Slides are available during the conference!

The Cutting Edge of First-Person Vision

Ryo Yonetani

Institute of Industrial Science

The University of Tokyo

東京大学 THE UNIVERSITY OF TOKYO

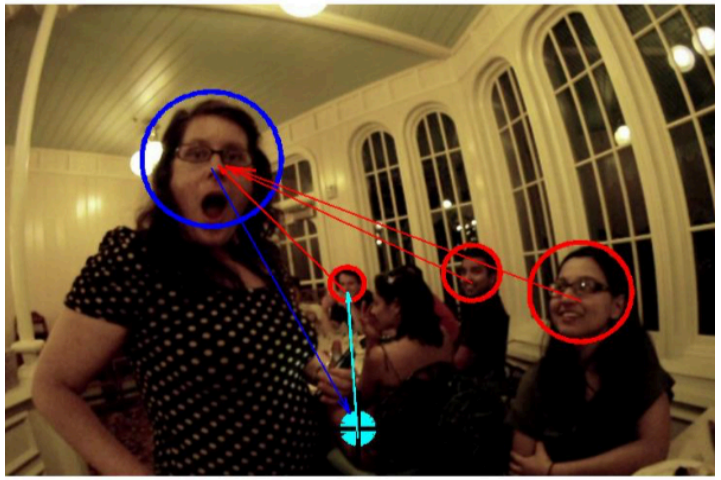
JST CREST Project
COLLECTIVE VISUAL SENSING

Human-Machine Harmonious Collaboration

First-Person Vision (FPV)

Computer vision using wearable cameras mounted on the head

Wearable camera

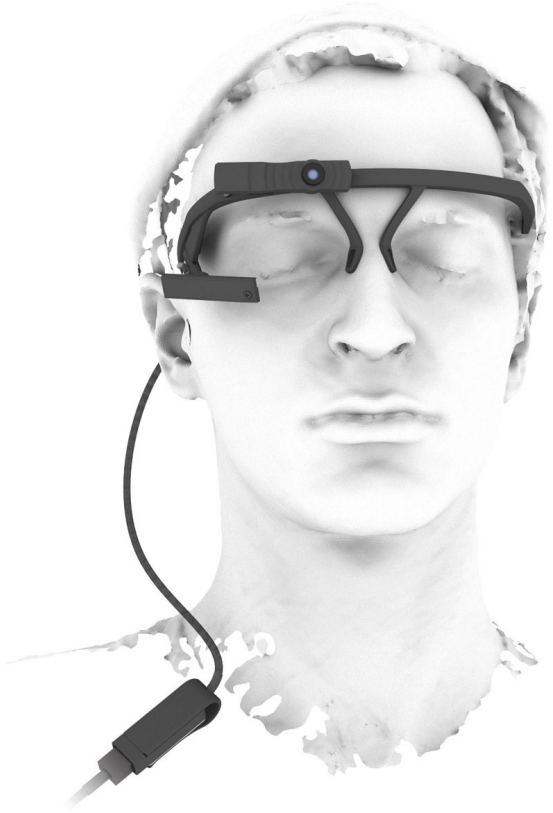


First-person points-of-view videos

FPV in Computer Vision Conferences

- Workshops:
  - Workshop on Egocentric Vision (CVPR2009, 2012, 2014, 2016)
  - Workshop on Egocentric Perception, Interaction and Computing (ECCV2016)
- Tutorials:
  - Group Behavior Analysis and Its Applications (CVPR2015)
  - First-person Visual Sensing: Theory, Models, and Applications (CVPR2016)
- 30 papers in CVPR2016

Startups on FPV



Pupil Labs

https://pupil-labs.com/pupil/

Read

The OrCam MyEye device can read printed text, in real time.

You can read newspapers and books, signs, labels on consumer products and even text on a computer or smartphone screen. Perfect for use at home and on the go.

Identify People

The OrCam MyEye device identifies known faces.

No more guessing - previously stored faces are identified and announced upon entering the camera's view. Less awkward situations, more control of your environment.

ORCAM

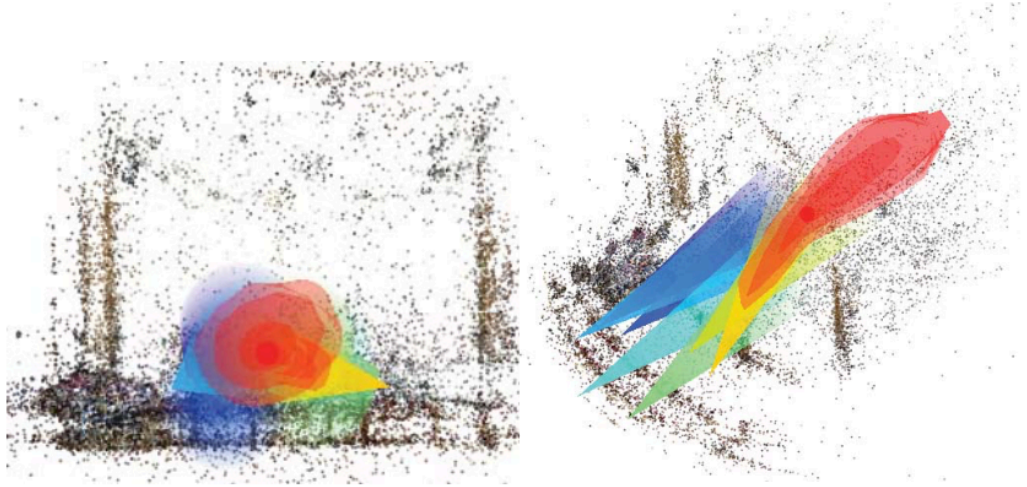
http://www.orcam.com/

Why First-Person Vision?

1. Measuring Attention

Wearable cameras can capture what people see clearly in the form of first-person videos

1. Measuring attention
2. Observing hands
3. Analyzing interactions
4. Leveraging ego-motion
5. Recording Everyday Life

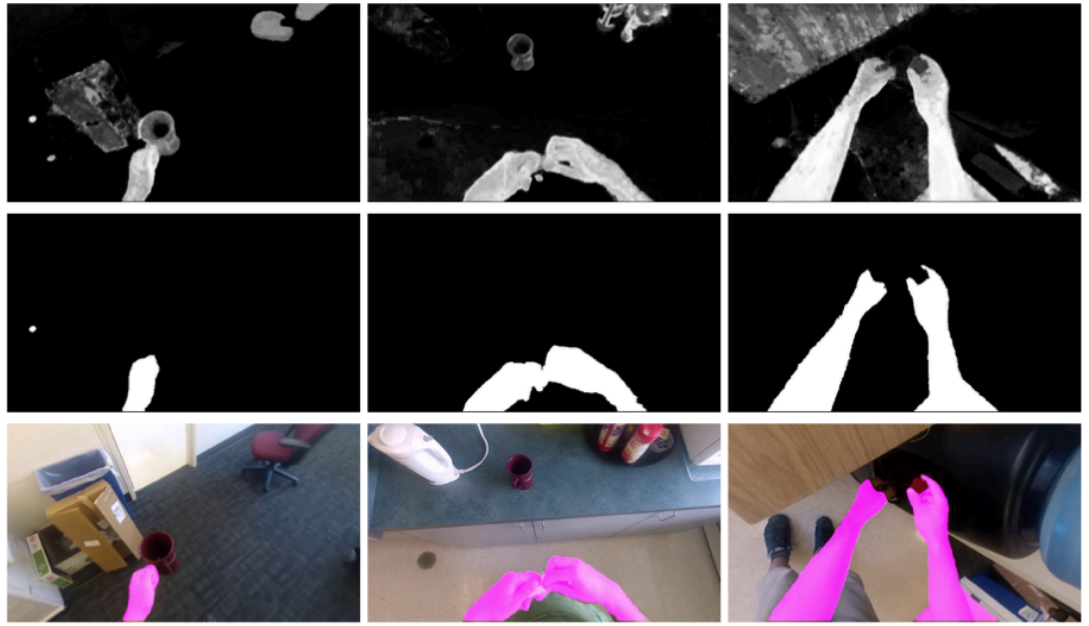


Social saliency [Park+, CVPR'15]

