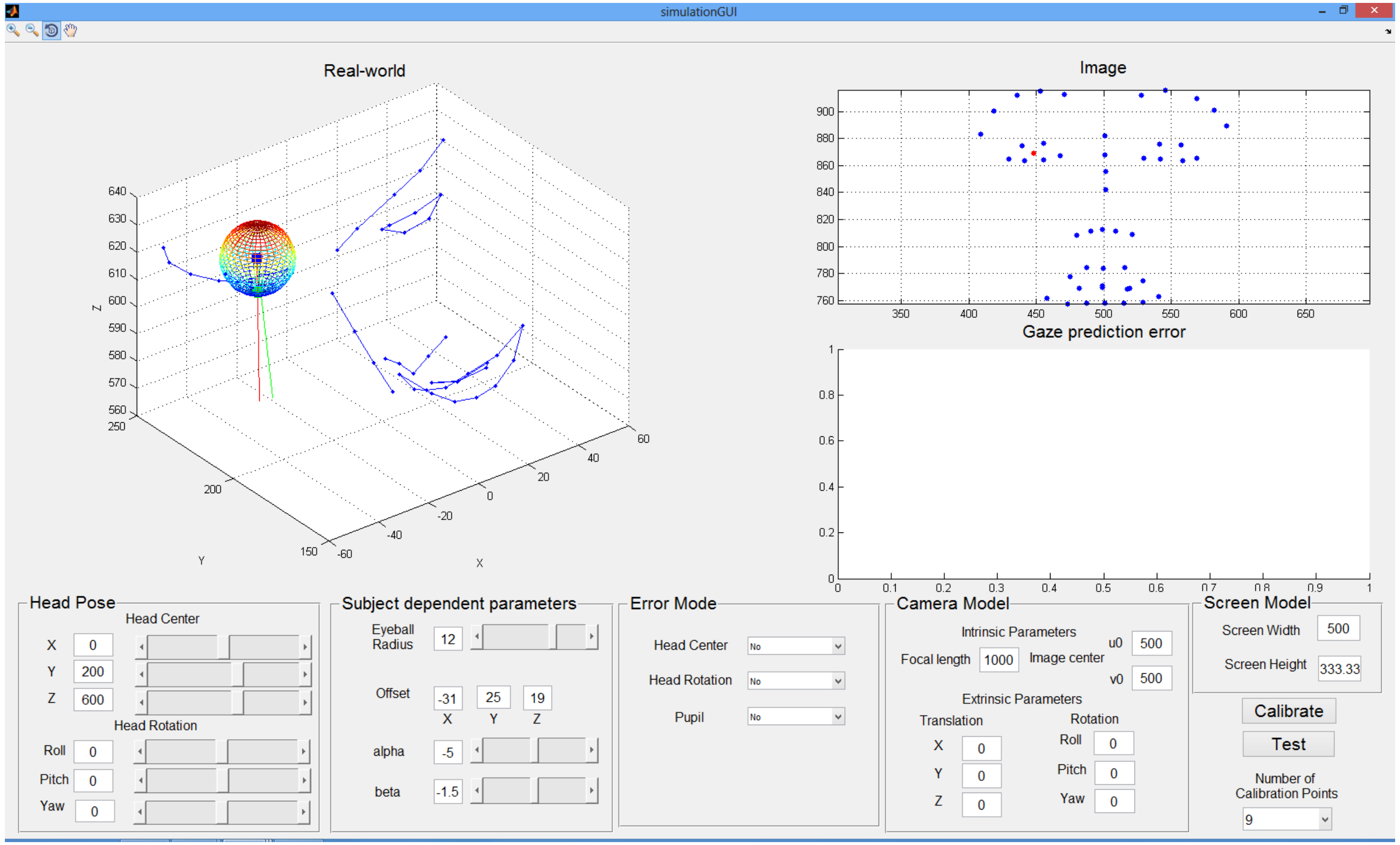
$\min \sum_i (1 - (\mathbf{R}_{\mathrm{vo}}\mathbf{t}_i)^T \mathbf{v}_i)^2$ over $\mathbf{R}_{\mathrm{vo}}, \overrightarrow{\mathbf{ae}}$, r

Outline

- Motivation
- Challenges
- Approach
- Results

Results

- Simulation

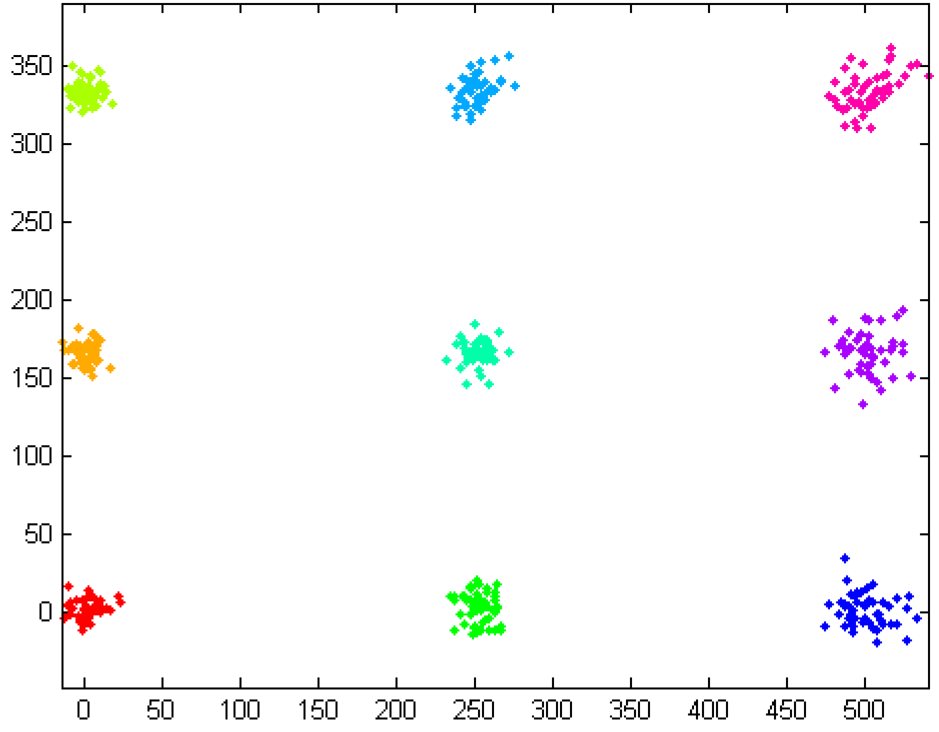


Error modeling

- Assuming perfect calibration (system and user)
- 3 sources of errors (assuming normal distribution with zero mean)
  - Head pose
  - Head center
  - Pupil
- Units
  - Head pose: degree
  - Head center: mm
  - Pupil: pixel

