Privacy-Preserving Localization and Recognition of Human Activities

Janusz Konrad

BOSTON UNIVERSITY

Boston University Electrical & Computer Engineering

IEEE Signal Processing Society

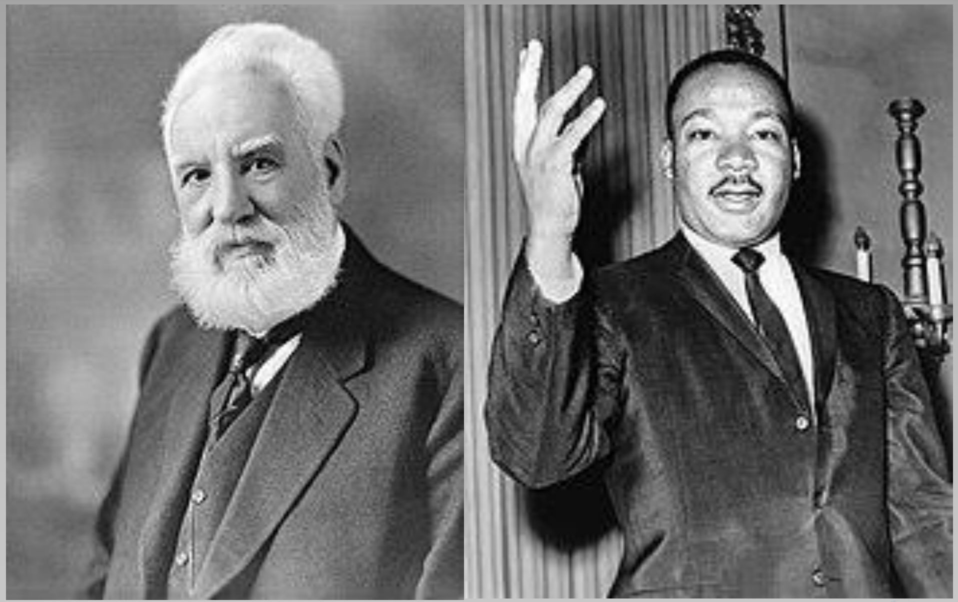
Outline

- Boston University and ECE
- Motivation
- Passive localization
- Passive activity recognition
- Localization via modulated light
- Final thoughts

Boston University

- Private, non-profit university:
  - Urban setting in major academic/metro area
  - 17,000 undergraduate students
  - 15,000 graduate students
  - 4,000 faculty
  - 18 schools and colleges
  - Major research university
- Department of Electrical and Computer Engineering:
  - 54 faculty
  - 465 undergraduate students
  - 221 Master's students
  - 144 PhD students

View of downtown across the Charles River



Main ECE research areas

Information and Data Sciences Group (ids.bu.edu):

- 20 faculty
- 50 PhD students

Acknowledgments

Prof. Prakash Ishwar

Dr. Jiawei Chen
(Oppo)

Jinyuan Zhao
(Google)

Dr. Behrouz Saghafi
(U. North Carolina)

Ji Dai (MS)

Doug Roeper (undergrad)

Natalia Frumkin (undergrad)

Research support: National Science Foundation and Boston University.
Lecture support: IEEE Signal Processing Society Distinguished Lecturer Program.

Paradigm shift

2019

2030?

Paradigm shift

2019

2030?

?

Energy only ?

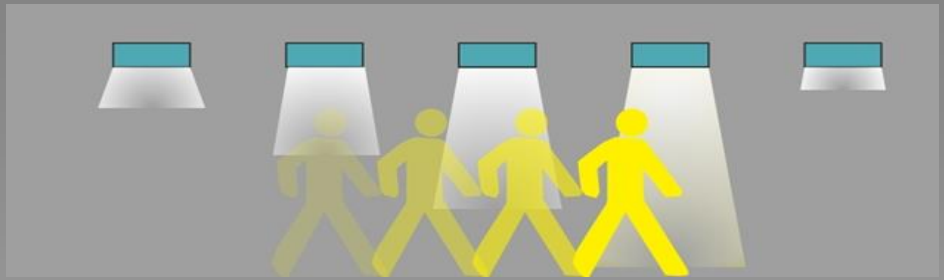
Can we better exploit this opportunity?

Lighting-Enabled Systems & Applications

- NSF Engineering Research Center
- $37M from over 10 years
- 3 universities
- 24 industrial members

Leveraging LEDs for …

Energy Savings



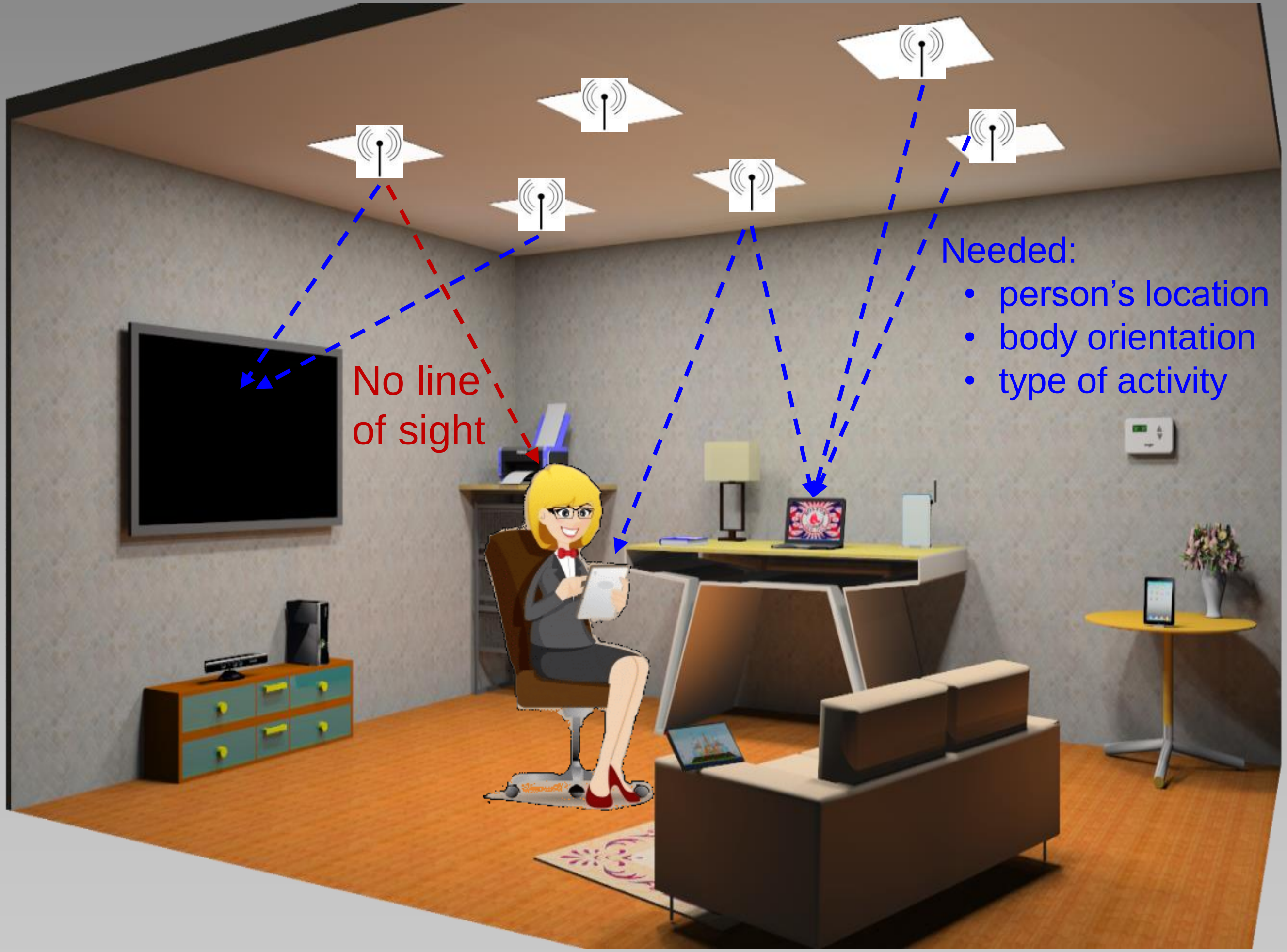
Health Benefits

Productivity Gains

Bit-rate up to 1Gb/s for a focused, directional beam

Much lower rates for diffused light

Context: Visible Light Communication



Other applications:

- Lighting control
- HVAC control
- Robotics
  (people avoidance)

Clear goal, but one caveat …

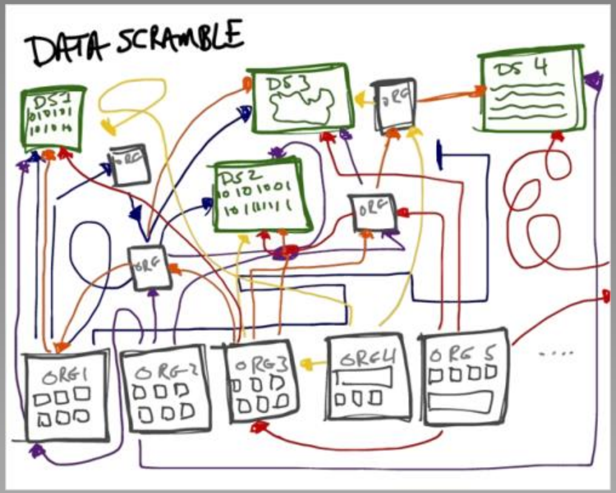
- Activity localization and recognition studied extensively over decades
- Excellent performance, even under challenging conditions, but …

requires cameras

no privacy

Approach I: Reversible methods

- **Data scrambling:**
  - Typical approach: data permutation
  - Domain: image, transform, bitstream
  - Vulnerable to attacks [*Macq and Quicquater, Proc. IEEE, 1995*]



- **Data encryption:**
  - Naïve methods: video bitstream = text data
  - Cryptographic algorithms: DES, AES, RSA
  - Extracting original information from encrypted data is challenging
  - Attacks are difficult but recently **deep learning** was successful in recognizing encrypted images [*Wang et al., MSSP 2017, Bachrach et al., ICML 2016*]

Approach II: Irreversible methods

Data degradation:

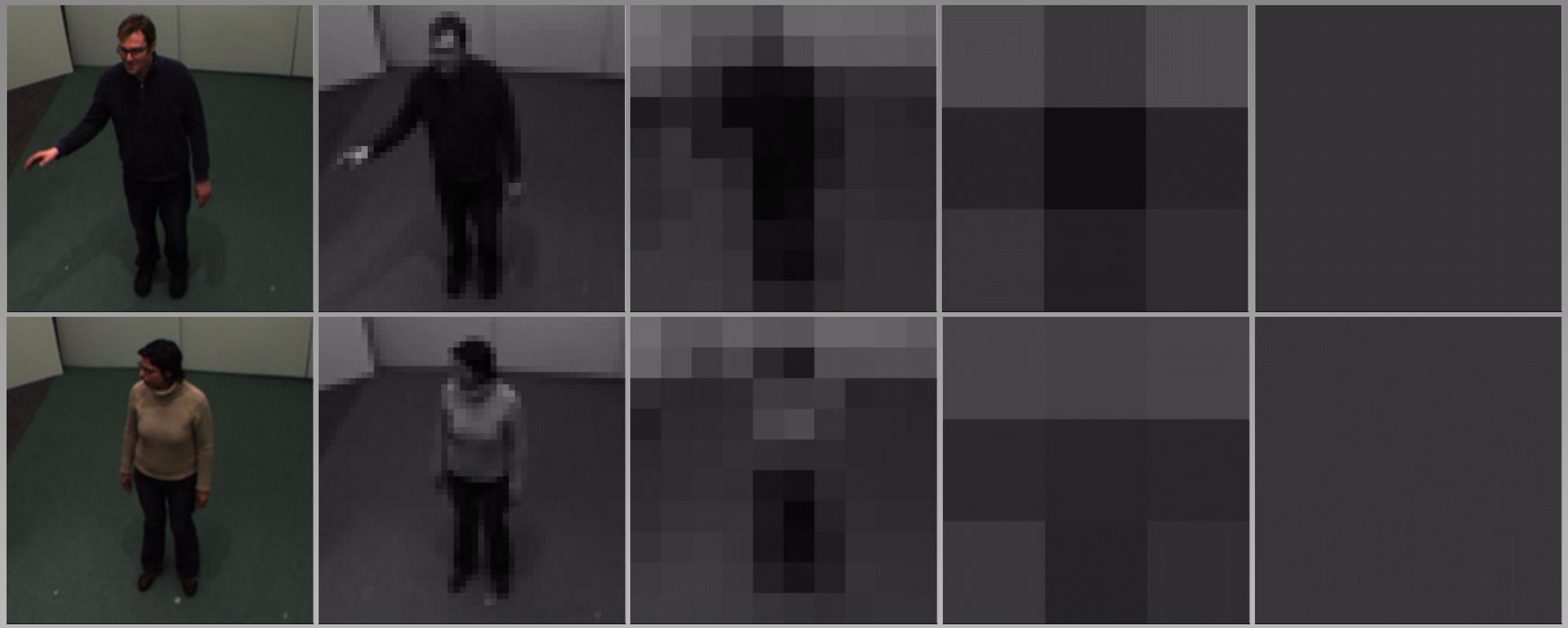
- Before acquisition (optically):

[*Pittaluga et al., CVPR, 2015*]
Coarse control due to optics

- After acquisition (digitally):