2. Observing hands

Wearable cameras can capture hands in front of the body

1. Measuring attention
2. Observing hands
3. Analyzing interactions
4. Leveraging ego-motion
5. Recording Everyday Life

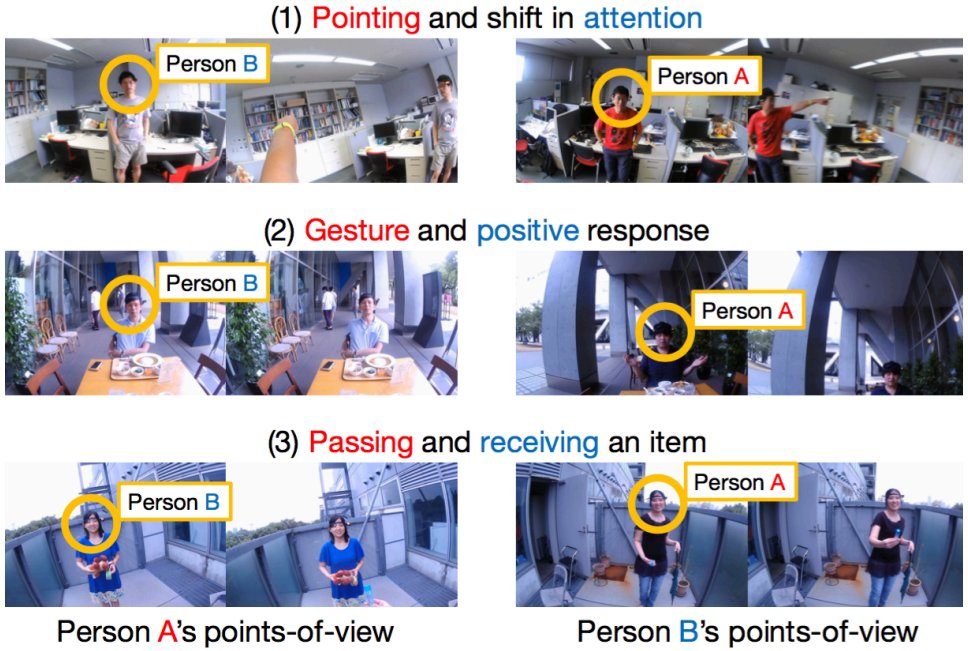


Pixel-level hand detection [Li+, ICCV'15]

3. Analyzing Interactions

Wearable cameras can capture interaction partners

1. Measuring attention
2. Observing hands
3. Analyzing interactions
4. Leveraging ego-motion
5. Recording Everyday Life

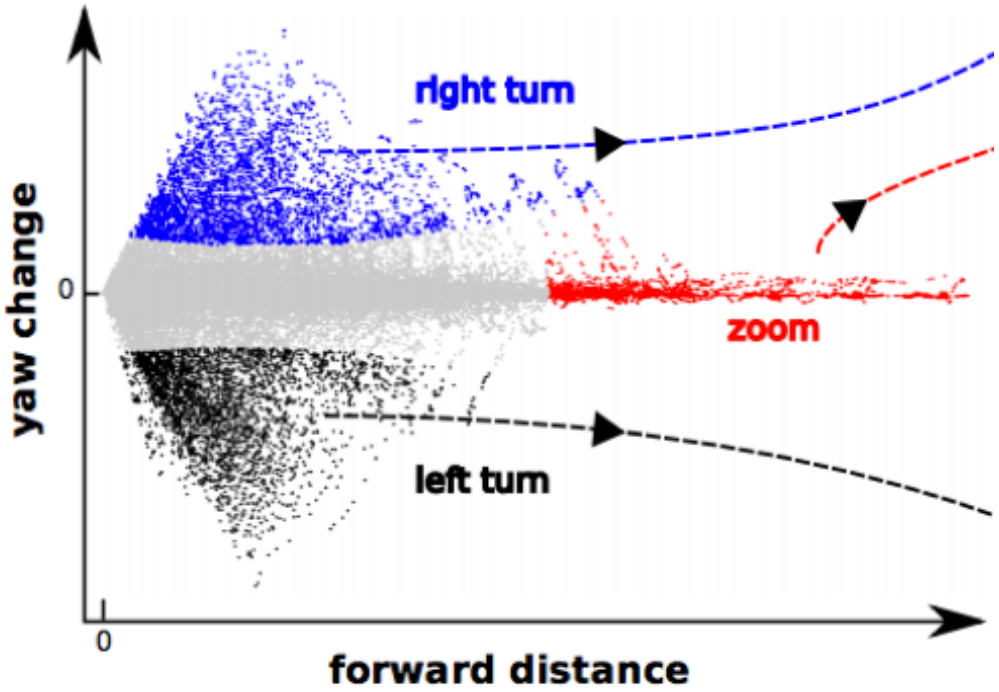


Micro-action [Yonetani+, CVPR'16]

4. Leveraging ego-motion

Camera ego-motion reflects how people move

1. Measuring attention
2. Observing hands
3. Analyzing interactions
4. Leveraging ego-motion
5. Recording Everyday Life



Learning ego-motion [Jayaraman+, ICCV'15]

5. Recording Everyday Life

Analyzing activities at various times and places

1. Measuring attention
2. Observing hands
3. Analyzing interactions
4. Leveraging ego-motion
5. Recording Everyday Life

Action map [Rhinehart+, CVPR'16]