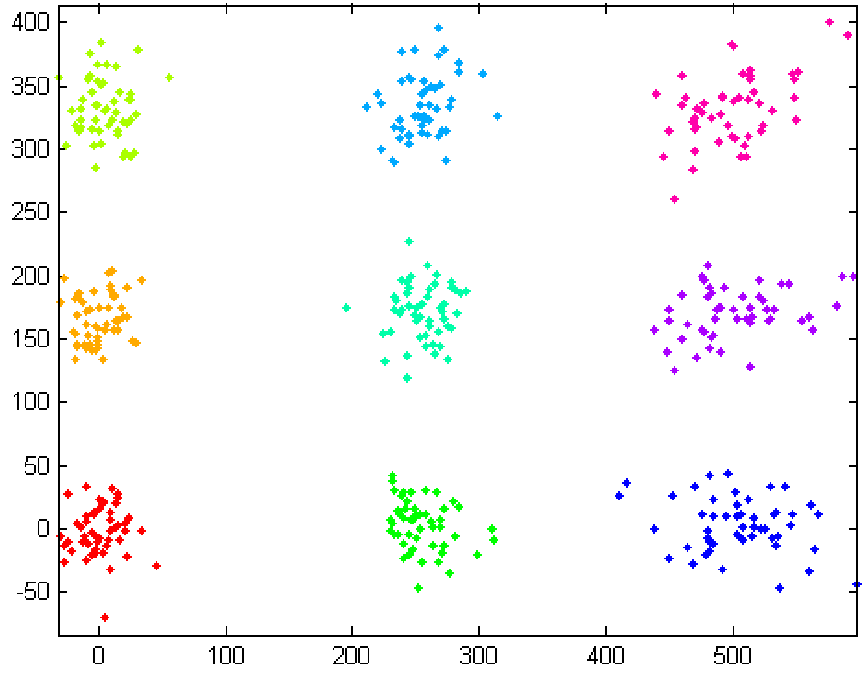
Simulation Result with low variances

- Variances – 0.1

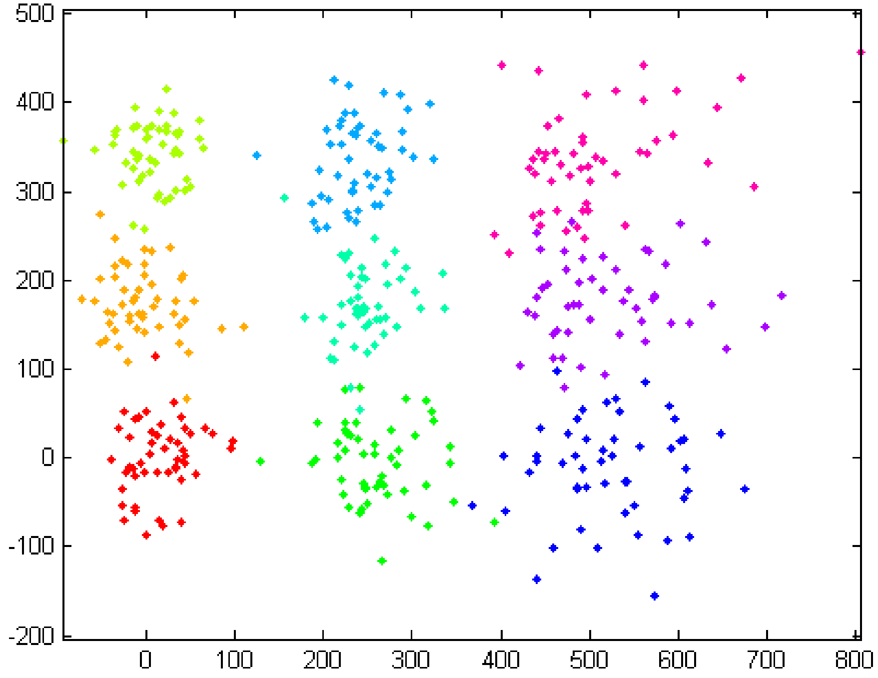


Back to reality

Variance – 0.25



Variance – 0.5



Real Data: Free head movement

(a)

(b)

9 calibration points

A subject with colored stickers

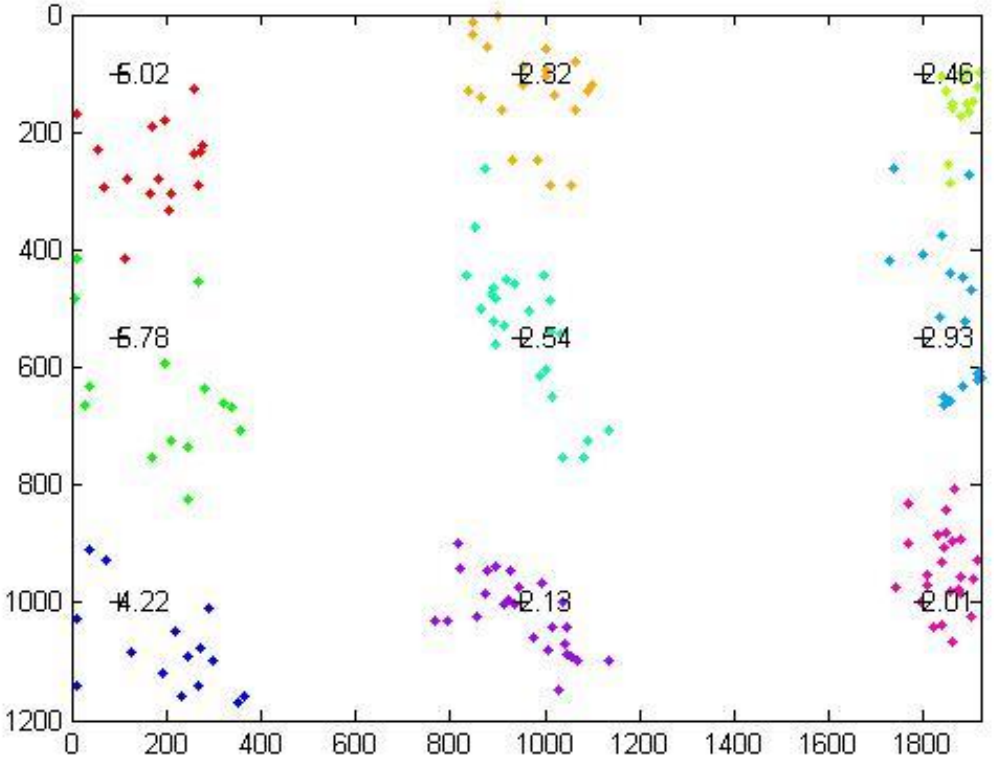
Experimental setup

- The monitor has a dimension of 520mm by 320mm.
- The distance between a test subject and the Kinect is between 600mm and 800mm.
- There are 9 subjects participated in the data collection.
- We collect three training sessions and two test sessions for each subject.