[*Winkler et al., AVSS, 2014*]
Potential for eavesdropping

Alternative approach: ultra-low res

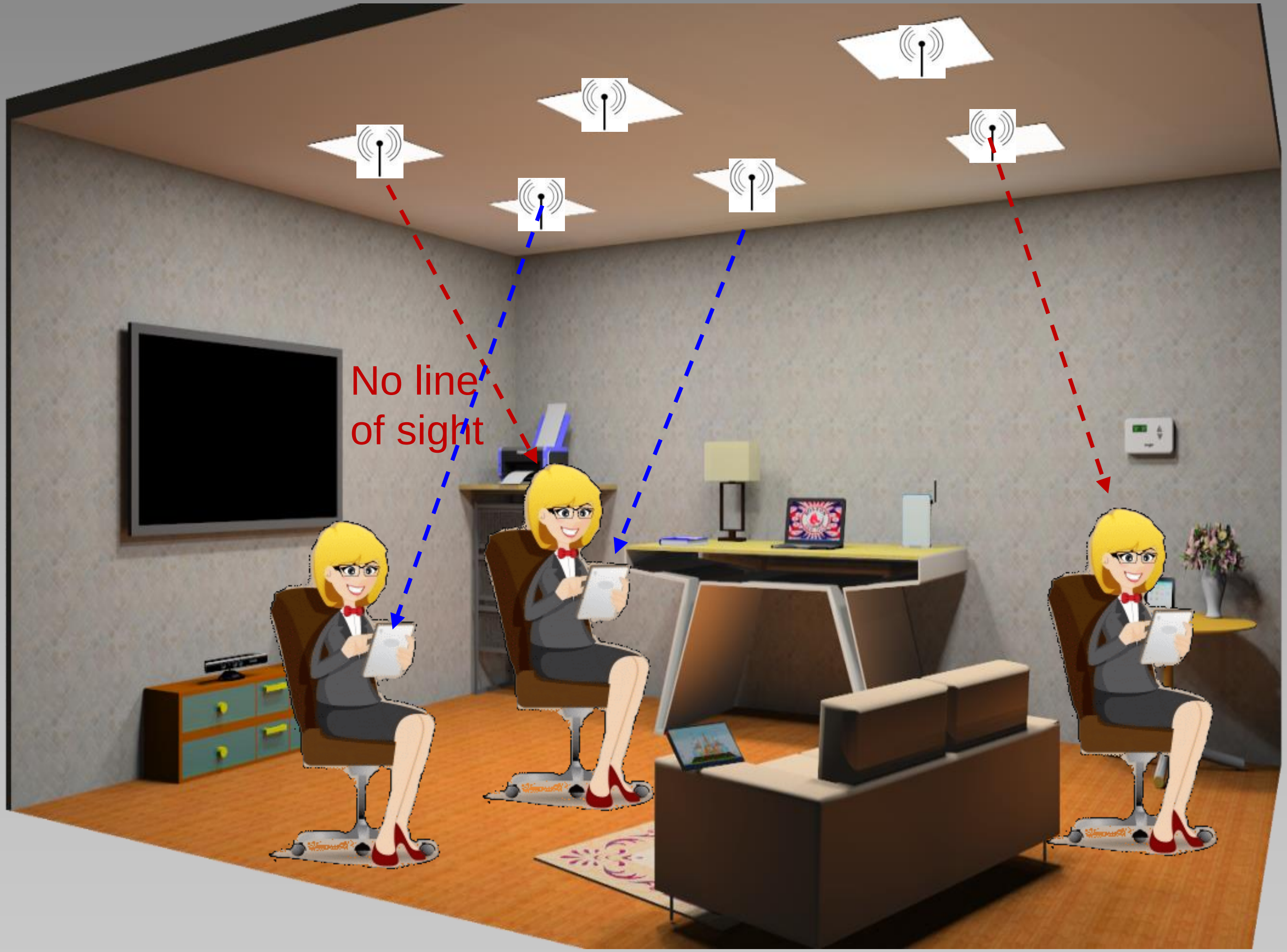


Challenge: Localize and recognize activities at **extremely-low resolution (eLR)**

Benefits: - eavesdropping will not threaten privacy

- low data transmission and computing costs

Task I: Person localization



Not difficult with video camera(s)

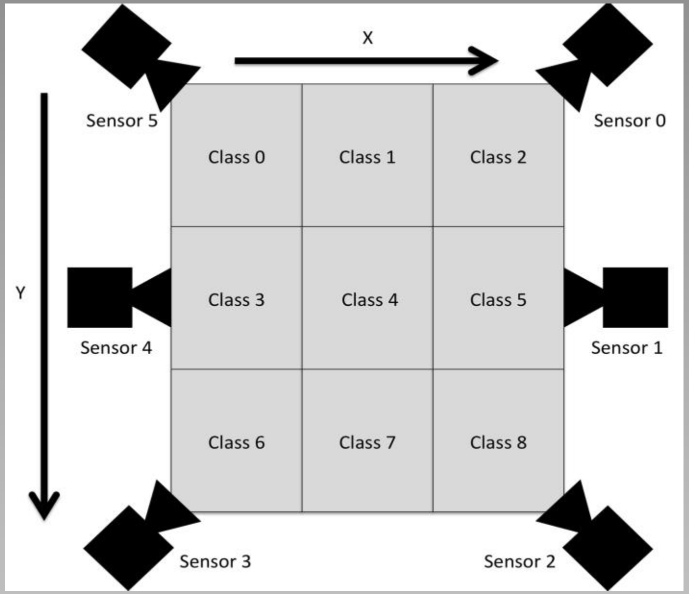
How about ``single-pixel'' cameras ?

1 sensor reading per frame

Testbed

- 6 single-pixel visible-light sensors (Taos-AMS TCS 34725)

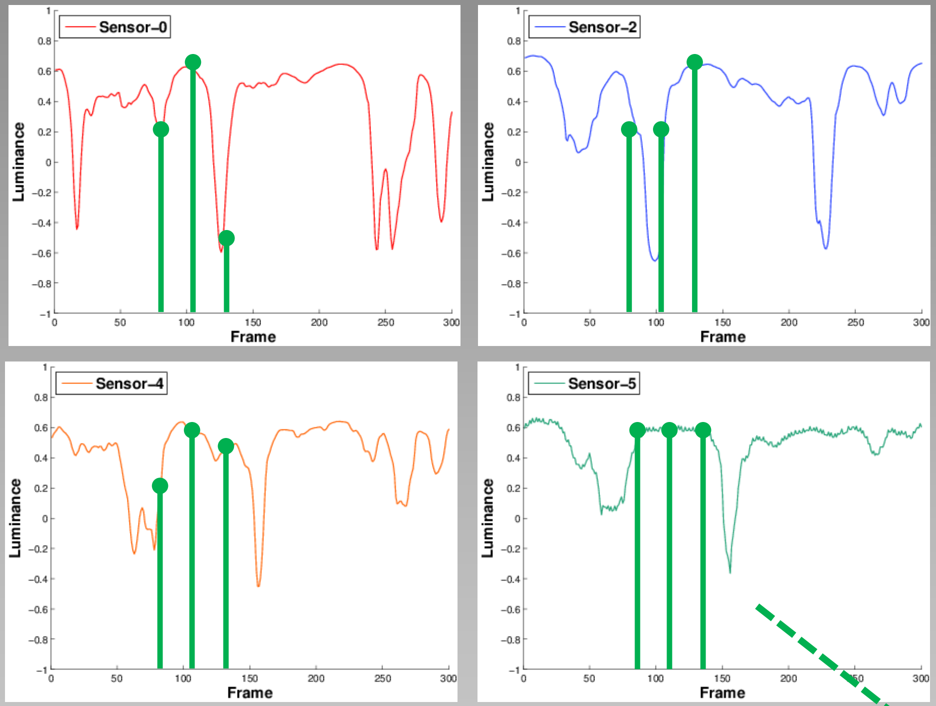
- Area: 2.37m x 2.72m



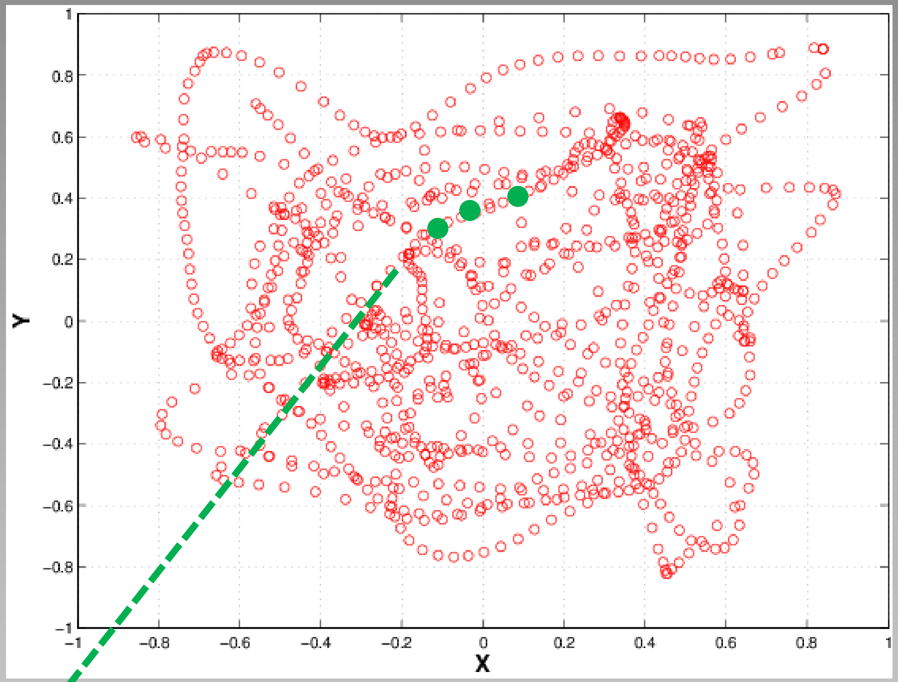
- Data-driven approach → Ground truth needed:
  Hollywood-style motion capture (IR light + markers)

Data-driven localization: First train

Simultaneous recording:



Single-pixel sensor light readings

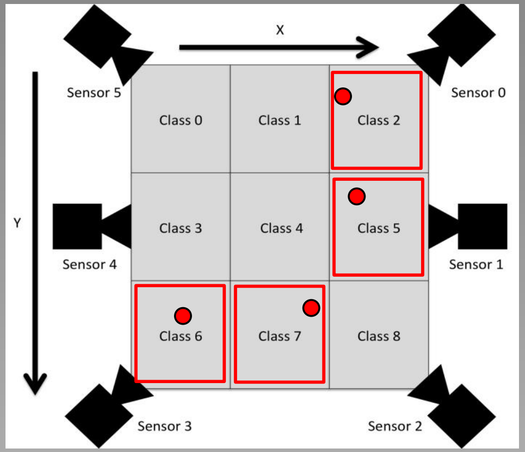


Ground-truth locations readings

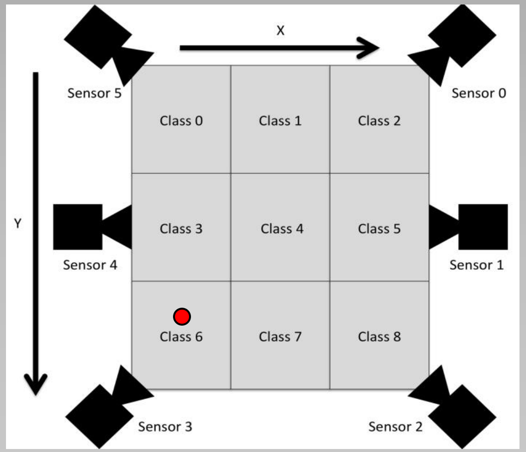
Machine Learning

Then, test

- Coarse-grained localization (classification):
  - Area divided into cells (3 x 3 grid)
  - Each cell is a class
  - SVM classifier (RBF kernel)



- Fine-grained localization (estimation):
  - Real-valued location coordinates estimated
  - Separate estimation for X and Y coordinates
  - SVM regressor (RBF kernel)



Usage scenarios and validation

- Dataset
  - Several random 90-second walks by 4 different people
  - Ground-truth locations: OptiTrack system
- Public setting (e.g., conference room)
  - New users appear often
  - System cannot be trained on all users
  - *Leave-one-person-out* cross-validation
- Private setting (e.g., home)
  - Same set of users
  - System can be trained on all users
  - *Leave-one-walk-out* cross-validation

[*Roeper et al., IEEE-AVSS, 2016*]