Tutorial Overview

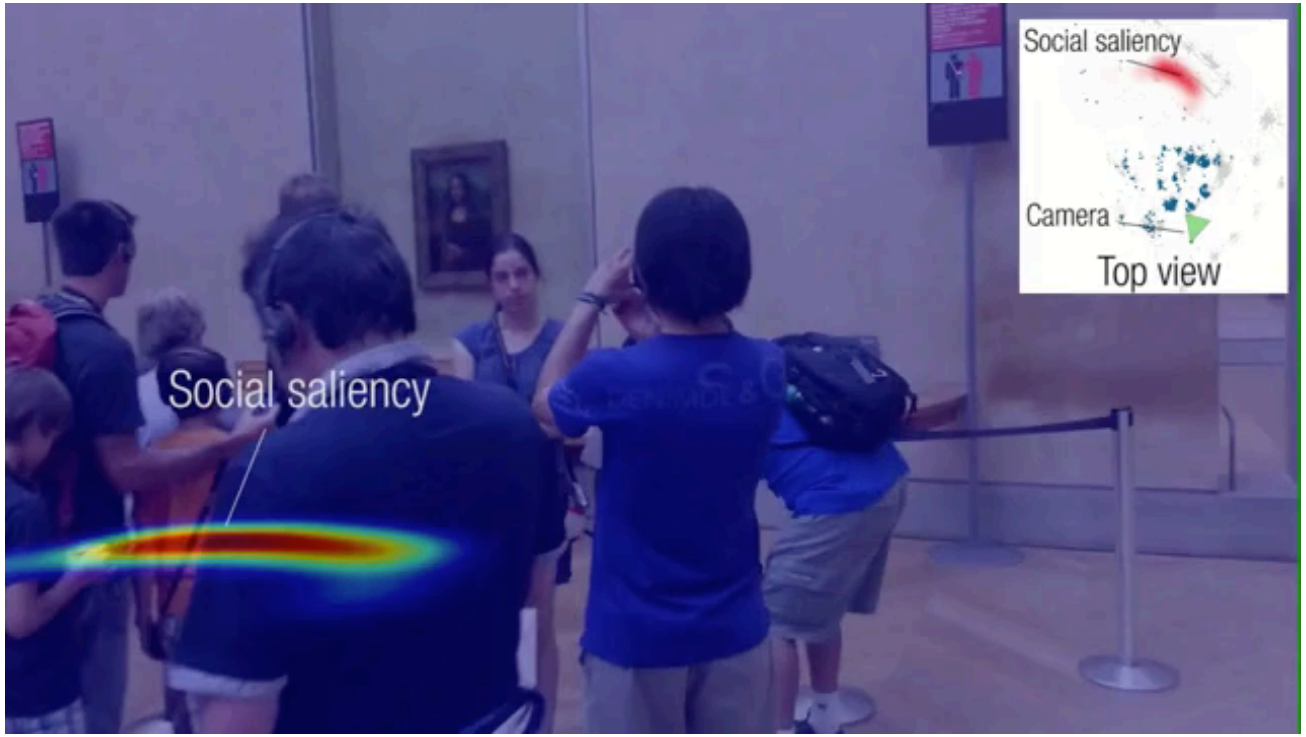
1. Measuring attention
2. Observing hands
3. Analyzing interactions
4. Leveraging ego-motion
5. Recording Everyday Life
6. Summary, future work

- 2~3 papers from recent conferences
- Discussion for future work

1. Measuring Attention

Social Saliency

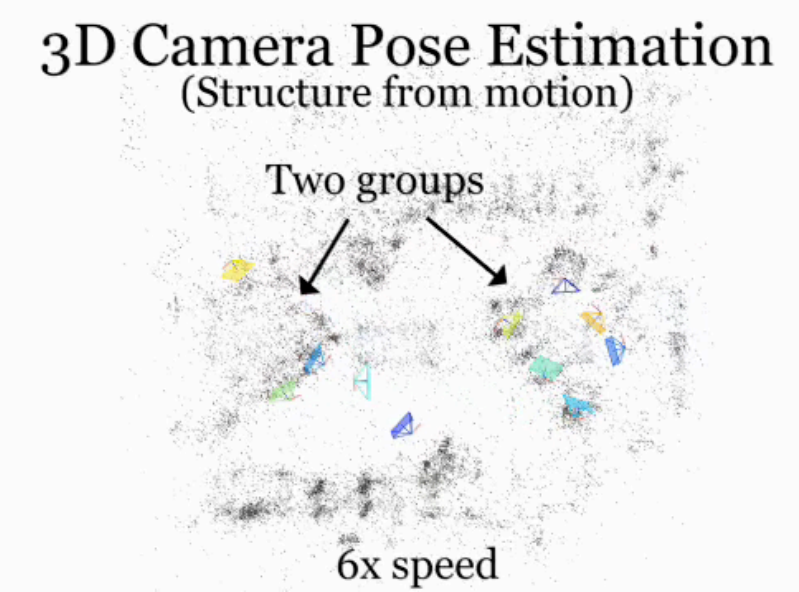
Park and Shi, "Social Saliency Prediction", CVPR'15

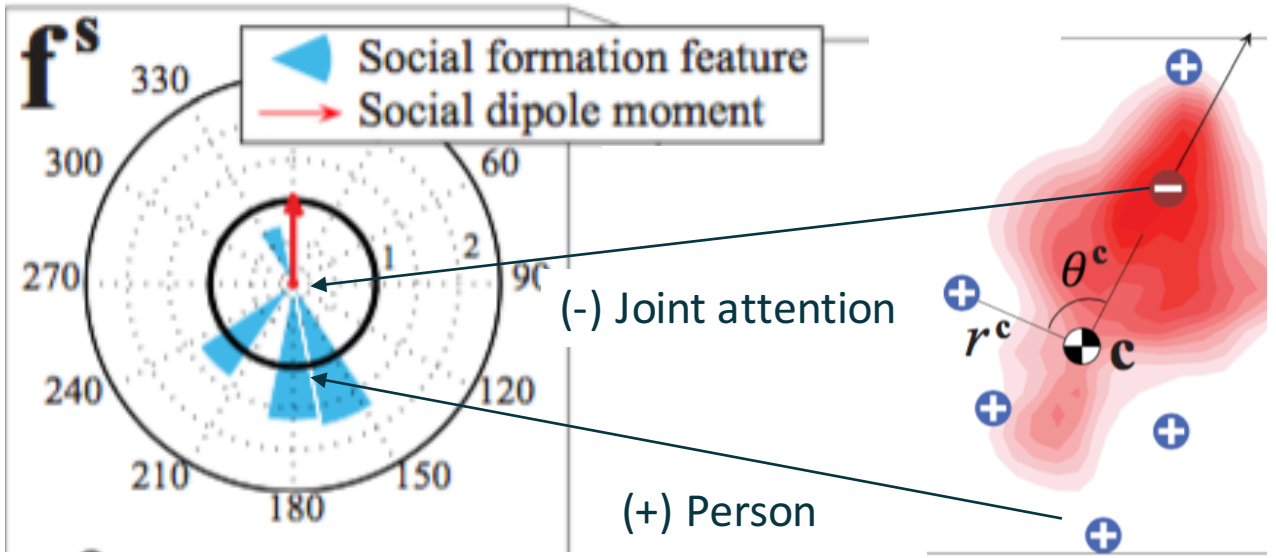


Given first-person videos, predicting where many people look at

Social Saliency

Park and Shi, “Social Saliency Prediction”, CVPR’15

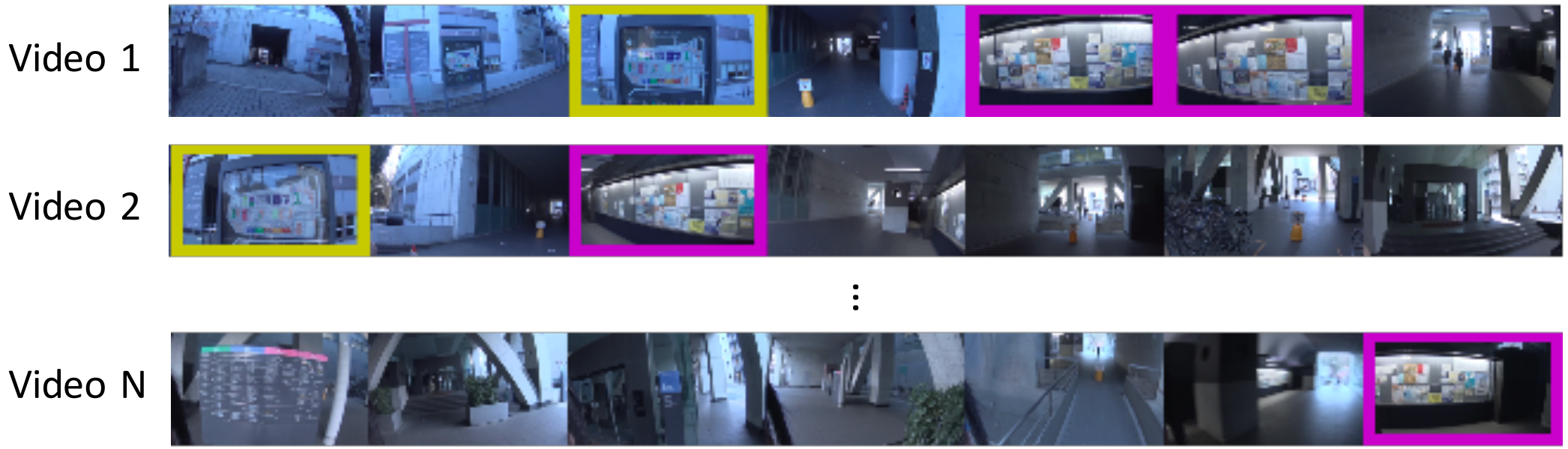




- Localizing people and joint attention via SfM [Park+, NIPS’12]
- Features: relative locations of people w.r.t. joint attention