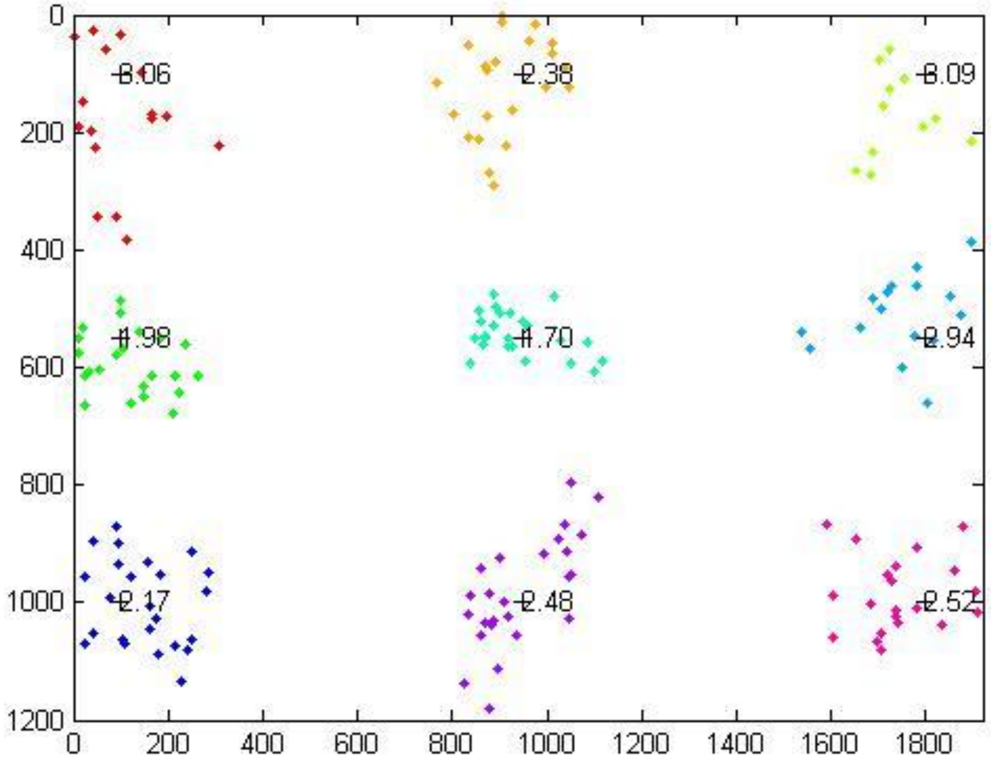
Best case scenario

Training error

Left eye

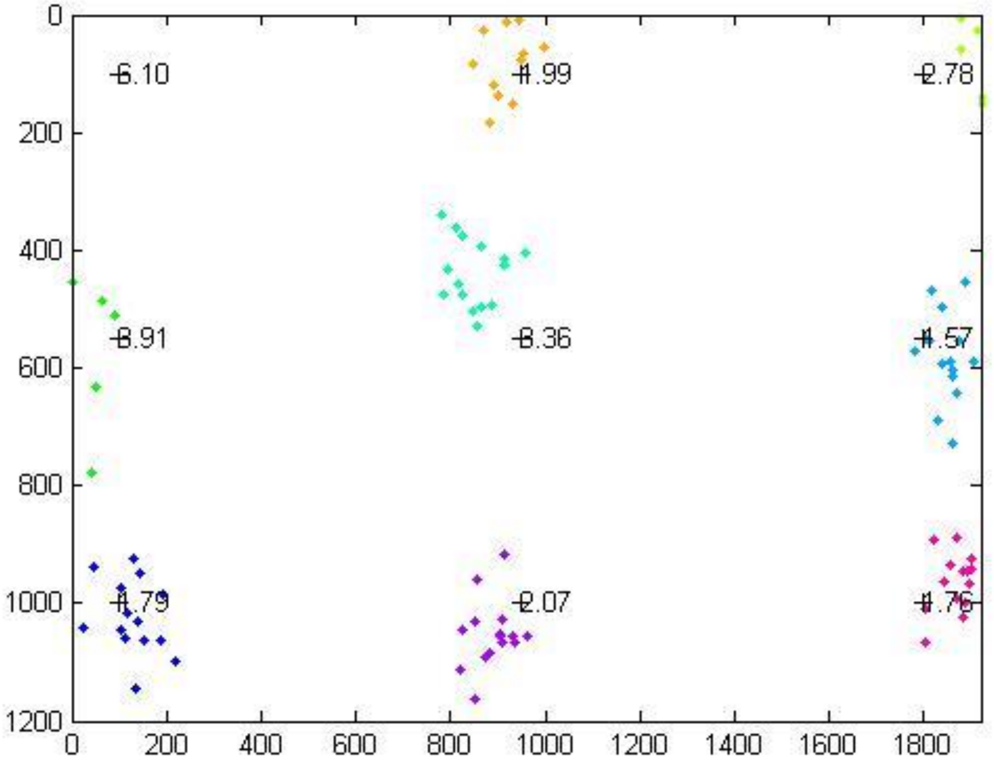


Right eye

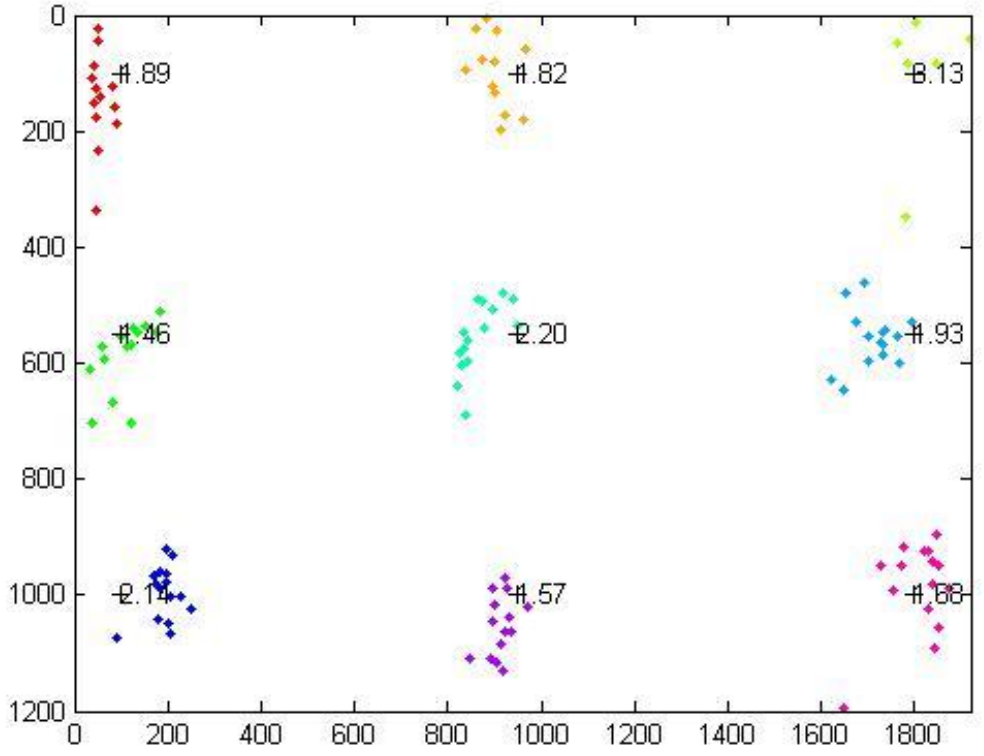


Testing error

Left eye

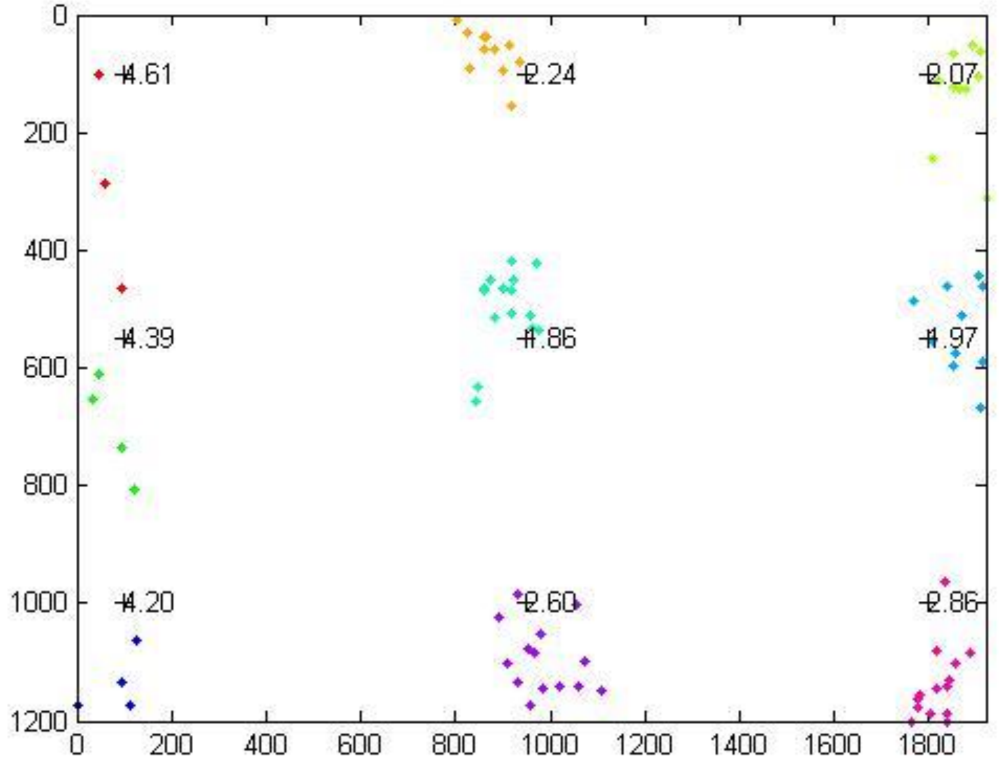


Right eye

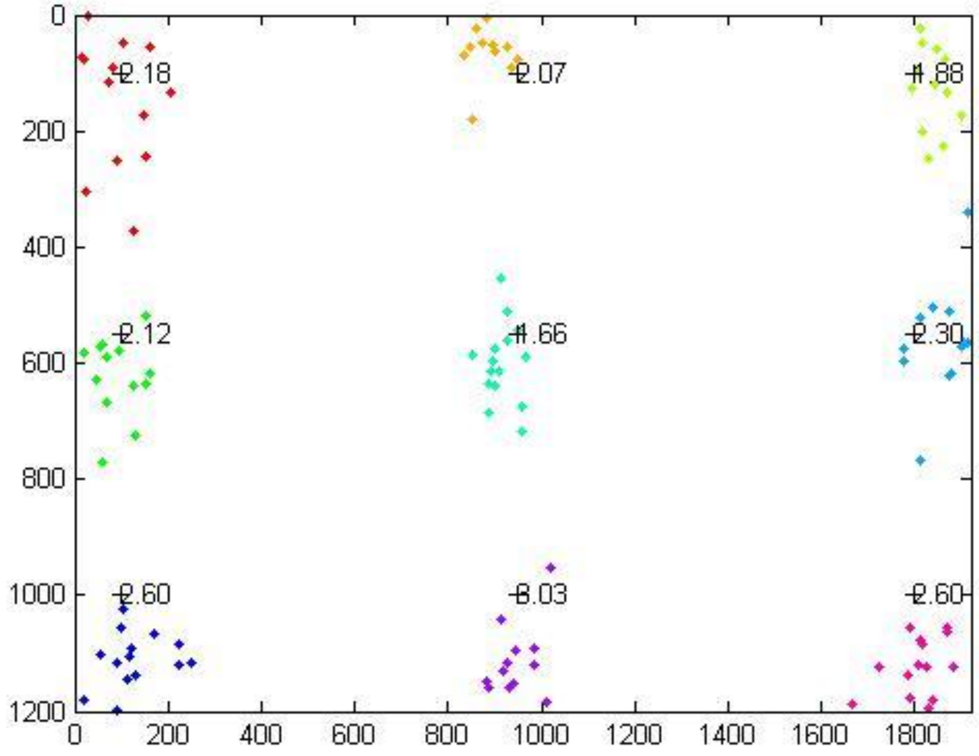


Testing error 2