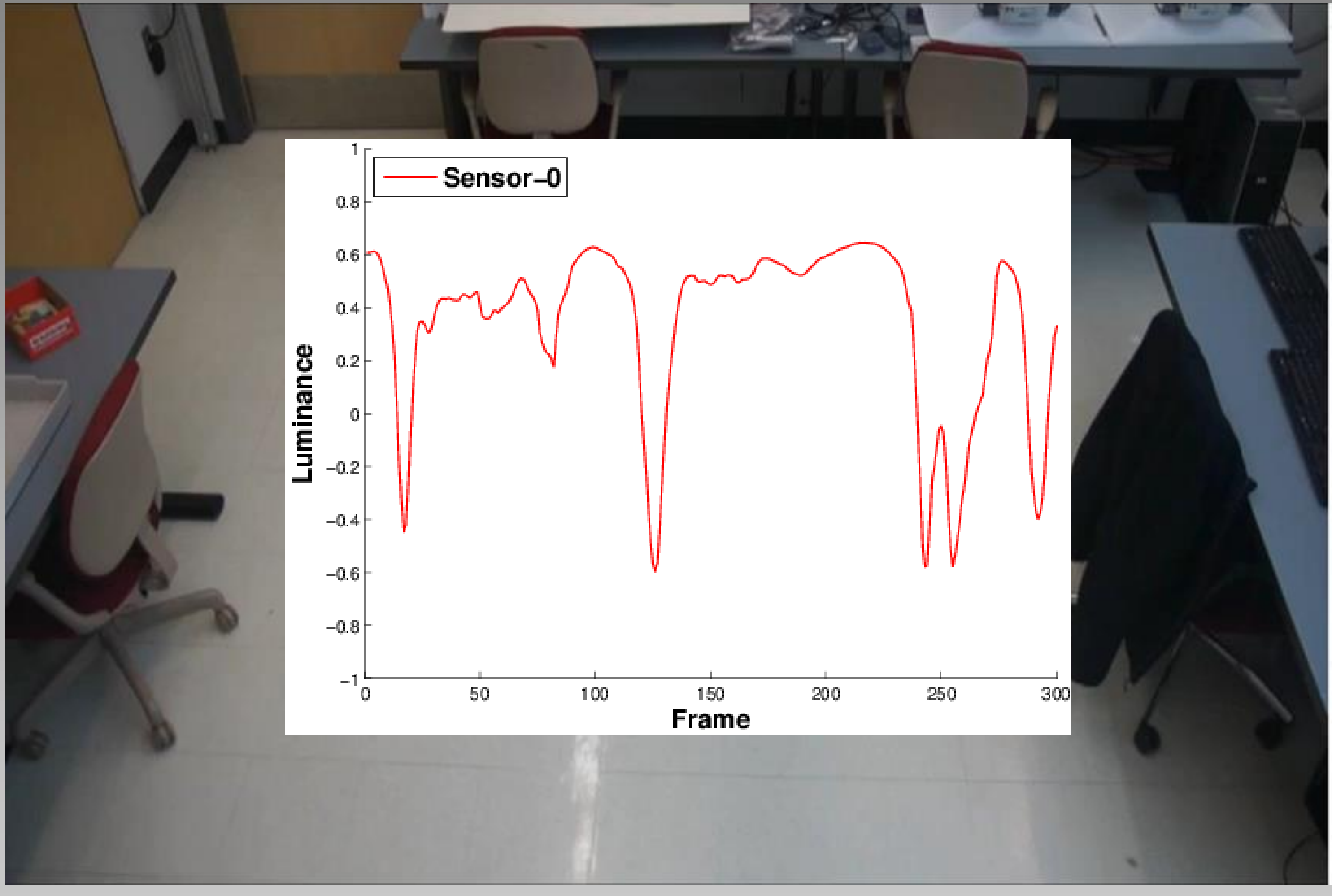
Coarse [CCR]	Fine [localization error]

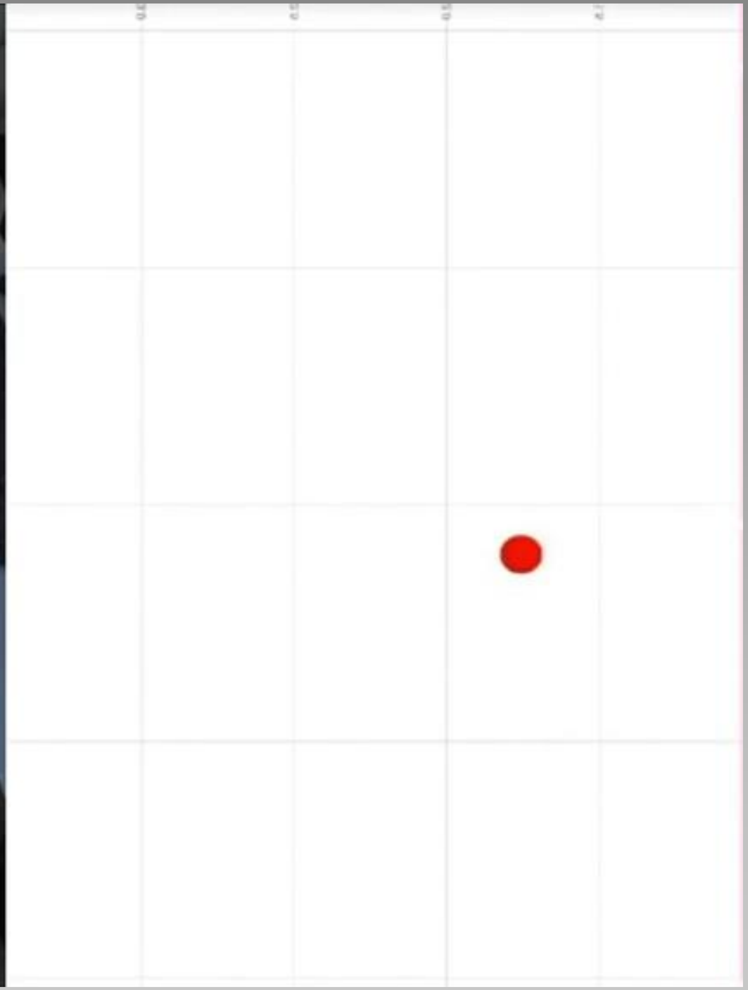
Public setting	
67%	**35cm** (±25cm)

Private setting	
72%	**31cm** (±22cm)

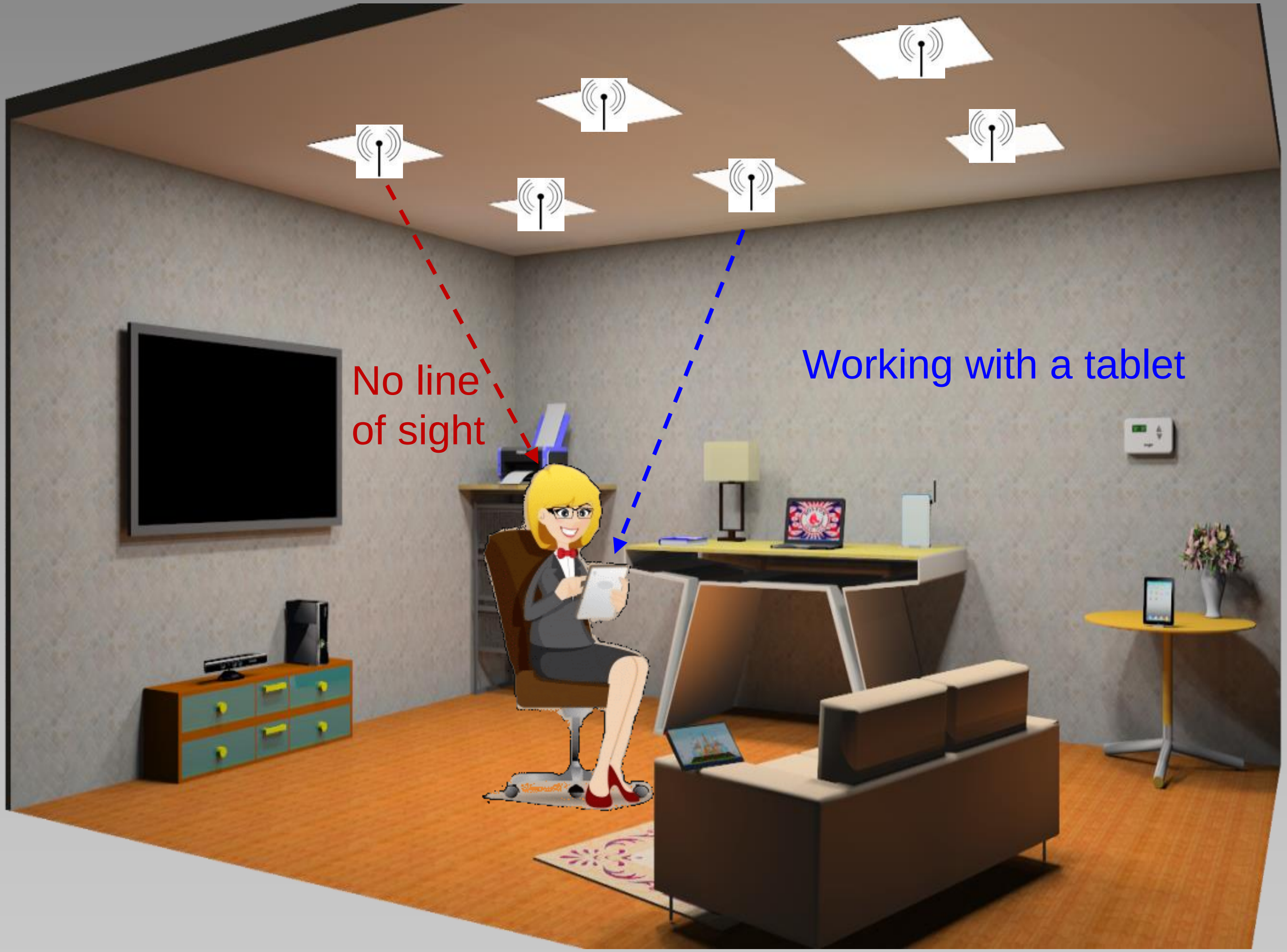
Results: tracking

Red – estimate from 6 single-pixel sensors





Task II: Activity Recognition



State-of-the-art

- **30-pixel** humans: optical flow, NN classifier, optical-flow correlation as distance metric [*Effros et al., 2003*]
  Optical flow unreliable at lower resolutions

- **32 x 48 images**: Hu moments from directional history images, kNN classifier [*Ahad et al. 2010*]
  Poor performance at lower resolutions

- 20 ceiling-mounted **binary IR sensors**: short-duration averages of binary values, SVM classifier [*Tao et al. 2012*]
  Unrealistic scenario leveraging strong correlation between action and location

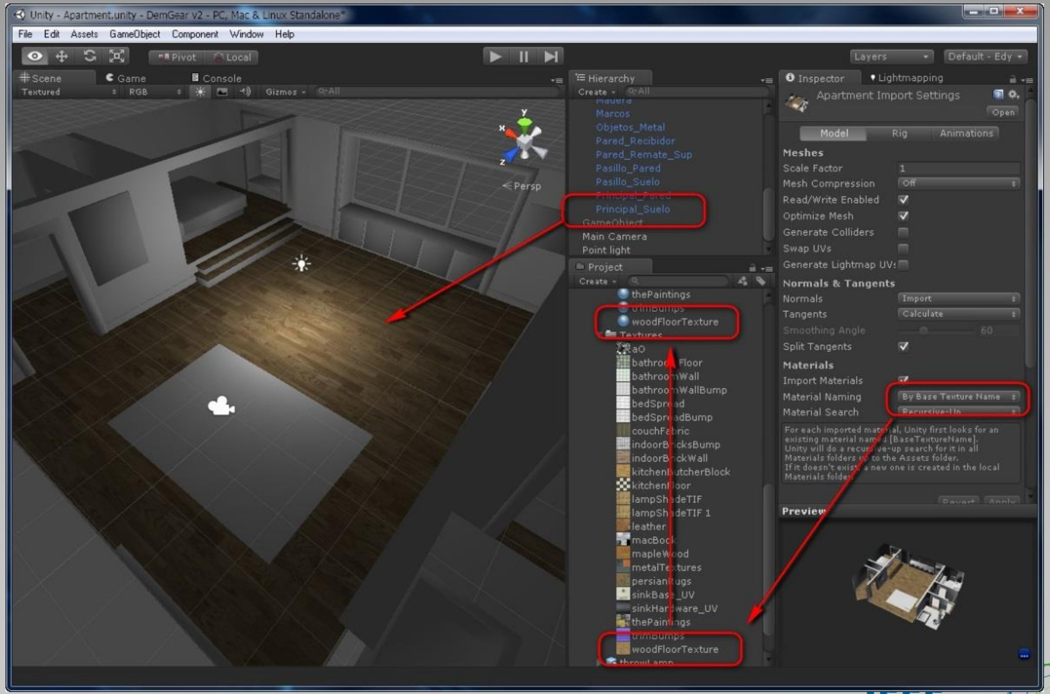
Can we go even lower in resolution but maintain recognition performance?