Visual Motif Discovery

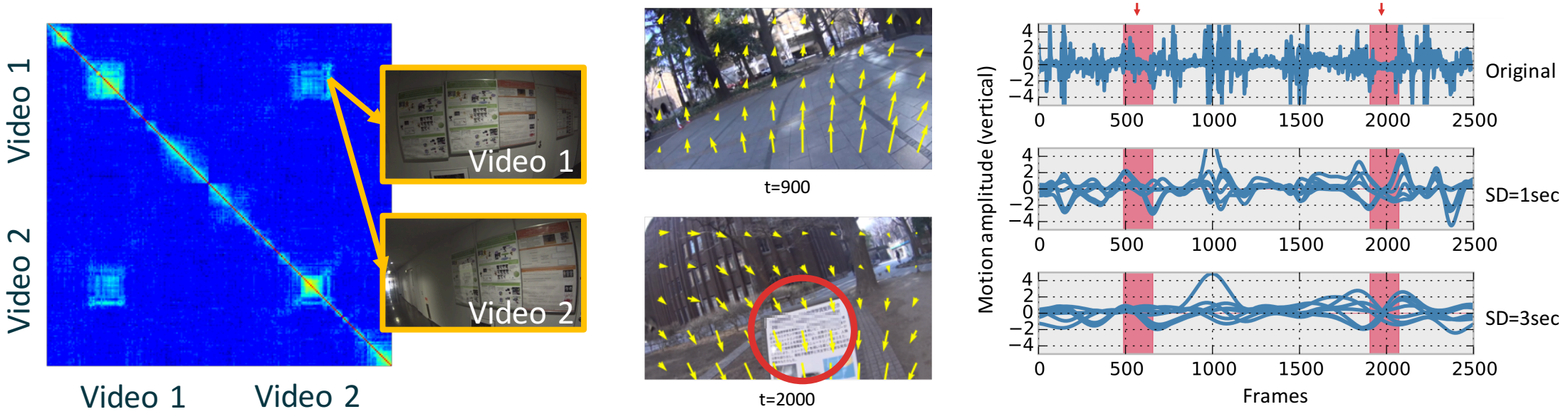
Yonetani, Kitani, and Sato, "Visual Motif Discovery via First-Person Vision", ECCV'16



Discovering significant scenes shared across many people from a collection of first-person videos

Visual Motif Discovery

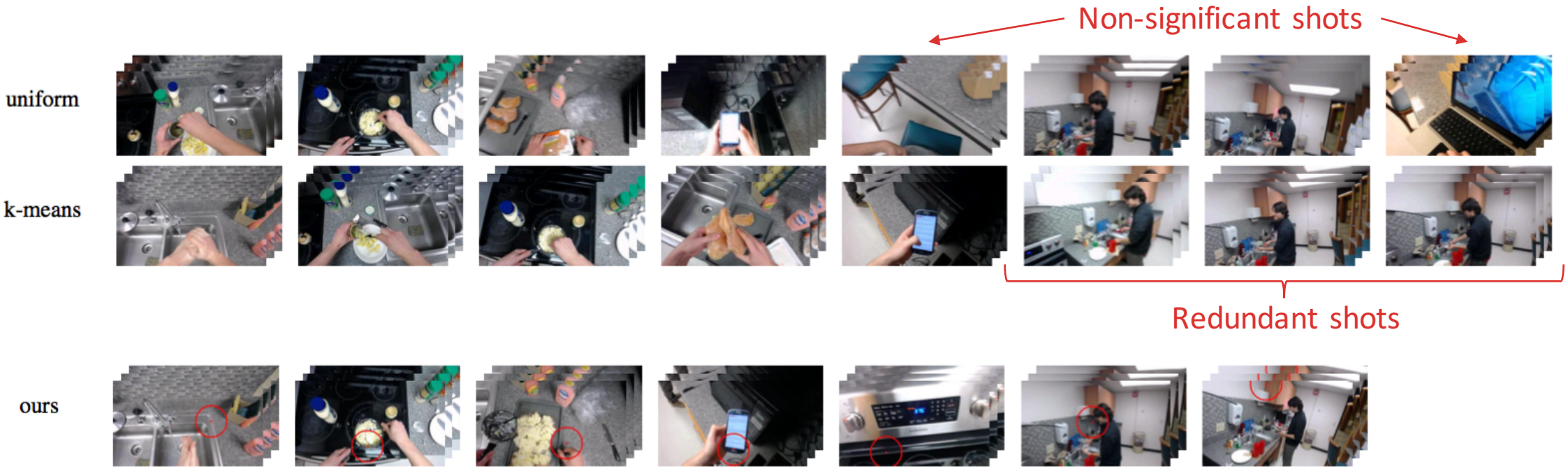
Yonetani, Kitani, and Sato, "Visual Motif Discovery via First-Person Vision", ECCV'16



- Commonality clustering to find similar scenes across multiple videos
- Egocentric action recognition to specify important moments

Gaze-enabled Video Summarization

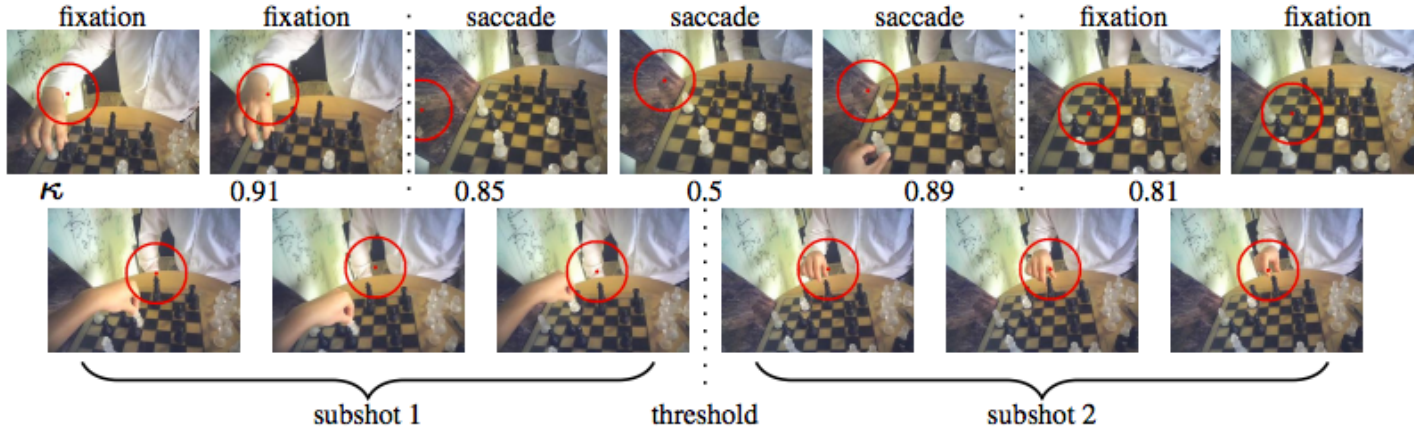
Xu, Mukherjee, Li, Warner, Rehg, and Singh, "Gaze-enabled Egocentric Video Summarization via Constrained Submodular Maximization", CVPR'15