Left eye



Right eye

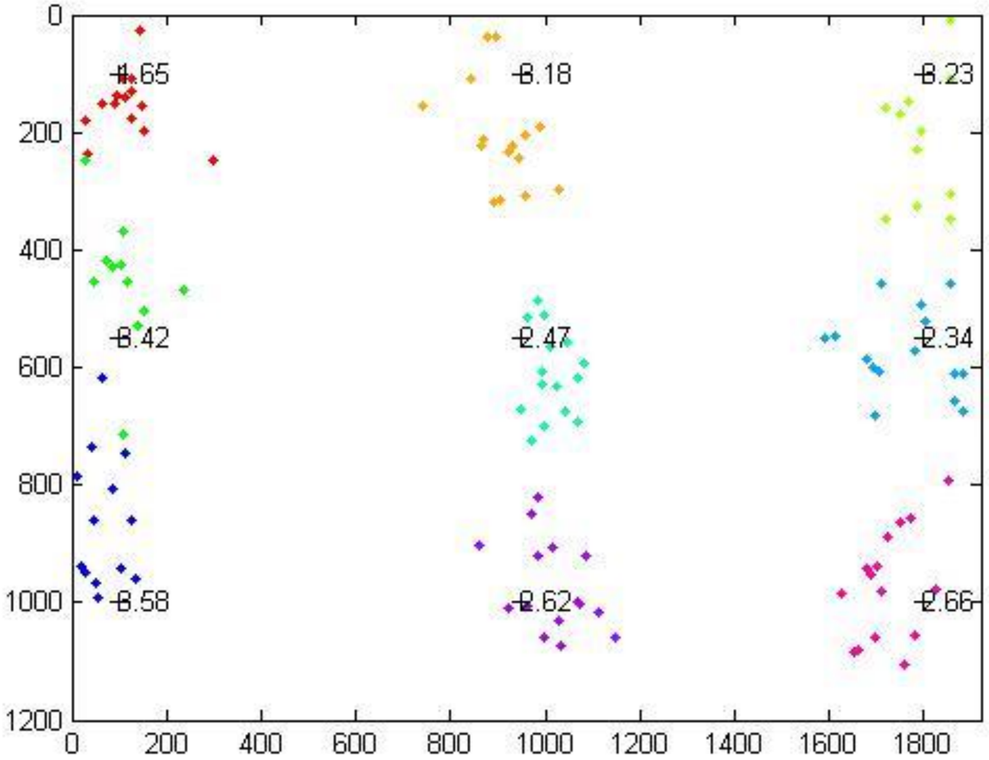


Sample Results Without Stickers

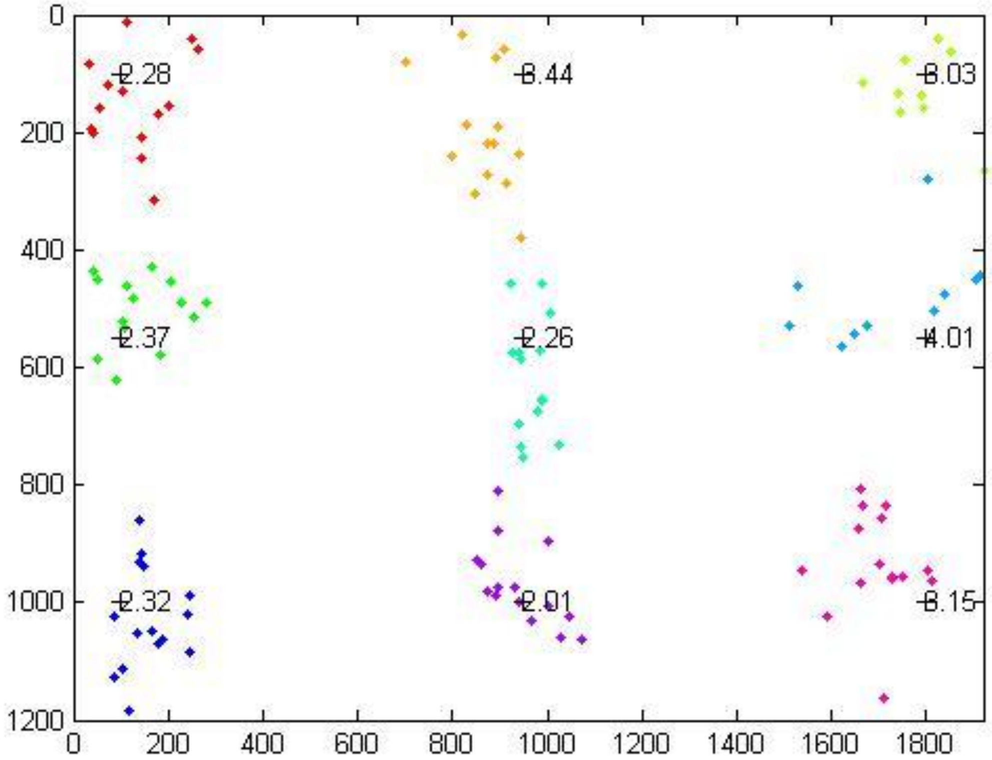
Qin

Qin – training error

Left eye

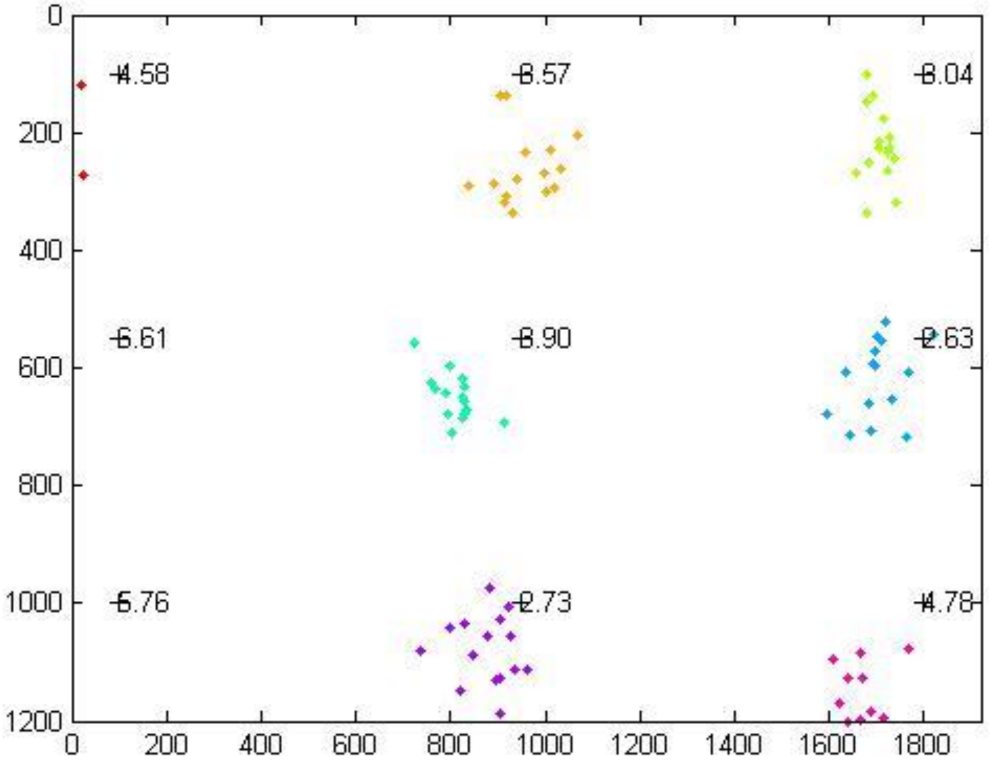


Right eye

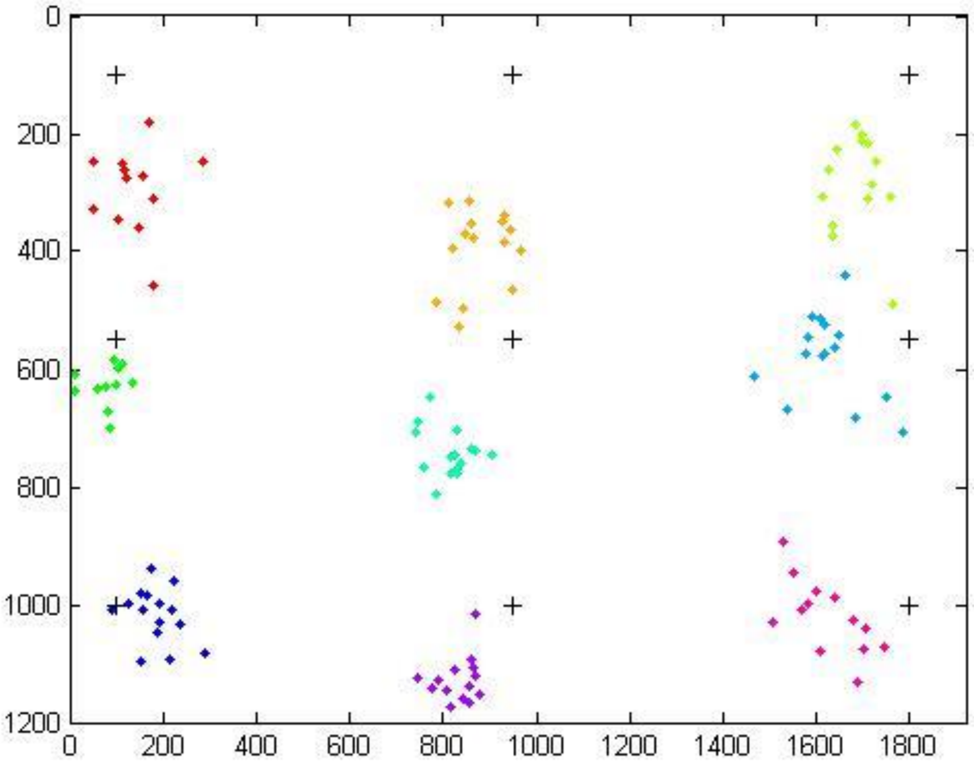


Qin – testing error

Left eye

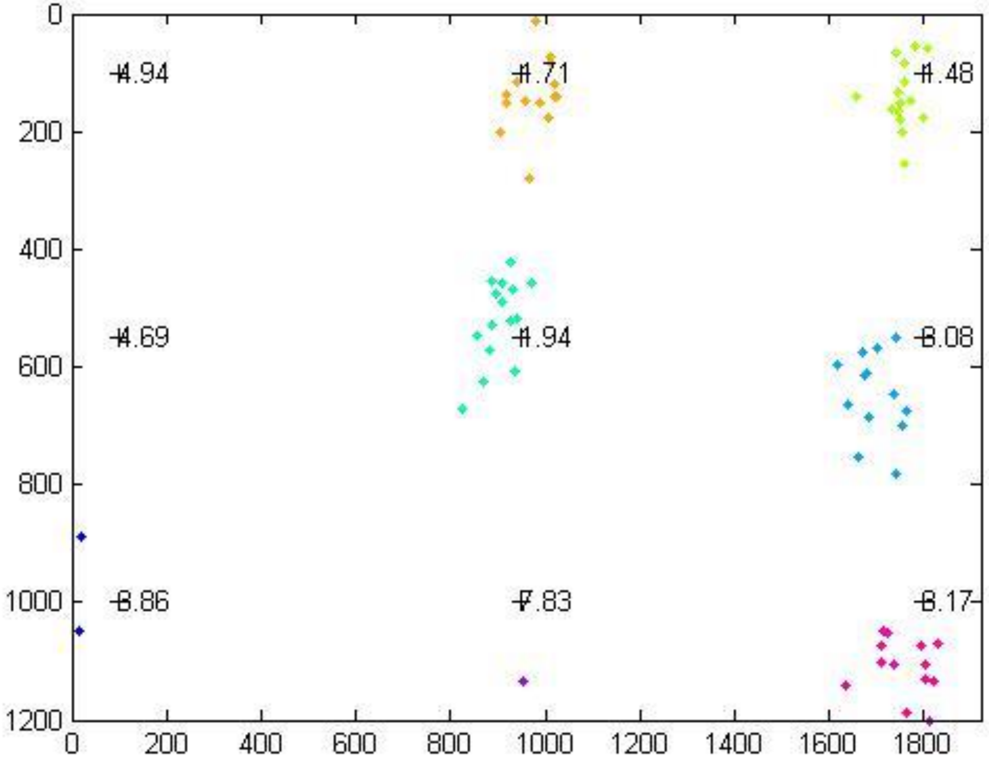