Phase I: Simulation

Virtual testbed:

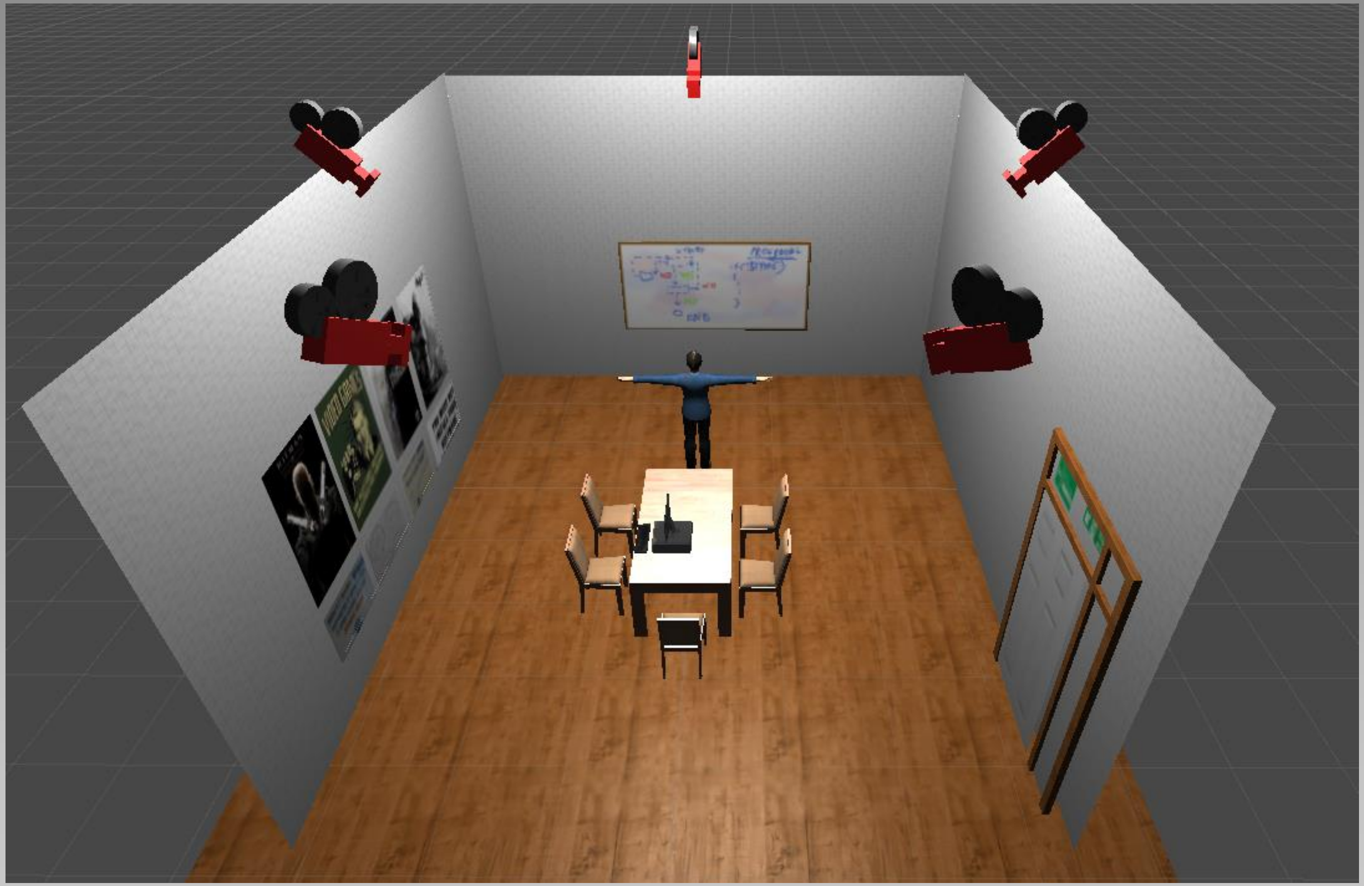
- Kinect v2: motion capture

- Unity3D©:
  - 3-D scene,
  - avatars,
  - animation by human motion



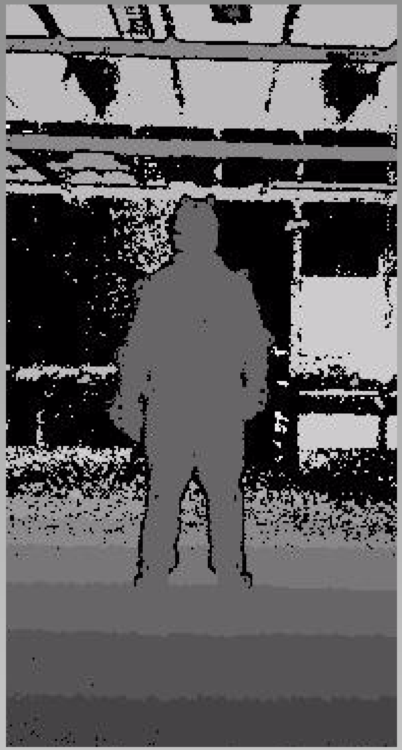
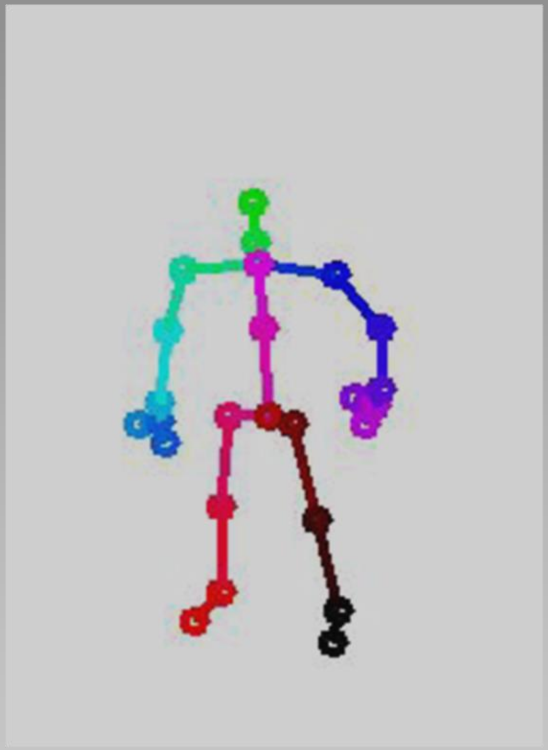
Step 1

Simulate a virtual 3-D scene and sensors



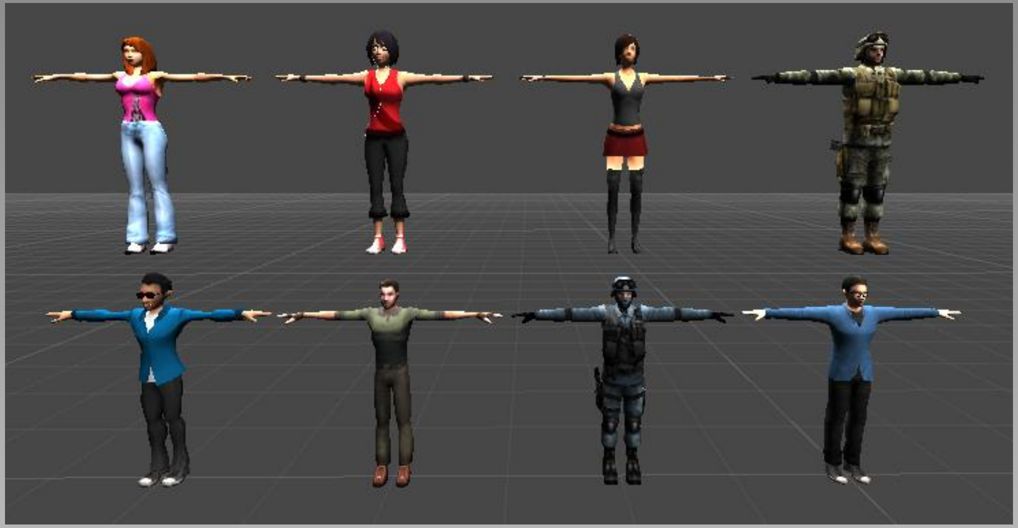
Step 2

Record humans with Kinect and extract skeletons

Step 3

Animate avatars using recorded skeletons



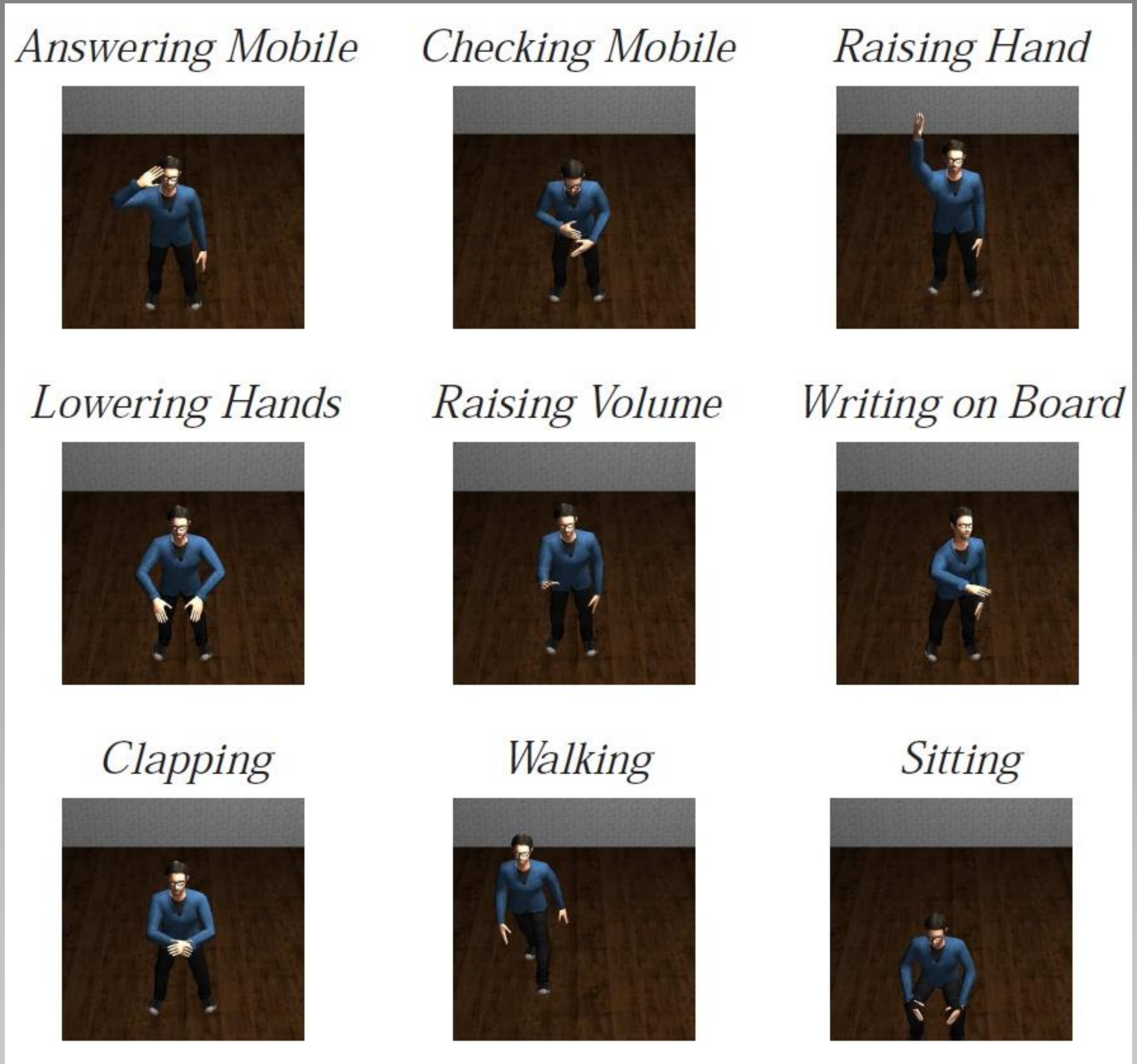
Step 4

Capture data from the virtual scene:

- at various resolutions,
- at various locations from various angles (field of view),
- of different type (luminance, color, depth),
- …

Collected actions

- 12 subjects:
  - 7 male,
  - 5 female
- 9 actions typical of a seminar-room scenario
- Single avatar in FOV



Features

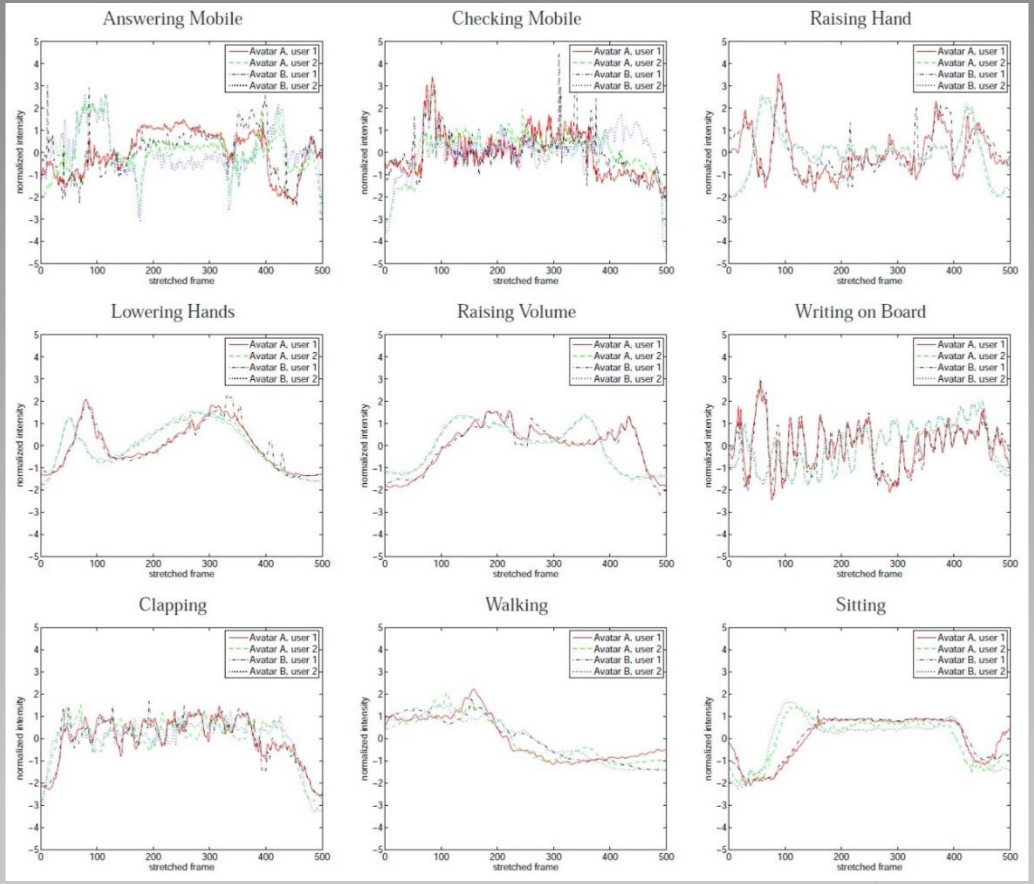
- Elaborate features cannot be extracted (too few pixels)