Using gaze fixations to find significant shots for video summarization

Gaze-enabled Video Summarization

Xu, Mukherjee, Li, Warner, Rehg, and Singh, "Gaze-enabled Egocentric Video Summarization via Constrained Submodular Maximization", CVPR'15



Shot segmentation based on gaze:
Detecting where large saccades occur

$$\text{Summary} = \underset{\text{Selected shots}}{\text{ArgMax}} \left\{ \text{Mutual information between selected and non-selected shots} + \text{Number of fixations in selected shots} \right\}$$

Finding a set of shots informative and containing many fixations

Measuring Attention – Other Work

- Action recognition using gaze [Fathi+, ECCV'12]
  - Gaze = strong cue for hand manipulations
- Predicting gaze in first-person videos [Li+, ICCV'13]
- Discovering joint attention [Kera+, CVPRW'16; PS2-64 (Aug. 3)]
  - Multiple first-person videos + gaze for detecting objects of joint attention

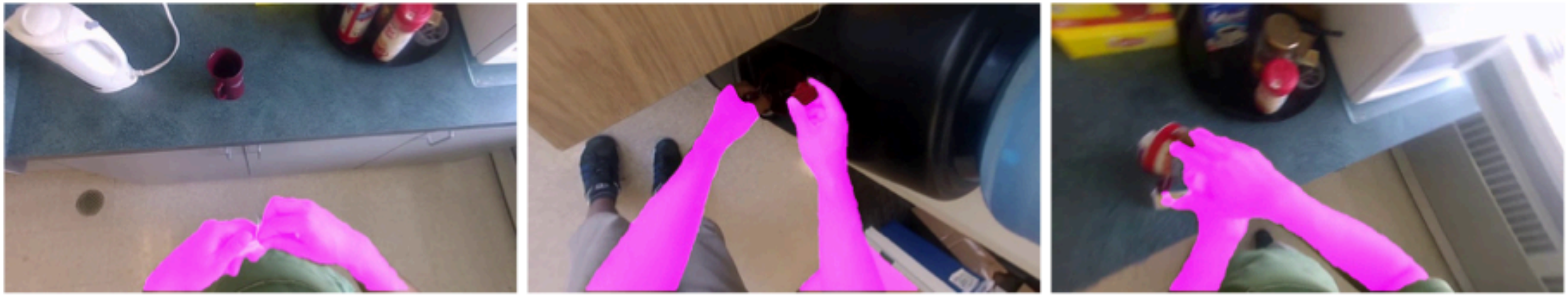
Measuring Attention – Discussion

- Analyzing attention in large-scale environments
  - Large-scale structure-from motion, efficient scene clustering
- Introducing gaze data
  - Pros: able to measure attention at pixel / object levels
  - Cons: more difficult to conduct large-scale experiments

2. Observing Hands

Pixel-Level Hand Detection

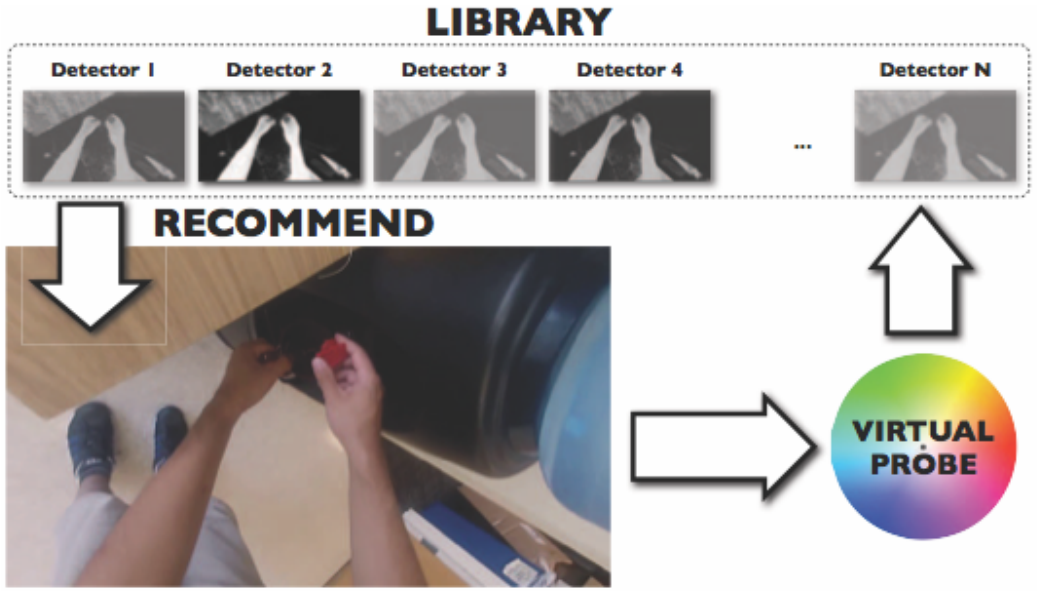
Li and Kitani, "Model Recommendation with Virtual Probes for Egocentric Hand Detection", ICCV'13

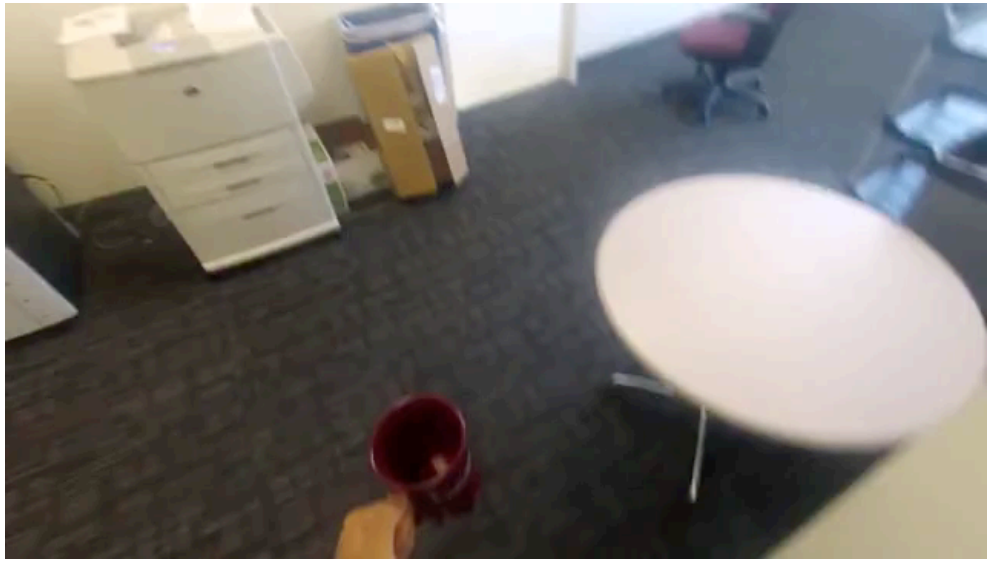


- Detecting hands at pixel-level
- Need to cope with variability in hand poses, luminance, etc.