Right eye

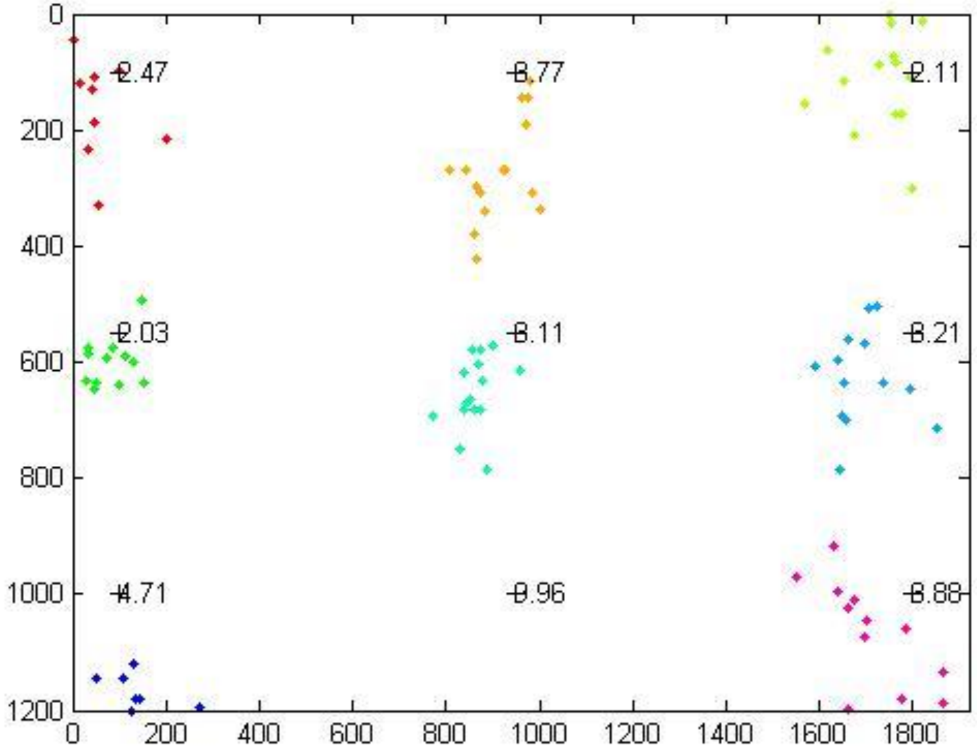


Qin – testing error 2

Left eye



Right eye

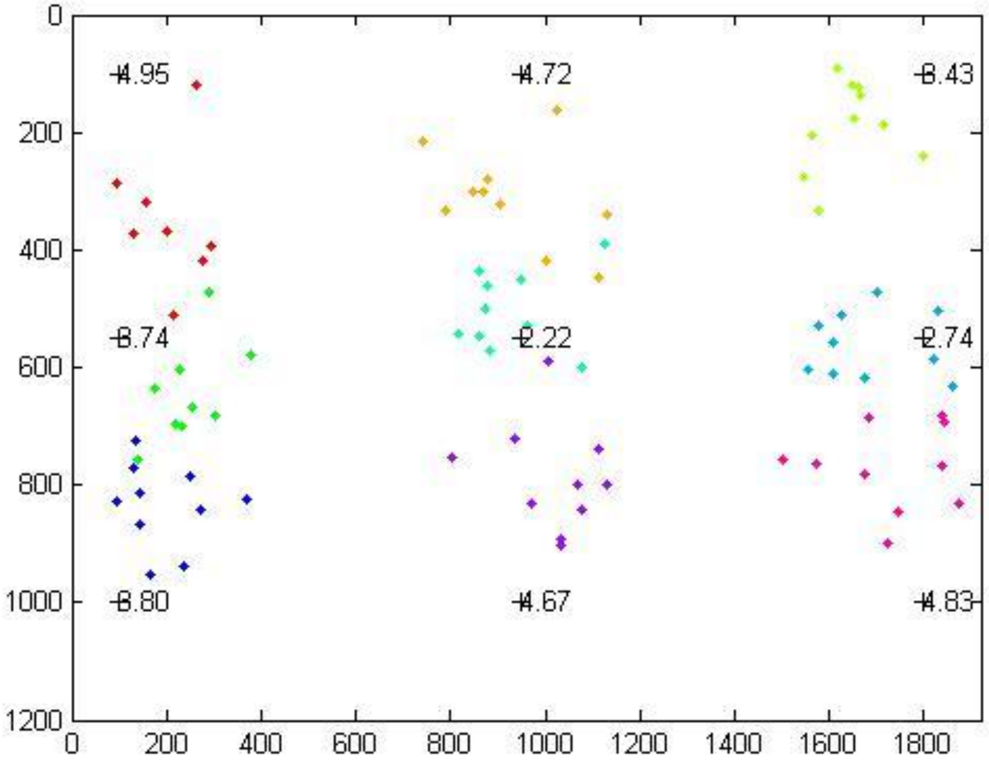


No (little) head movement

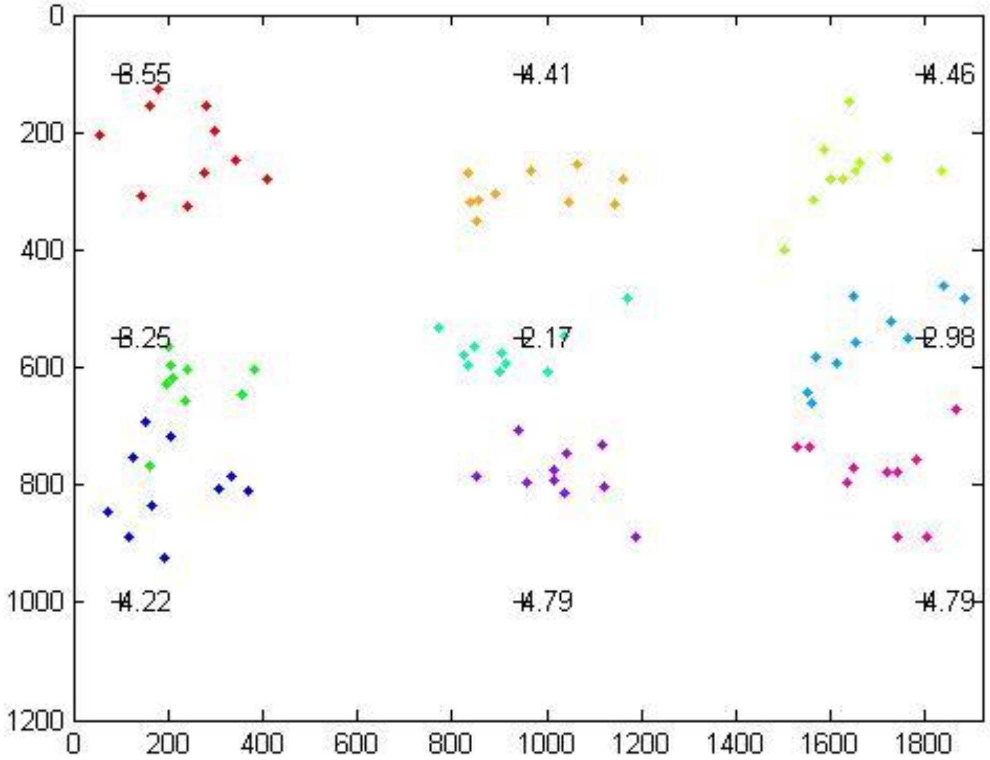
Best case scenario

Training error

Left eye



Right eye

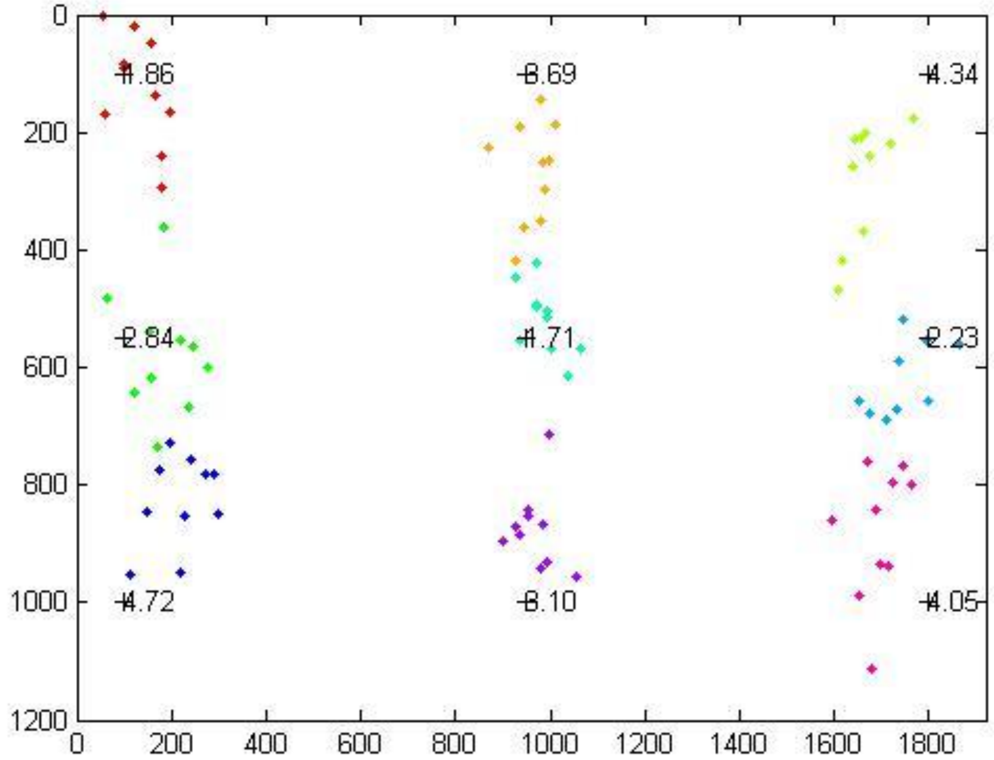