- Feature = grayscale value at each pixel:

$$I_{i,j,k}[t]$$

(i,j)	spatial location
k	camera number
t	time instant

- Mean-variance equalization to focus on dynamics and reduce impact of clothing



Classification

- **Given:** query sample $\hat{I}$ and dictionary samples:

$$\{\hat{V}^{m,l}\} \quad l = 1, \dots, L, \qquad m = 1, \dots, M_l$$

# of samples per class

# of classes

- **Find nearest neighbor:**

$$(\hat{m}, \hat{l}) = \arg\min_{m,l} d(\hat{I}, \hat{V}^{m,l})$$

Winning class

under l_1 distance metric:

$$d(\hat{I}, \hat{V}) = \sum_k \sum_t \sum_i \sum_j |\hat{I}_{i,j,k}[t] - \hat{V}_{i,j,k}[t]|$$

Results

Description	Configuration	CCR
Best	10 x 10, 30 Hz, 5 cams	89.60%
Low frame rate	10 x 10, 2 Hz, 5 cams	86.49%
Single camera	10 x 10, 30 Hz, 1 cam	77.96%
Low spatial resolution	1 x 1, 30 Hz, 5 cams	75.50%
Everything low	1 x 1, 2 Hz, 1 cam	48.39%

[*Dai et al., IEEE-ICIP, 2015*]

- CCRs around 90% possible at privacy-preserving resolutions
- No need for high frame rate (for these actions)
- More sensors needed at extremely low spatial resolutions

Phase II: Real-camera data

- So far, proof of concept validated on synthetic data:
  - no noise,
  - no illumination variations,
  - known subject location.
- Test on a real-camera dataset?
- IXMAS-ROI actions dataset (64 x 48 pixels, 25 Hz):
  - 12 actions, 10 subjects, 5 cameras.

IXMAS-ROI results

- Various decimation factors

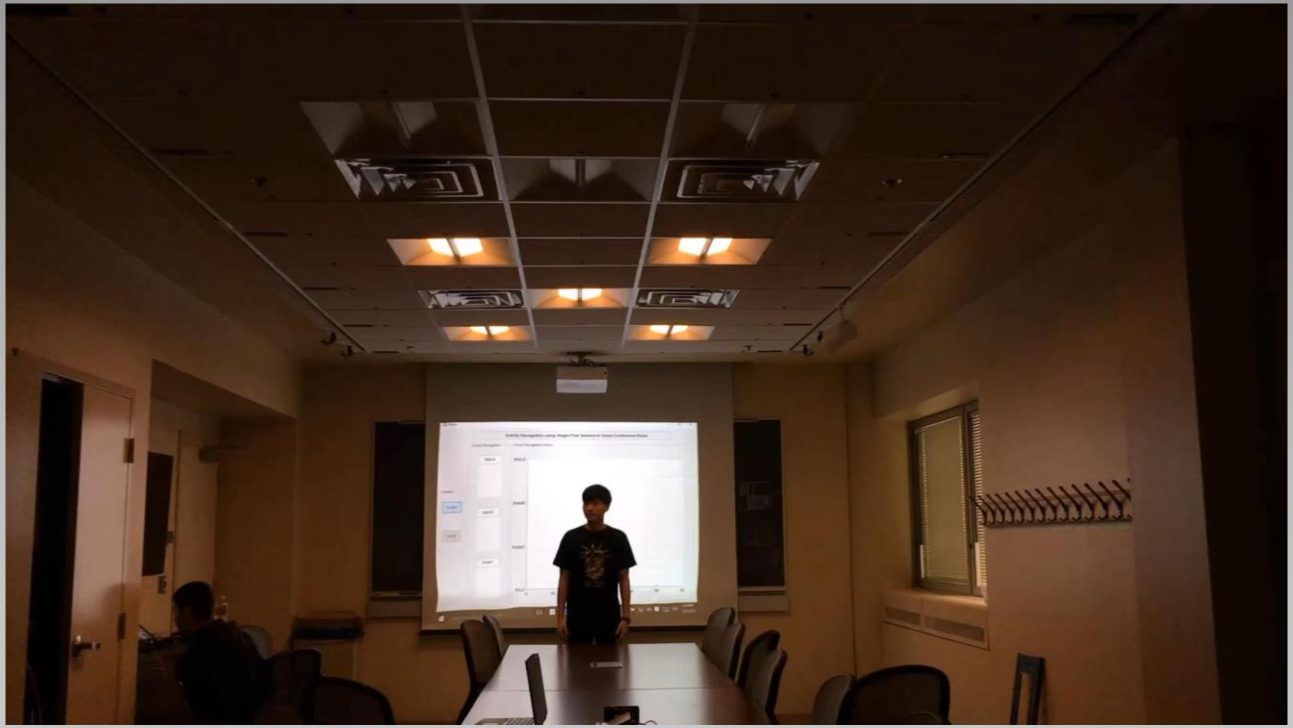
Description	Configuration	CCR
Best	16 x 12, 25 Hz, 5 cams	80.00%
-	8 x 6, 25 Hz, 5 cams	77.78%
-	4 x 3, 25 Hz, 5 cams	76.94%
Low frame rate	16 x 12, 2 Hz, 5 cams	74.35%
Single camera	16 x 12, 25 Hz, 1 cam	67.11%
Low spatial resolution	1 x 1, 25 Hz, 5 cams	63.33%
Everything low	1 x 1, 2 Hz, 1 cam	29.21%

[*Dai et al., CVPR-AMFG, 2015*]

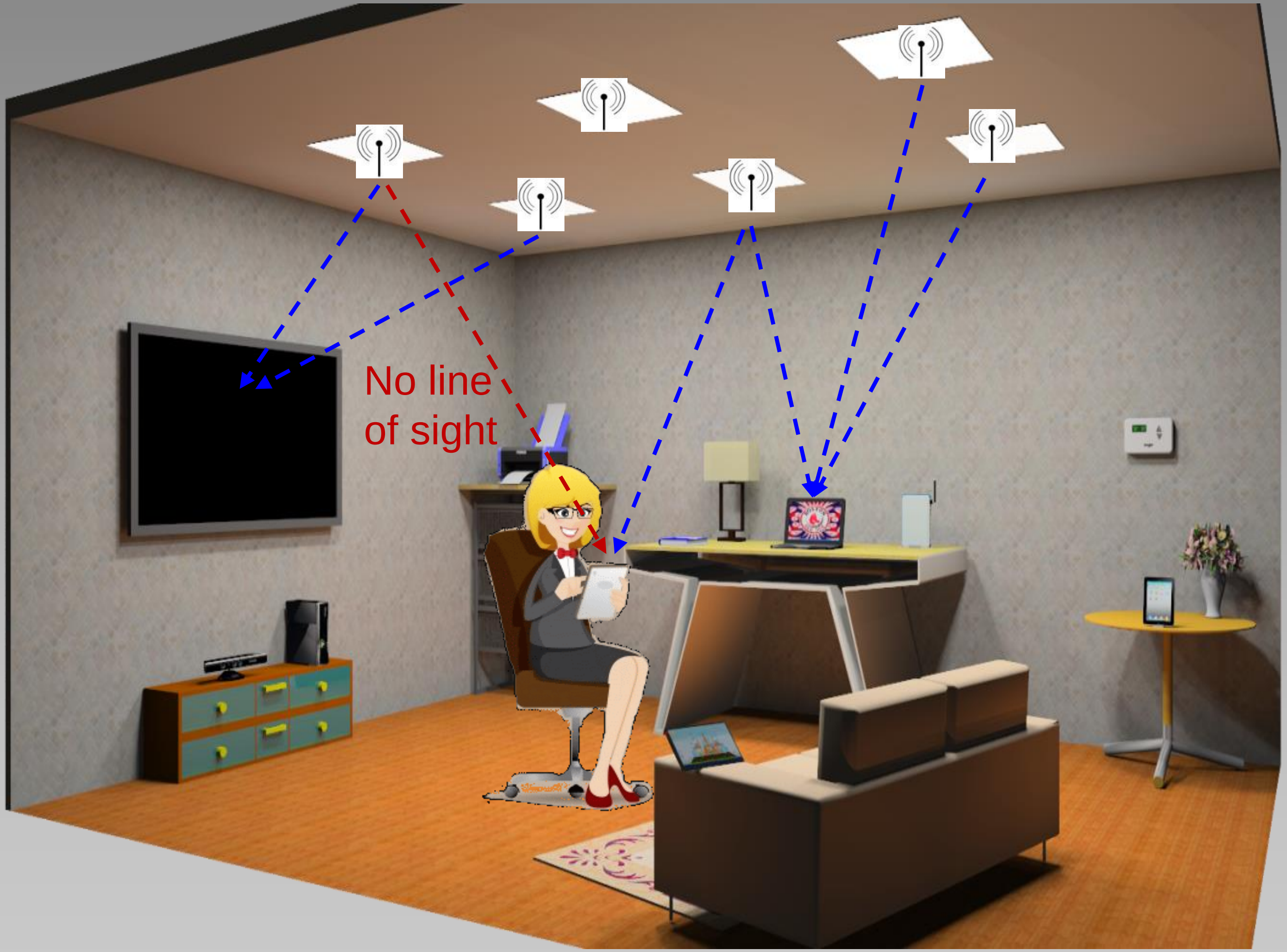
- 7-10% CCR drop compared to avatar data, but …
  the same trends are observed

Phase III: Physical testbed

- 12 single-pixel sensors, 10 fps
- POE data transmission and power
- Real-time algorithm in *Matlab* on a laptop

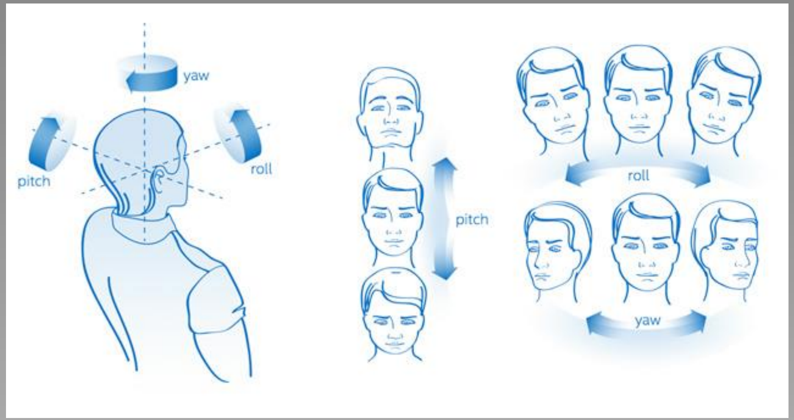


Task III: Body orientation estimation



Where is this tablet?