Pixel-Level Hand Detection

Li and Kitani, "Model Recommendation with Virtual Probes for Egocentric Hand Detection", ICCV'13

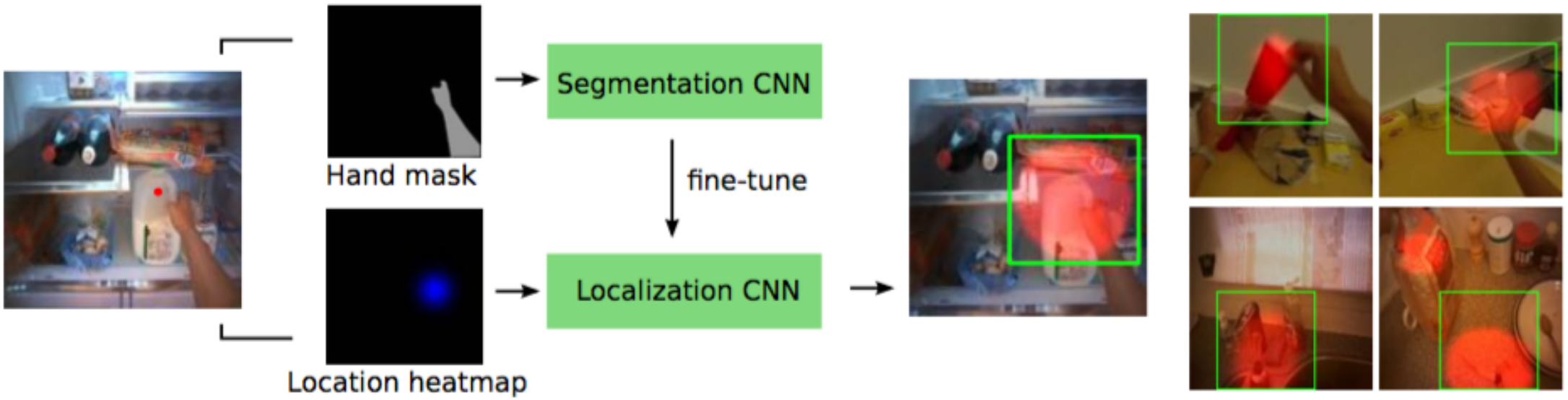




Recommending n-best hand detectors using probes (e.g., global appearance) from testing images

Egocentric Object Recognition

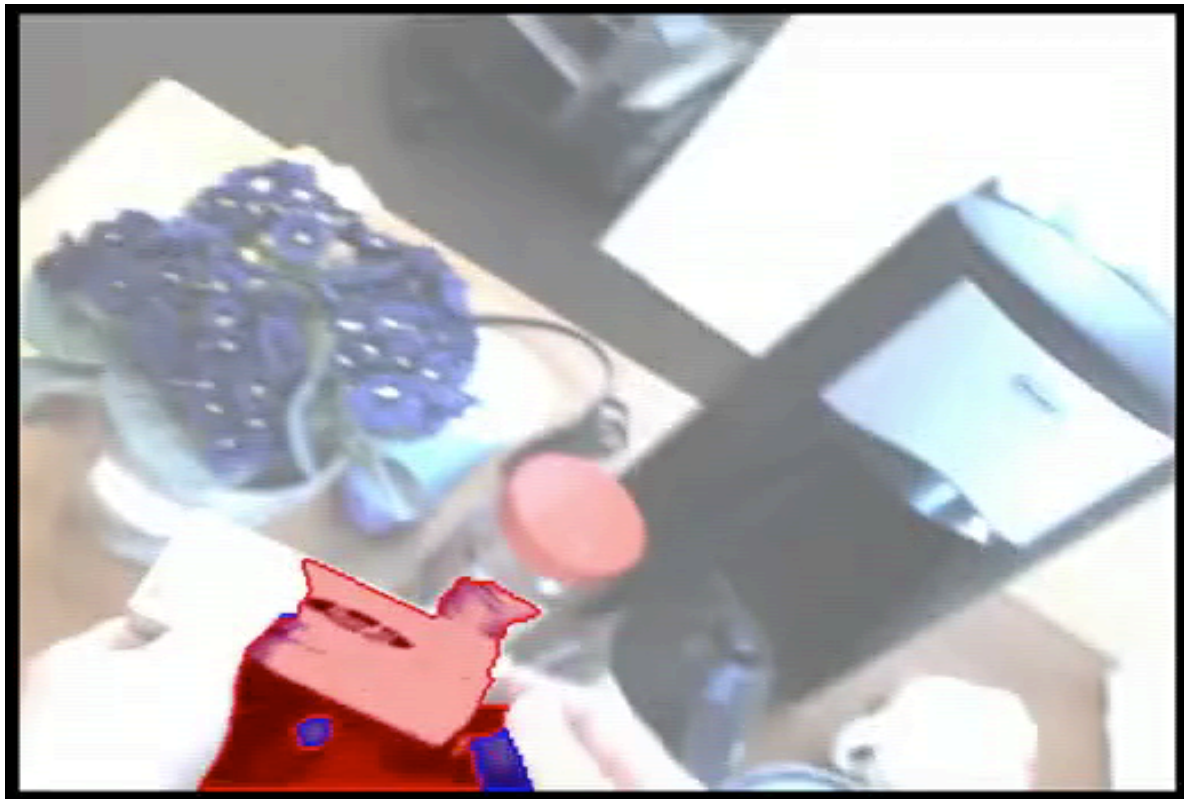
Ma, Fan, and Kitani, "Going Deeper into First-Person Action Recognition", CVPR'16



- Learning hand detectors to improve object localization via fine-tuning
- Improving state-of-the-art handled object recognition (61% -> 76%)

Egocentric Activity Recognition

Ma, Fan, and Kitani, "Going Deeper into First-Person Action Recognition", CVPR'16

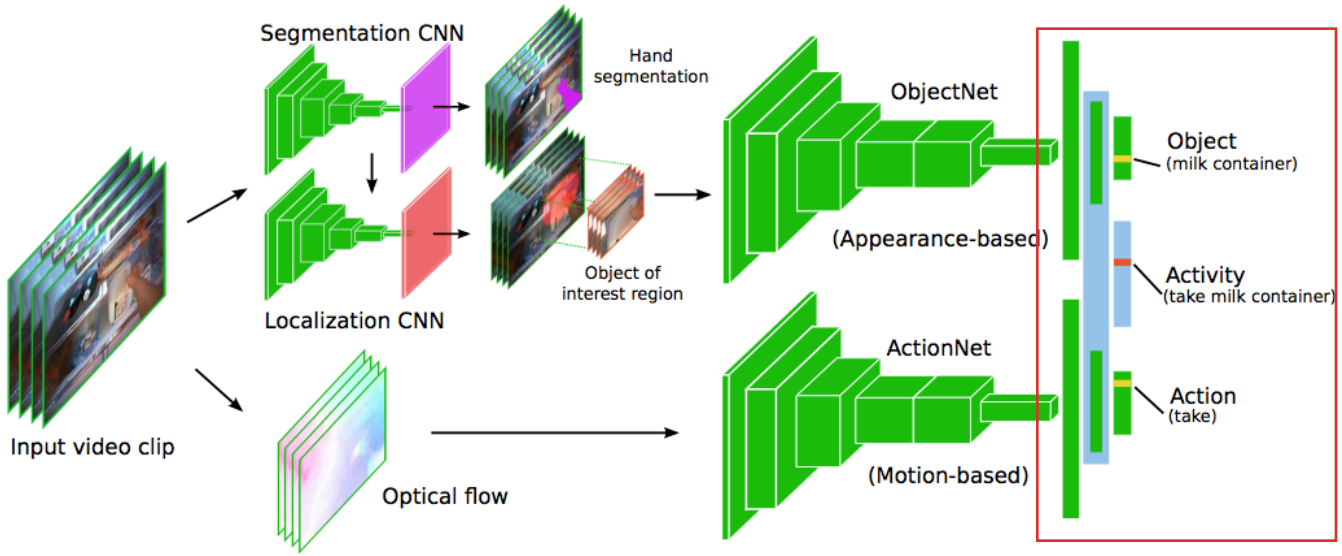


GeogiaTech Egocentric Activities [Fathi+, CVPR'11]