Sample Results Without Stickers

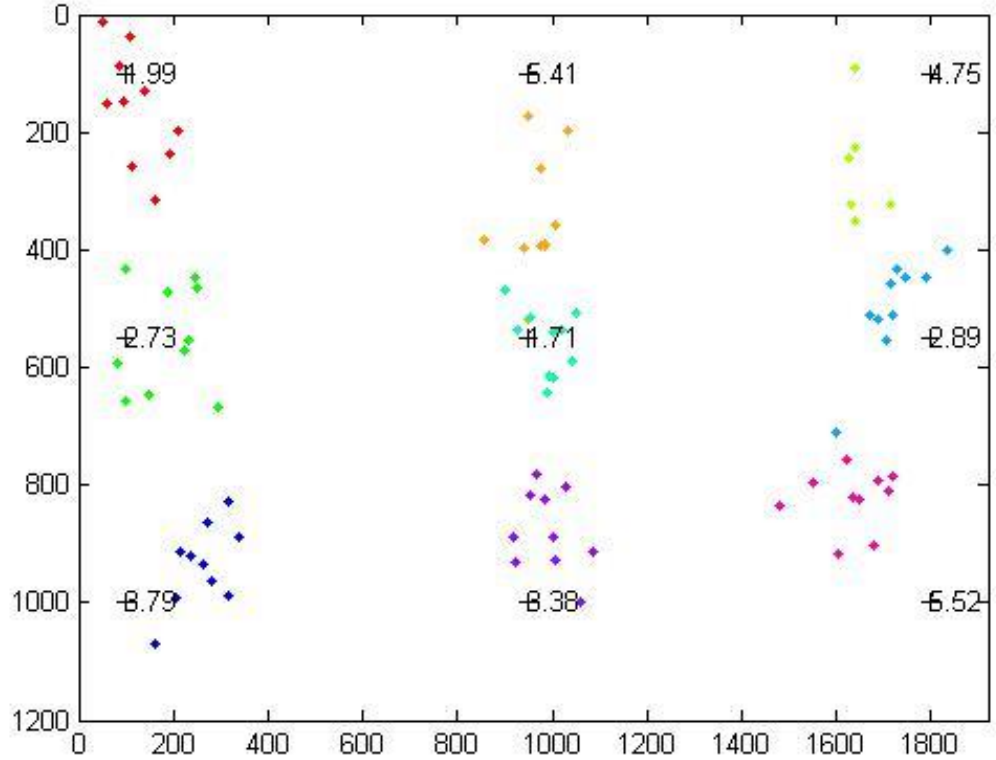
Qin

Qin – training error

Left eye

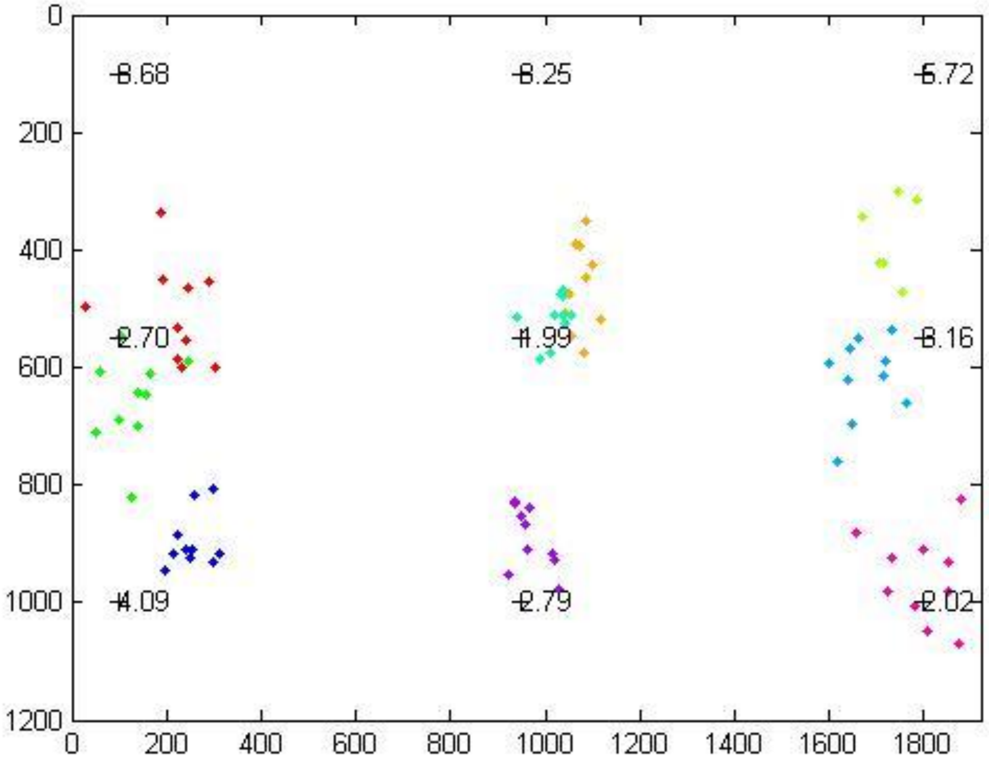


Right eye

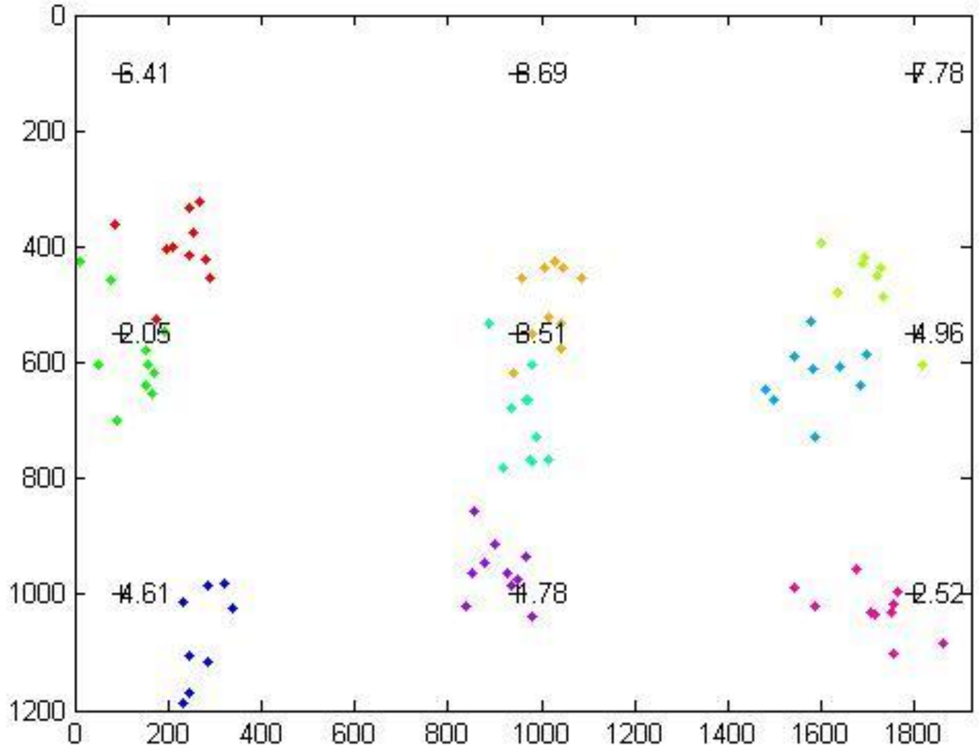


Qin – testing error

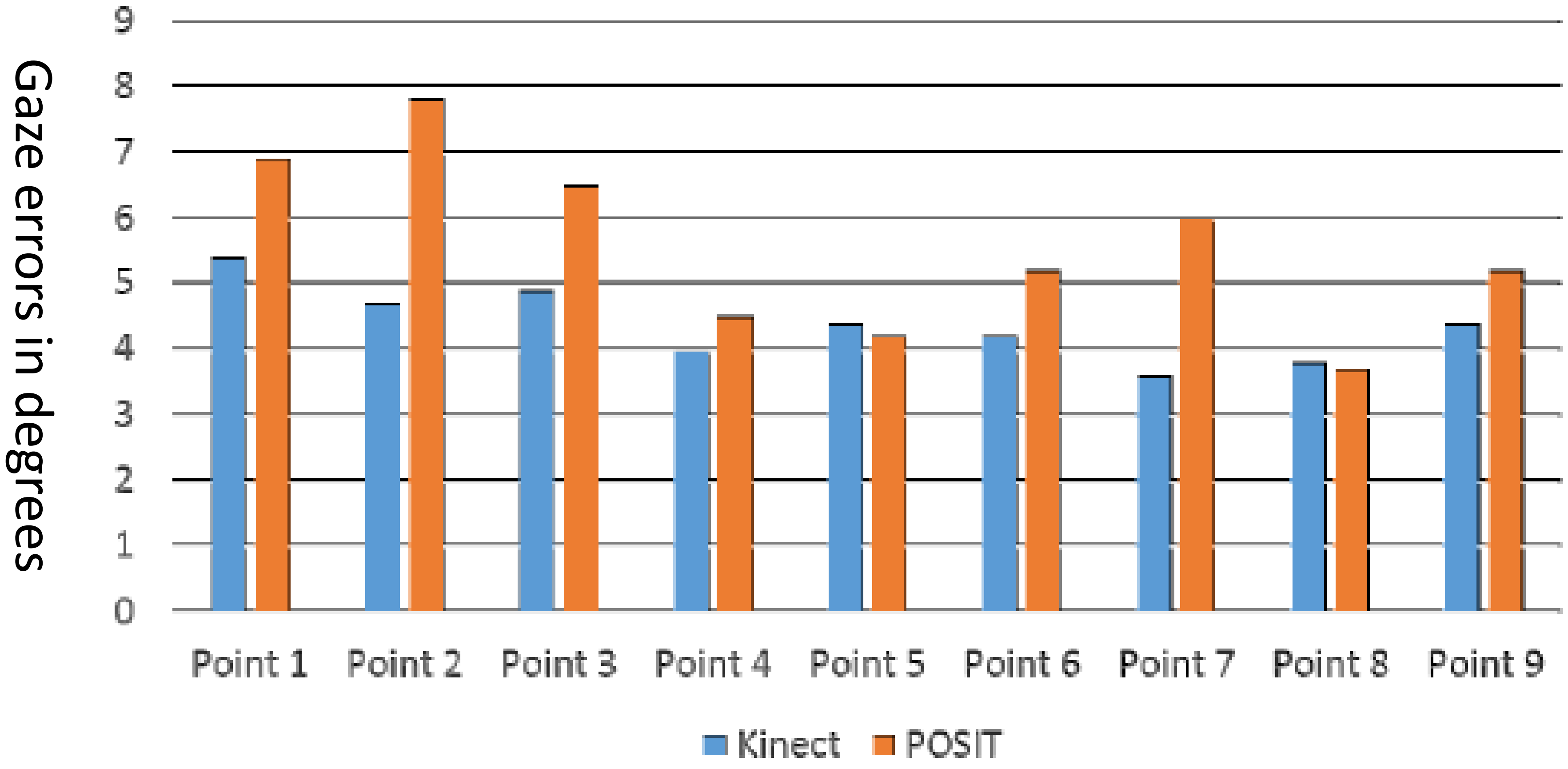
Left eye



Right eye

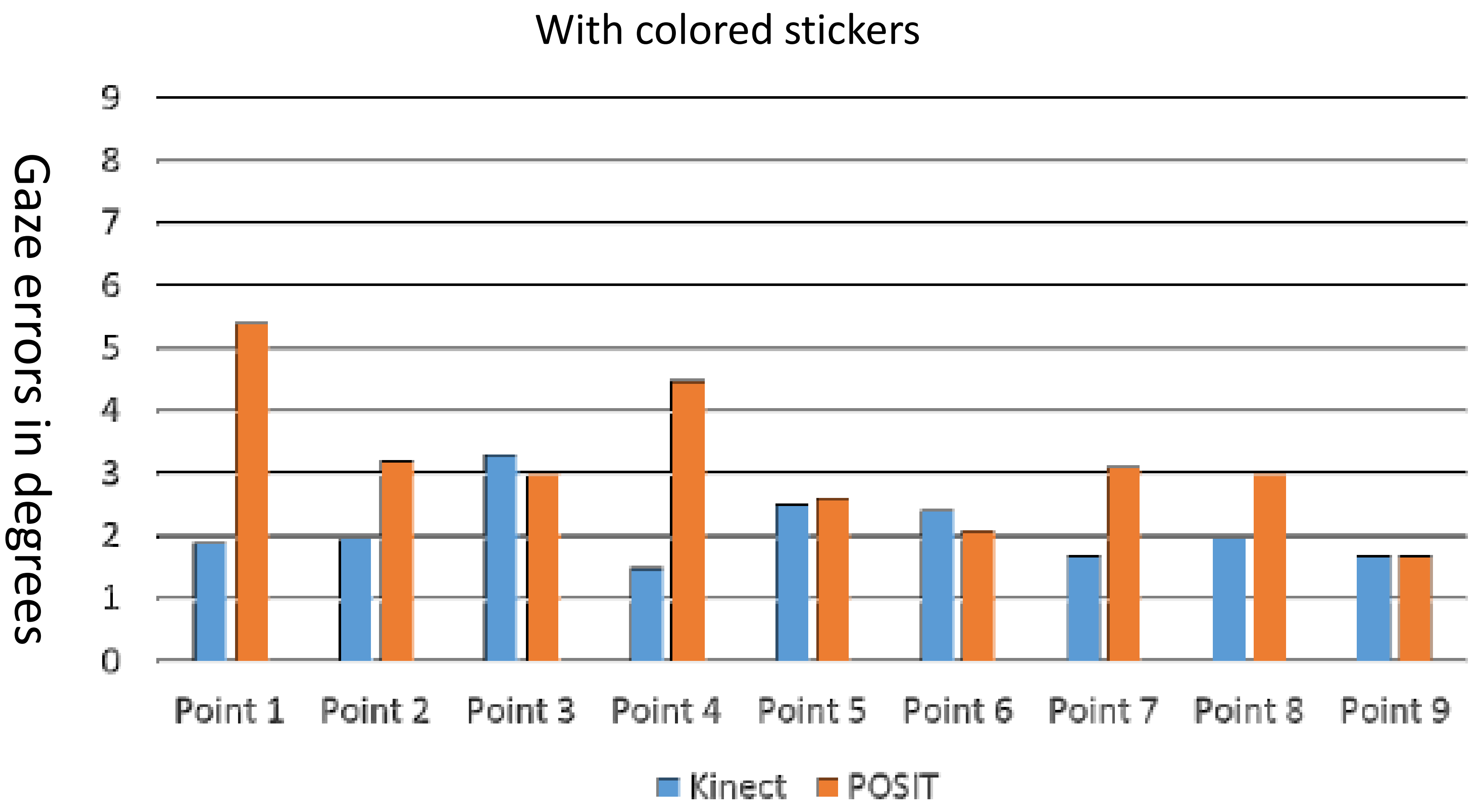


Gaze errors on real-world data



Average errors: 4.6 degrees with RGBD, and 5.6 degrees with RGB

Low-bound of gaze errors



Average errors: 2.1 degrees with RGBD, and 3.2 degrees with RGB

Conclusions

- Using depth information directly from Kinect provides more accurate gaze estimation compared with the one from only RGB images.

- The lower bound for gaze error is around 2 degrees with RGBD and 4 degrees with RGB

- Future work
  - Better RGBD sensor -> lower gaze error
  - Leverage two eyes

Zhengyou Zhang, Qin Cai, Improving Cross-Ratio-Based Eye Tracking Techniques by Leveraging the Binocular Fixation Constraint, in ETRA 2014.

Thank You