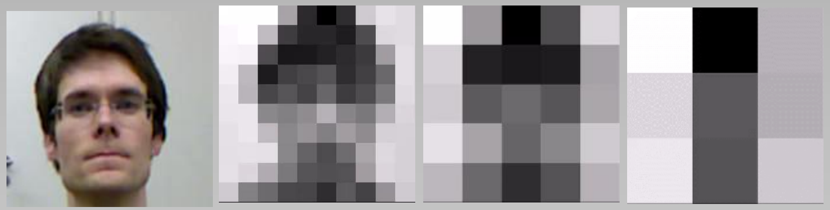
Study case: Head pose estimation



Traditional methods

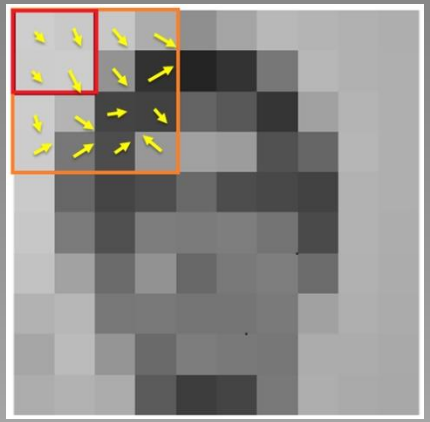
Our approach

Full resolution



Ultra-low resolution

Features



- Histogram of Gradients (HOG):

  $\boldsymbol{f}$ = concatenated histograms

Resolution	Cell Size	Block Size	Length of f
10 × 10	2 × 2 (pixel)	2 × 2 (cell)	576
5 × 5	1 × 1 (pixel)	2 × 2 (cell)	576
3 × 3	1 × 1 (pixel)	2 × 2 (cell)	144

- New gradient-based pixel-wise feature: $\boldsymbol{f} = [\boldsymbol{g}_{1,1}, \boldsymbol{g}_{2,1}, \boldsymbol{g}_{3,1}, \dots]$

$$\boldsymbol{g}_{i,j} = \left(\frac{\partial \hat{I}_{i,j}}{\partial x}, \frac{\partial \hat{I}_{i,j}}{\partial y}, \|\nabla \hat{I}_{i,j}\|, \arg(\nabla \hat{I}_{i,j})\right)$$

Resolution	Length of f
10 × 10	400
5 × 5	100
3 × 3	36

Estimation *via* non-linear regression

- **Support Vector Regression**: given a training set $\{(\boldsymbol{f}_j, \theta_j), j = 1, \dots, N\}$, learn functional mapping:

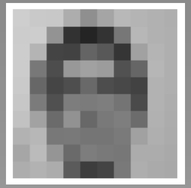
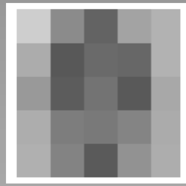
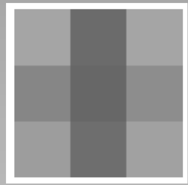
$$\hat{\theta}(\boldsymbol{f}) = \sum_{j=1}^{N} w_j K(\boldsymbol{f}_j, \boldsymbol{f}) + b$$

by minimizing a regularized ϵ–insensitive loss function:

$$\min_{b,\boldsymbol{w}} \frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum_{j=1}^{N} \max(0, |\theta_j - \hat{\theta}(\boldsymbol{f}_j)| - \epsilon)$$

- **One regressor for each pose angle:** pitch, yaw, roll

Results: Mean-Absolute Error on 15k images

Method/Resolution	Pitch Error°	Yaw Error°	Roll Error°
SVR:HoG$_{rgb}$ 10 × 10	12.9±17.2	9.9±12.4	6.9±9.8
SVR:Grad$_{rgb}$ 10 × 10	14.1±18.6	12.4±15.6	7.2±10.3
SVR:HoG$_{rgb}$ 5 × 5	16.1±20.1	15.2±19.3	7.6±10.9
SVR:Grad$_{rgb}$ 5 × 5	13.7±17.6	11.2±14.4	7.7±10.9
SVR:HoG$_{rgb}$ 3 ×3	18.7±23.1	22.8±28.9	7.6±11.6
SVR:Grad$_{rgb}$ 3 ×3	15.9±20.2	16.3±20.8	8.0±11.5
Median -	18.8±15.9	23.9±18.9	7.4±8.9
[12] HoG$_d$ + SIFT$_{rgb}$ Full resolution	8.5±11.1	8.8±14.3	7.4±10.8
[3] HoG$_{rgb}$ Full resolution	5.7±6.1	4.9±5.1	4.8±5.9
[3] HoG$_{rgb}$ + HoG$_d$ Full Resolution	5.0±5.8	3.9±4.2	4.3±4.6

≈10° at 10x10 pixels

≈11° at 5x5 pixels

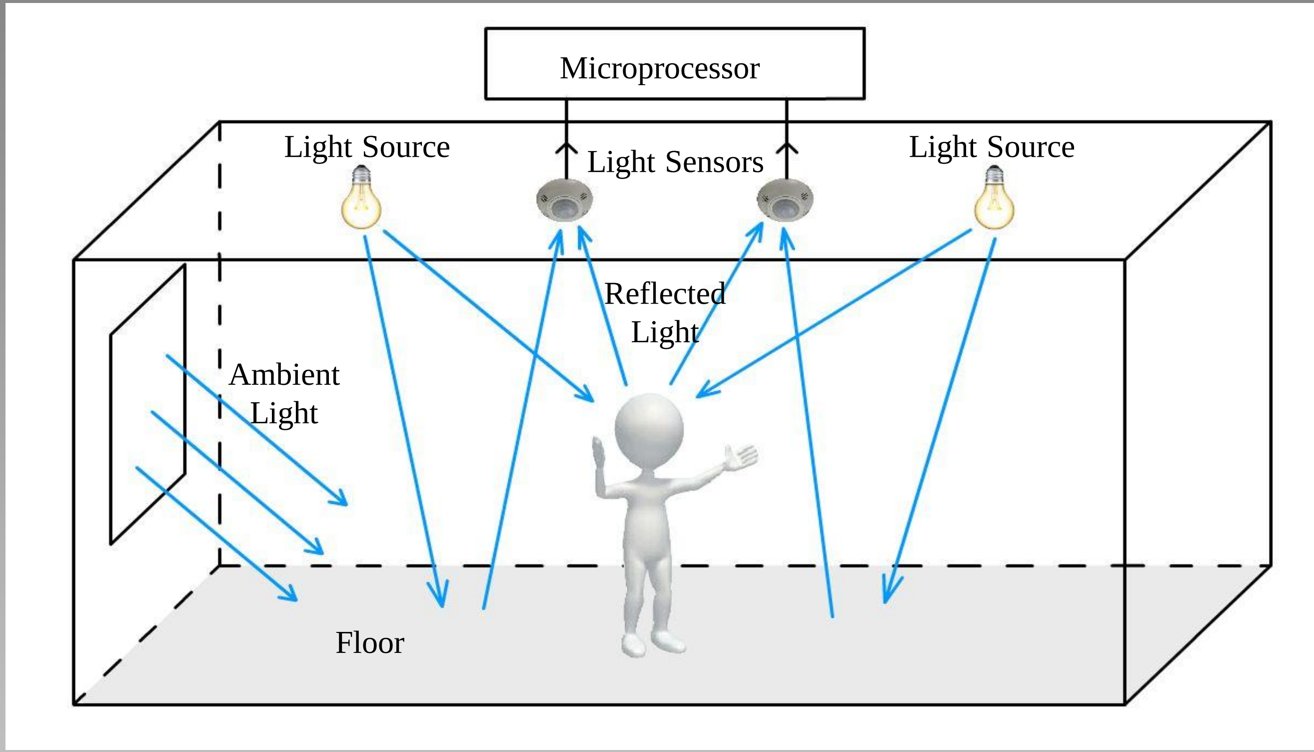
Grad better than HOG

≈13° at 3x3 pixels

≈5° at 640x480 pixels

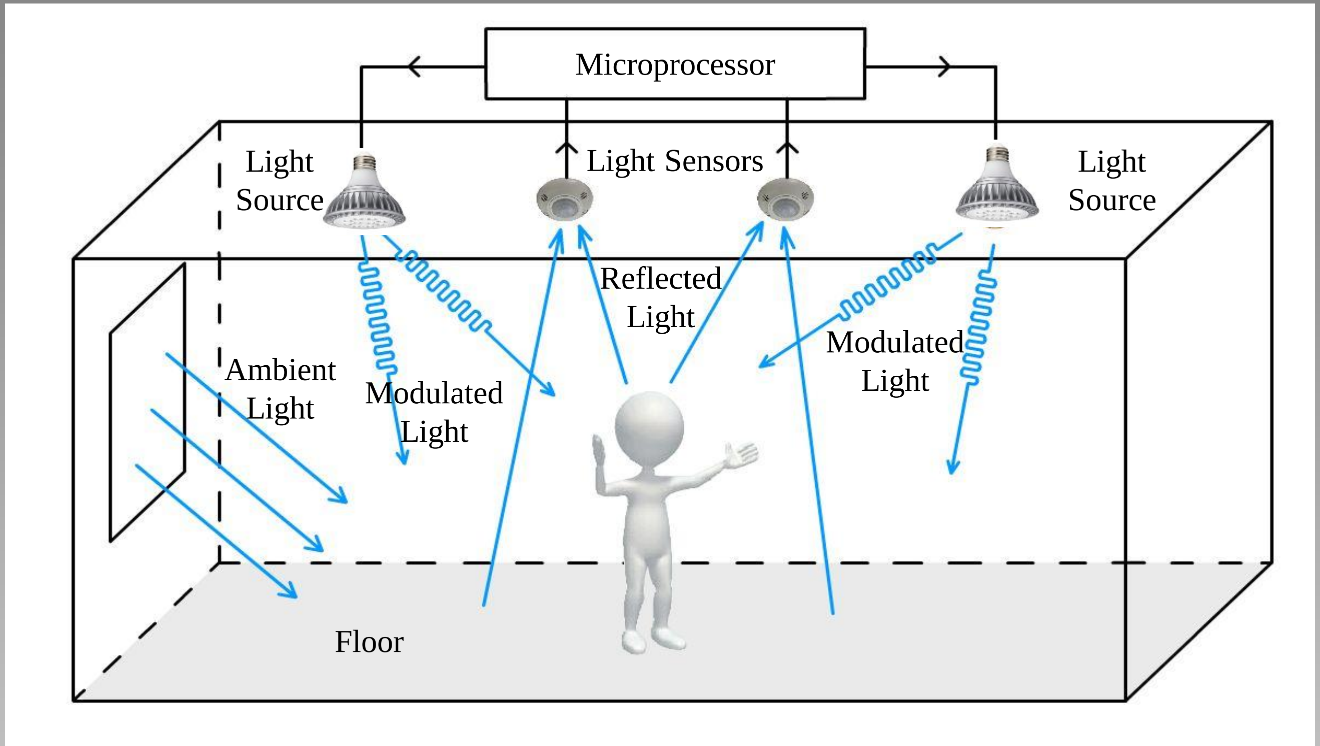
[Chen et al., IEEE-SSIAI, 2016]

Localization thus far: Passive light sensing



- Light sources not controllable (incandescent, fluorescent)
- Algorithms rely on reflected light measurements: high sensitivity to changes in illumination

Alternative localization: Active light sensing



- Precisely-modulated LED light sources (frequency > 60 Hz)
- Algorithms use both reflected light **and** modulation pattern
- Robust to illumination changes

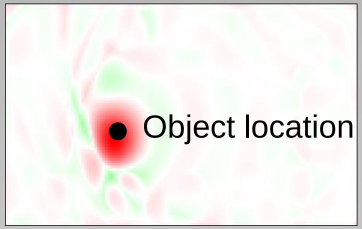
Algorithm overview

- Relationship between light modulation and sensor response is captured by **light transport matrix** A
- Object presence **changes** light transport matrix A
- Algorithm estimates **floor reflectivity change** from the change in light transport matrix
- Region of largest reflectivity change identifies **object location**

Object

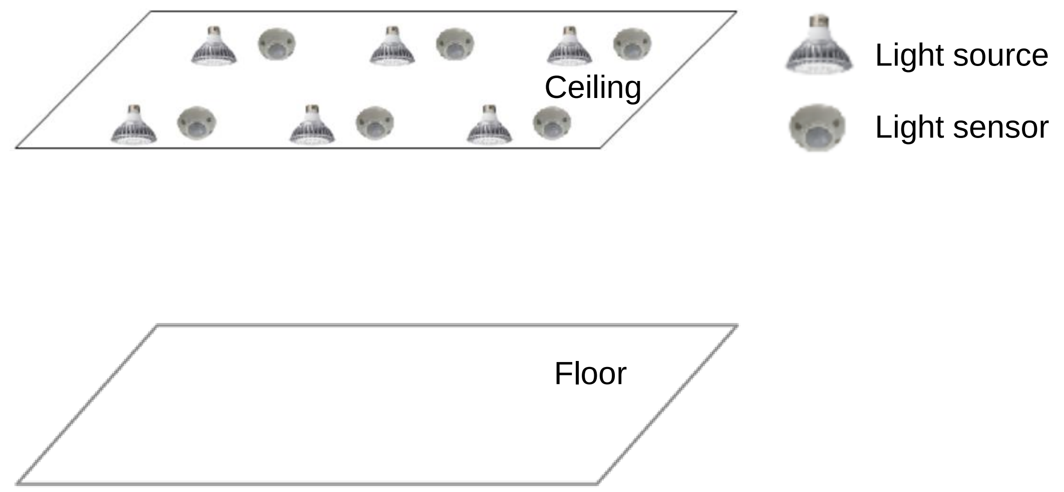
$$\begin{array}{cc} & \text{Sources} \\ \text{Sensors} & \begin{bmatrix} \Delta A(1,1) & \cdots & \cdots \\ \vdots & \Delta A(i,j) & \vdots \\ \cdots & \cdots & \Delta A(N_s,N_f) \end{bmatrix} \end{array}$$

Light transport matrix change ΔA

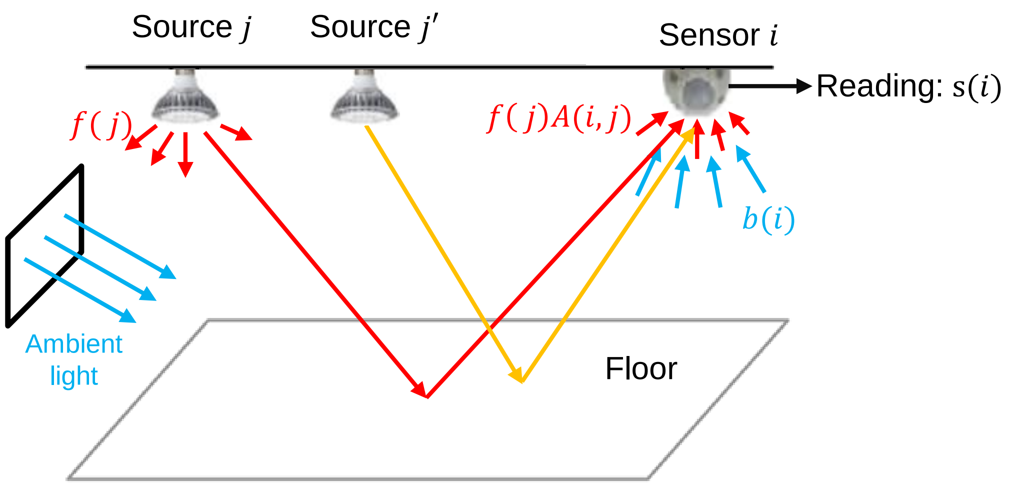


Floor reflectivity change

Light transport matrix A



Light transport matrix A



$$s(i) = b(i) + \sum_j f(j) \cdot A(i,j)$$

$s(i)$	$b(i)$	$f(j)$	$A(i,j)$
Sensor i reading	Ambient light	Source j intensity	Light contribution Source $j \rightarrow$ Sensor i