- 71 activity categories (e.g., close coffee, put jam)
- 4 subjects

Egocentric Activity Recognition

Ma, Fan, and Kitani, "Going Deeper into First-Person Action Recognition", CVPR'16



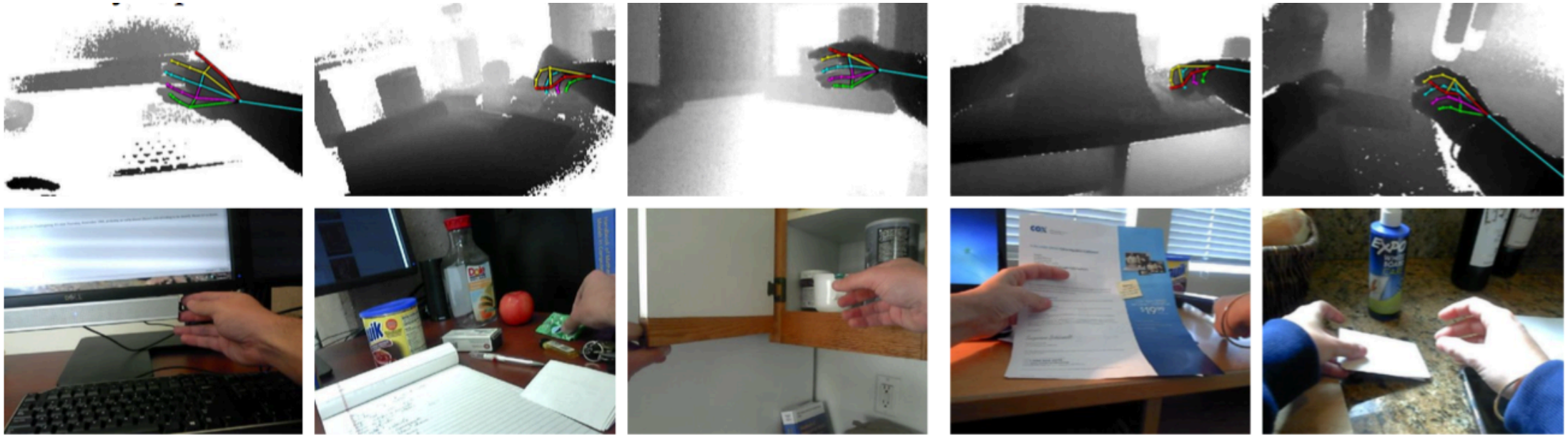
	Methods	GTEA(61)*
Li *et al.*[20]	O+M+E+H	61.10
	O+M+E+G	N/A
	O+E+H	66.80
S. & Z.[30]	temporal-cnn	34.30
	spatial-cnn	53.77
	temporal+spatial-svm	46.51
	temporal+spatial-joint	57.64
Ours	object-cnn	60.02
	motion+object-svm	53.01
	motion+object-joint	**75.08**

Fusion layer to recognize activities

- Jointly training object recognition and action recognition networks
- Improving state-of-the-art egocentric activity recognition (67% -> 75%)

3d Hand Pose Recognition

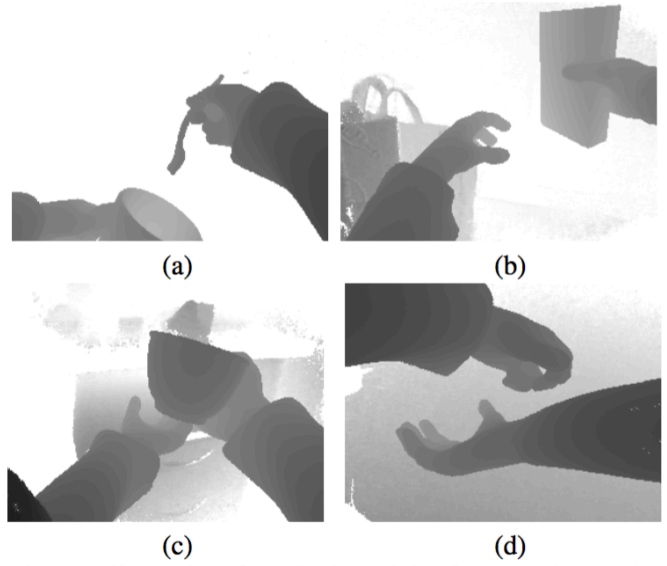
Rogez, Spancic III, and Ramanan, "First-Person Pose Estimation using Egocentric Workspaces", CVPR'15

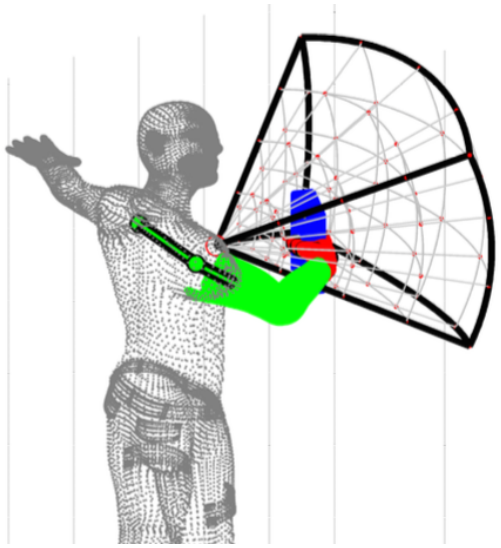


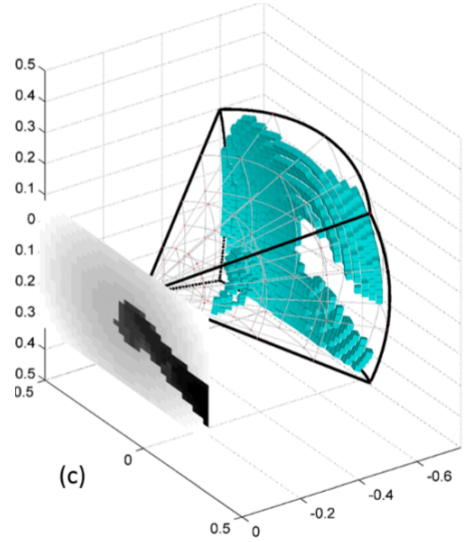
Estimating the 3D pose of arms and hands from a depth camera

3d Hand Pose Recognition

Rogez, Spancic III, and Ramanan, "First-Person Pose Estimation using Egocentric Workspaces", CVPR'15







- Synthesize samples with various configurations of arms, hands, objects, and background
- Learning arm and hand poses from binarized volumetric features (occupancy for each spatiotemporal bin in egocentric workspaces)