Estimating matrix A via light modulation

$$\text{Sensor } i \longleftarrow \begin{bmatrix} s(1) \\ \vdots \\ s(i) \\ \vdots \\ s(N_s) \end{bmatrix} = \begin{bmatrix} A(1,1) & \cdots & \cdots \\ \vdots & A(i,j) & \vdots \\ \cdots & \cdots & A(N_s,N_f) \end{bmatrix} \begin{bmatrix} f(1) \\ \vdots \\ f(j) \\ \vdots \\ f(N_f) \end{bmatrix} \longrightarrow \text{Source } j \; + \underbrace{\begin{bmatrix} b(1) \\ \vdots \\ b(N_s) \end{bmatrix}}_{\text{Ambient light}}$$

$$\boldsymbol{s} = A \boldsymbol{f} + \boldsymbol{b}$$

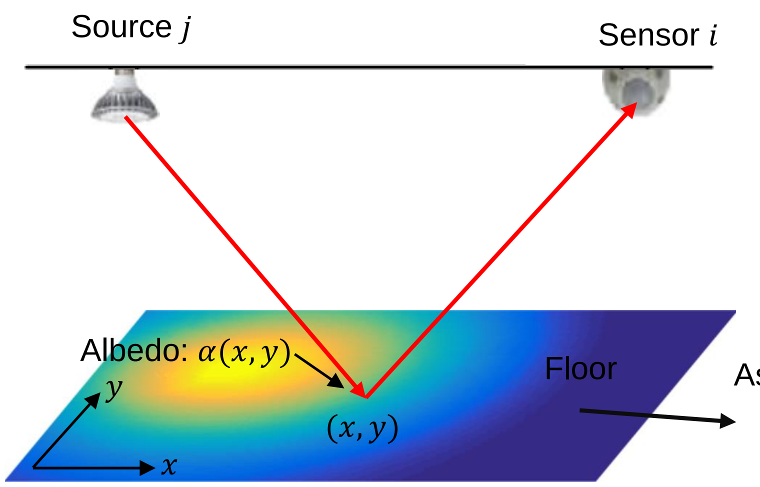
$$\boldsymbol{s} + \Delta\boldsymbol{s} = A \underbrace{(\boldsymbol{f} + \Delta\boldsymbol{f})}_{\text{Perturbation}} + \boldsymbol{b}$$

$$\Delta\boldsymbol{s} = A \, \Delta\boldsymbol{f}$$

$$\underbrace{[\Delta\boldsymbol{s}(t_1) \quad \cdots \quad \Delta\boldsymbol{s}(t_n)]}_{\Delta S} = A \underbrace{[\Delta\boldsymbol{f}(t_1) \quad \cdots \quad \Delta\boldsymbol{f}(t_n)]}_{\Delta F \text{ (designed)}}$$

$$\boxed{\Rightarrow \quad A = \Delta S \Delta F^{\mathrm{T}} (\Delta F \Delta F^{\mathrm{T}})^{-1}}$$

Light transport matrix A in detail



Assumption: Lambertian floor
$C(i,j;*,*)$: Map of floor contributions to the sensor reading

Light contribution
Source $j \rightarrow$ Sensor i

$$A(i,j) = \int_{(x,y)} \underbrace{C(i,j;x,y)}_{\text{Function of room geometry}} \; \underbrace{\alpha(x,y)}_{\text{Floor albedo}} \; dxdy$$

Change in $A \leftrightarrow$ Change in Floor Albedo

Initial state (empty): $A_0(i,j) = \int_{(x,y)} C(i,j;x,y)\alpha_0(x,y)\,dxdy$

New state (occupied): $A(i,j) = \int_{(x,y)} C(i,j;x,y)\alpha(x,y)\,dxdy$

Change: $\underbrace{A(i,j) - A_0(i,j)}_{\Delta A(i,j)} = \int_{(x,y)} C(i,j;x,y)\big(\underbrace{\alpha(x,y) - \alpha_0(x,y)}_{\Delta\alpha(x,y)}\big)\,dxdy$

Matrix-vector form:

$$\begin{array}{c} \text{source-} \\ \text{sensor} \\ \text{pairs} \\ (i,j) \end{array} \left\{ \underbrace{\begin{bmatrix} \Delta\boldsymbol{A} \end{bmatrix}}_{\text{vector}} \right. = \begin{array}{c} \text{source-} \\ \text{sensor} \\ \text{pairs} \\ (i,j) \end{array} \left\{ \underbrace{\overbrace{\begin{bmatrix} \quad \boldsymbol{C} \quad \end{bmatrix}}^{\text{floor positions } (x,y)}}_{\text{matrix}} \right. \underbrace{\begin{bmatrix} \Delta\boldsymbol{\alpha} \end{bmatrix}}_{\text{vector}} \left. \right\} \begin{array}{c} \text{floor} \\ \text{positions} \\ (x,y) \end{array}$$

- Known: $\Delta\boldsymbol{A}$, $\boldsymbol{C}$; solve for floor albedo change $\Delta\boldsymbol{\alpha} \rightarrow$ location of change

Localization algorithm

- **Step 0:** Ridge regression over the whole floor [Zhao et al., IEEE-ICASSP, 2017]

$$\Delta\boldsymbol{\alpha}_0{}^* = \underset{\Delta\boldsymbol{\alpha}}{\arg\min}\ \left(\|\Delta\boldsymbol{A} - \boldsymbol{C}\Delta\boldsymbol{\alpha}\|_{l_2}{}^2 + \sigma\|\Delta\boldsymbol{\alpha}\|_{l_2}{}^2\right)$$

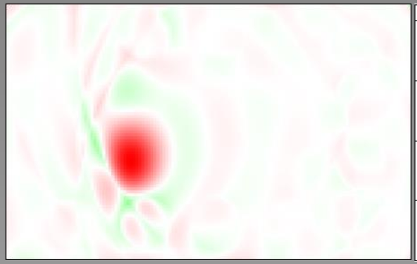
- **Step 1:** Threshold → coarse localization

$$Q = \{(x,y): |\Delta\boldsymbol{\alpha}_0{}^*(x,y)| \geq \tau\}$$

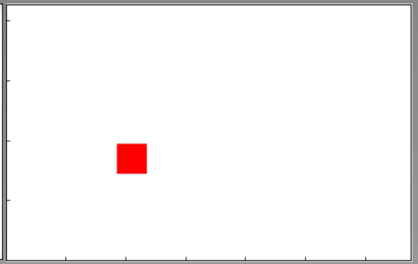
- **Step 2:** Ridge regression inside region of interest Q

$$\Delta\boldsymbol{\alpha}^* = \underset{\Delta\boldsymbol{\alpha}}{\arg\min}\ \left(\|\Delta\boldsymbol{A} - C\Delta\boldsymbol{\alpha}\|_{l_2}{}^2 + \sigma\|\Delta\boldsymbol{\alpha}\|_{l_2}{}^2\right)$$
$$\text{s.t.}\ \ \Delta\alpha(x,y) = 0, \forall(x,y) \notin Q$$

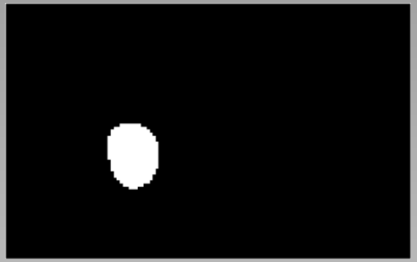
- **Step 3:** Estimated location: centroid of $|\Delta\boldsymbol{\alpha}^*|$



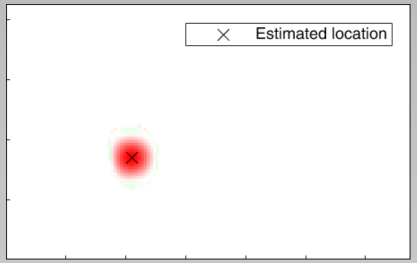
Step 0: Map of $\Delta\boldsymbol{\alpha}^{(0)}$. Red: positive, green: negative



Ground truth of $\Delta\boldsymbol{\alpha}$



Step 1



Steps 2 and 3

Passive localization (model based)

$$\text{Sensor } i \longleftarrow \begin{bmatrix} s(1) \\ \vdots \\ s(i) \\ \vdots \\ s(N_s) \end{bmatrix} = \begin{bmatrix} A(1,1) & \cdots & \cdots \\ \vdots & A(i,j) & \vdots \\ \cdots & \cdots & A(N_s,N_f) \end{bmatrix} \begin{bmatrix} f(1) \\ \vdots \\ f(j) \\ \vdots \\ f(N_f) \end{bmatrix} \longrightarrow \text{Source } j \; + \underbrace{\begin{bmatrix} b(1) \\ \vdots \\ b(N_s) \end{bmatrix}}_{\text{Ambient light}}$$

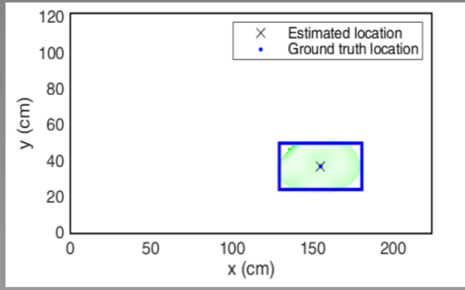
$$\boldsymbol{s} = A\,(\boldsymbol{no\ object})\,\boldsymbol{f} + \boldsymbol{b}$$
$$\boldsymbol{s} + \Delta\boldsymbol{s} = A\,(\boldsymbol{with\ object})\,\boldsymbol{f} + \boldsymbol{b}$$

$$\Delta\boldsymbol{s} = \underbrace{(A(\boldsymbol{with\ object}) - A(\boldsymbol{no\ object}))}_{\text{Depends on location}}\,\boldsymbol{f}$$

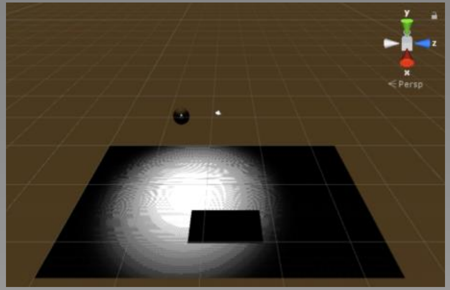
Taking ratios $\frac{\Delta s_i}{\Delta s_j} = f(x_0, y_0)$ for N_s-1 sensor pairs we get rid of $\boldsymbol{f}$ and use constrained least squares to solve for x_0, y_0.

Experiments

- Validation:



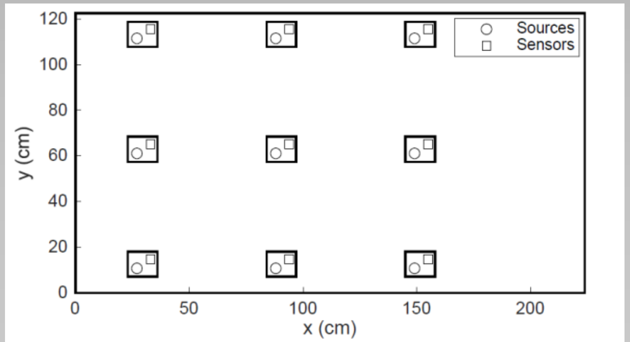
MATLAB simulation

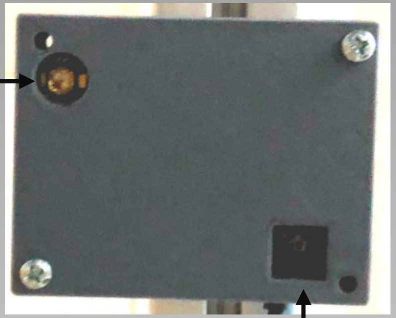

Unity3D simulation


Small-scale real testbed

Idealized — Realistic

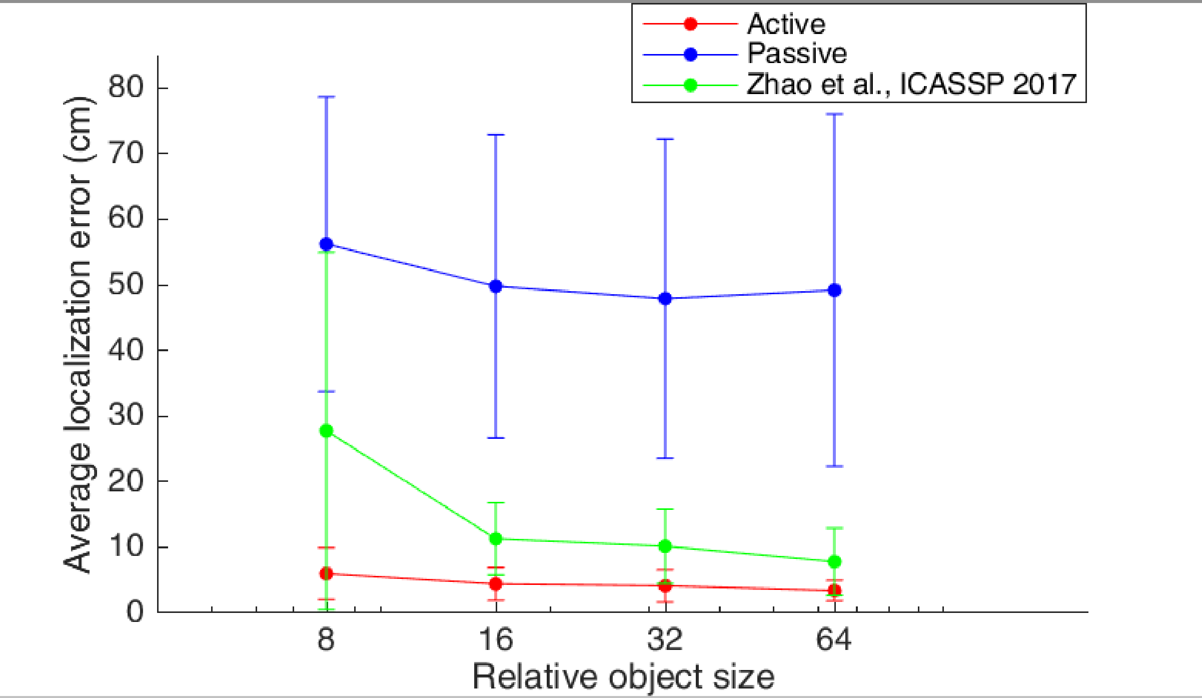
- Room: 1.2m (W) × 2.2m (L) × 0.7m (H)
- Flat objects: from 3cm x 4cm to 26cm x 51cm
- 3x3 layout of light sources/sensors:





Results: Small-scale testbed

- Localization error vs. object size (little ambient light)



Error sources:

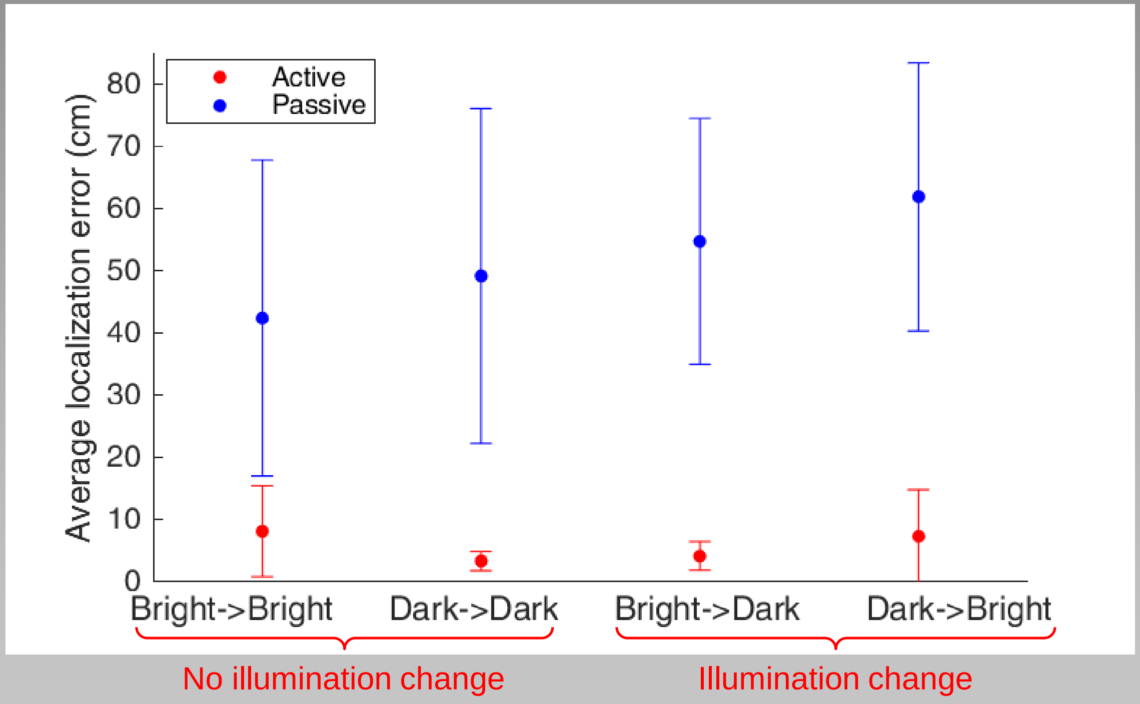
- LED light noise
- Sensor noise
- Interfering objects
- Non-Lambertian floor
- Indirect light

[*Zhao et al., CVPR-COPS, 2018*]

- Active illumination works well in real testbed

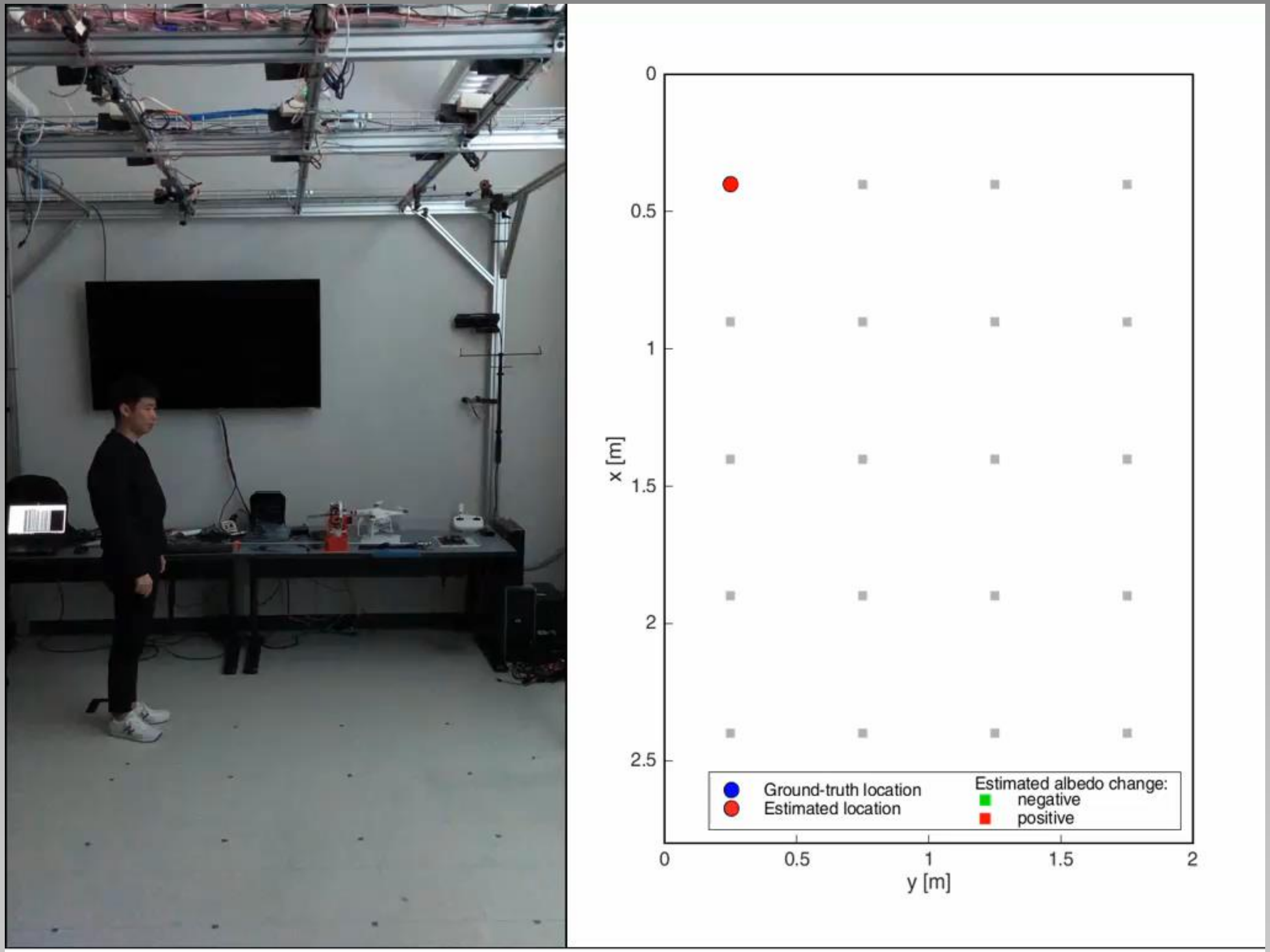
Results: Small-scale testbed

- Localization error vs. illumination change between empty and occupied states (fluorescent light on or off)



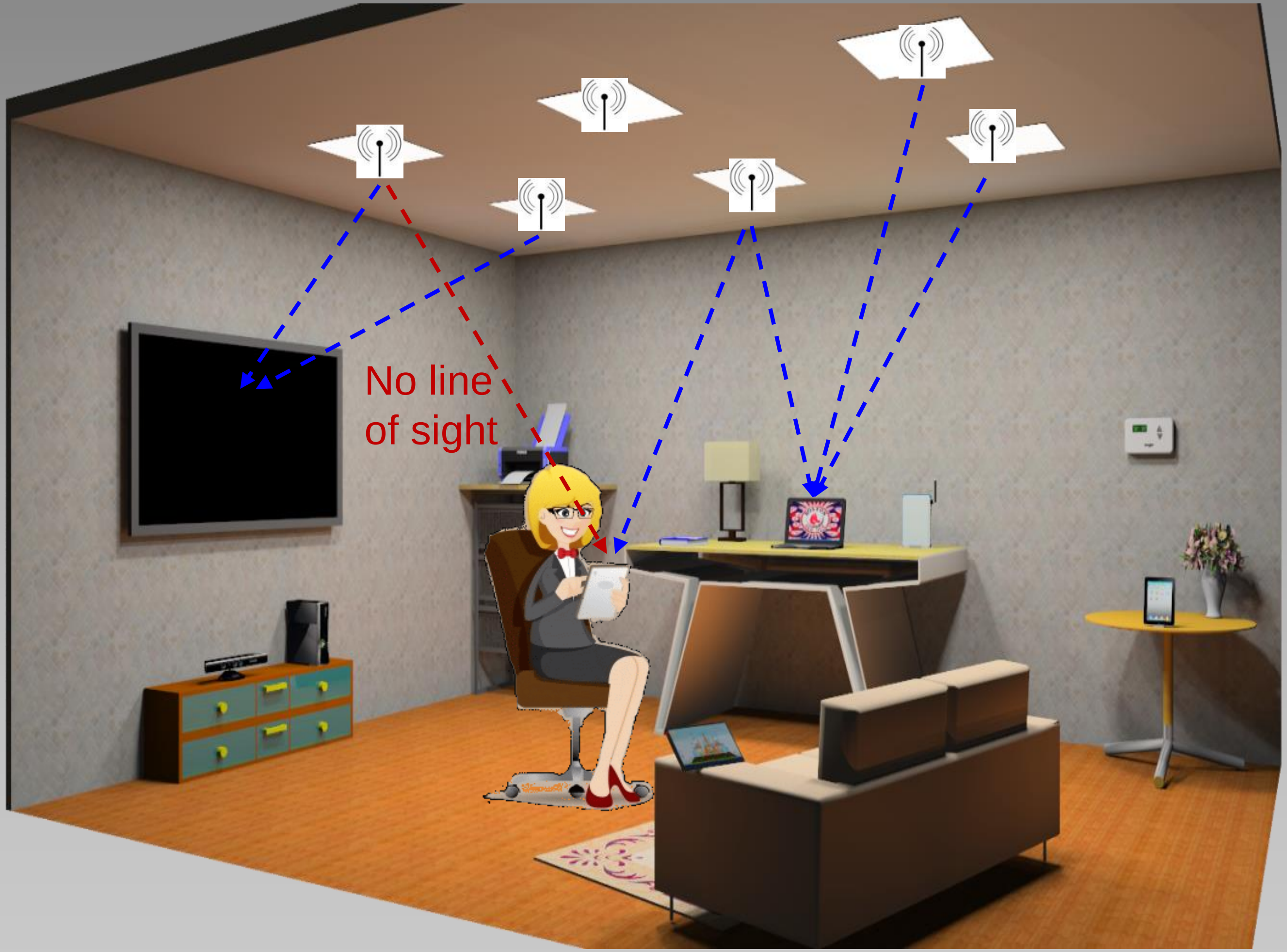
- Active illumination is still robust

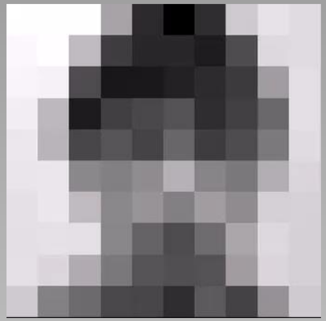
Full-scale testbed



Video speed: ×8. Data collected over 3 light cycles for accuracy. A fully-developed system will produce no visible flicker.

Private enough ?





Can deep learning
disclose visual identity ?

Final thoughts

- Sensors will ot
- Privacy conc
- **Solution:** Us alize and recognize vis llow learning"?
- **Bonus:** Sens th
- **Challenge:** H such sensors