Observing Hands – Discussion

- Objects, actions, and hands are correlated
- Many 3d datasets available

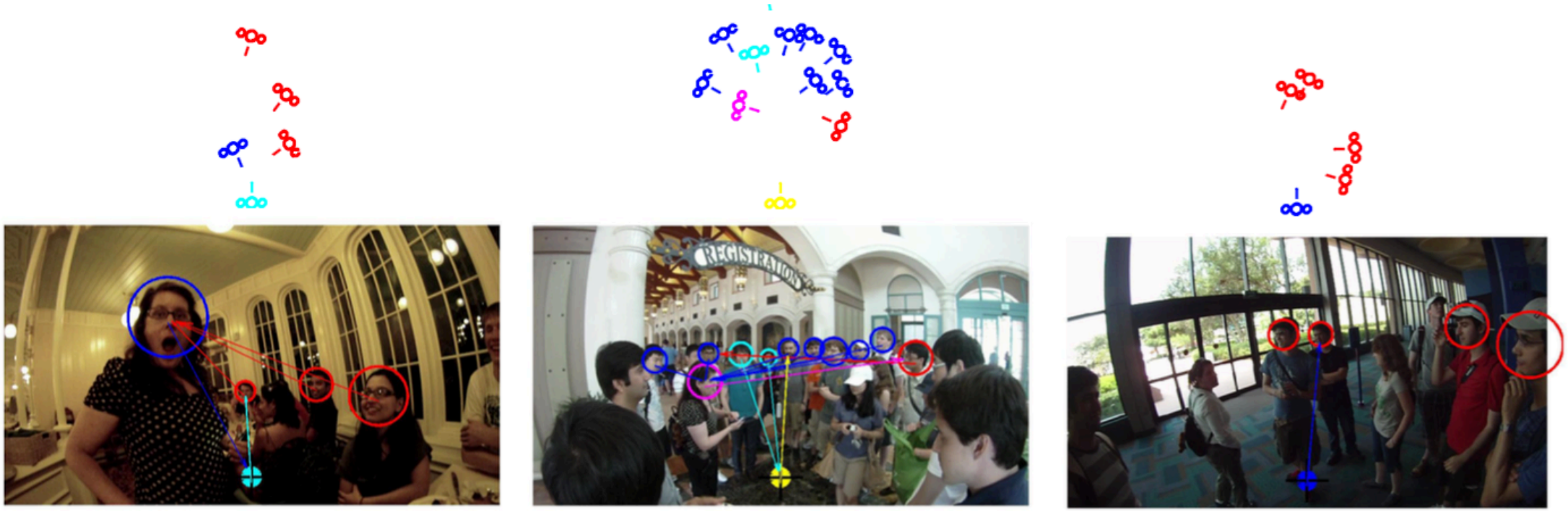
Dataset	Chal.	Scn.	Annot.	Frms.	Sub.	Cam.	Dist. (mm)
ASTAR [51]	A	1	435	435	10	ToF	270-580
Dexter 1 [42]	A	1	3,157	3,157	1	Both	100-989
MSRA [33]	A	1	2,400	2,400	6	ToF	339-422
ICL [45]	A	1	1,599	1,599	1	Struct	200-380
FORTH [28]	AV	1	0	7,148	5	Struct	200-1110
NYU [47]	AV	1	8,252	8,252	2	Struct	510-1070
KTH [30]	AVC	1	0	46,000	9	Struct	NA
UCI-EGO [35]	AVC	4	364	3,640	2	ToF	200-390
Ours	AVC	10+	23,640	23,640	10	Both	200-1950

"Depth-based hand pose estimation: methods, data, and challenges" [Spancic III+, arXiv]

3. Analyzing Interactions

Social Interaction

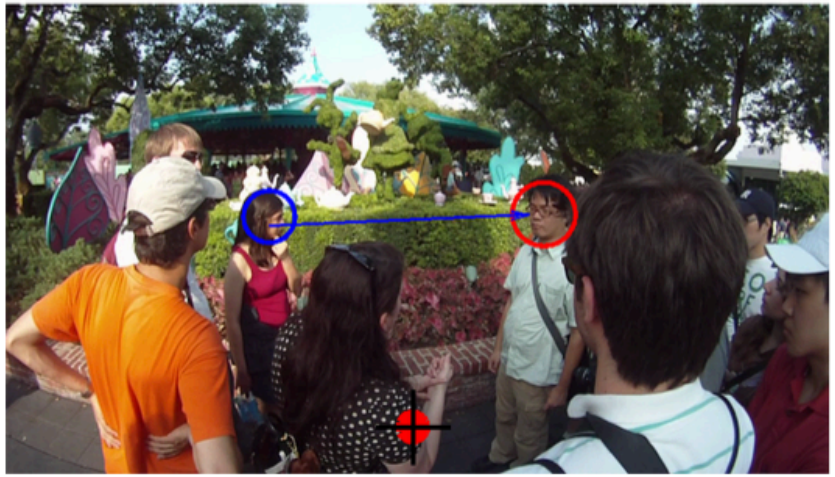
Fathi, Hodgins, and Rehg, "Social Interaction: A First-Person Perspective", CVPR'12

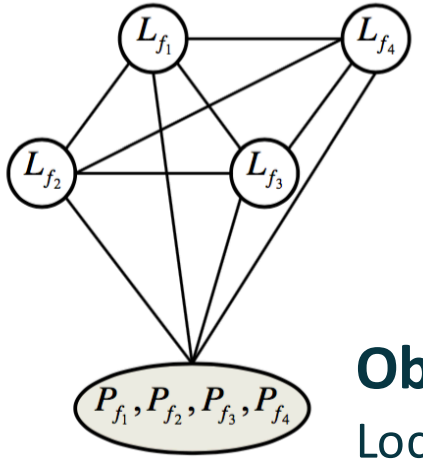


Recognizing types of social interactions based on attention

Social Interaction

Fathi, Hodgins, and Rehg, "Social Interaction: A First-Person Perspective", CVPR'12





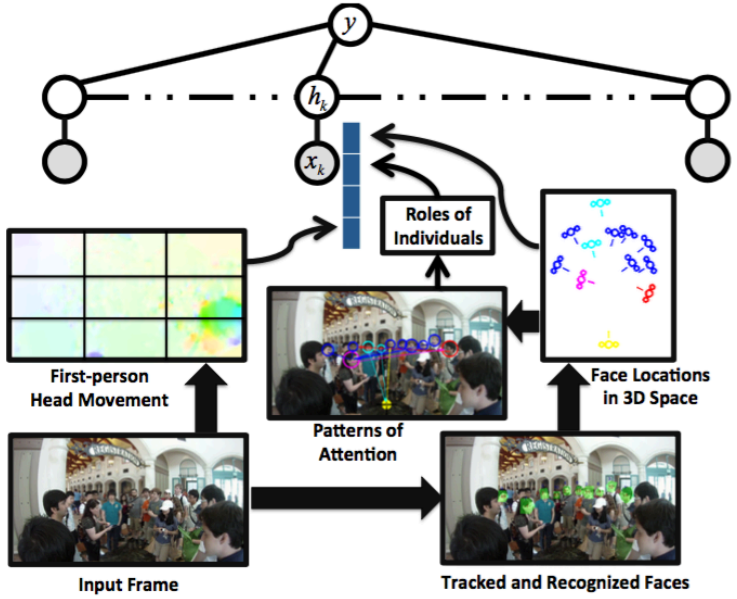
Nodes:
Attention location of each person

Observations:
Location and orientation of all faces

- MRF for attention estimation that considers:
  - Unary: attention locations should be on faces
  - Pairwise: attention locations should be the same across people

Social Interaction

Fathi, Hodgins, and Rehg, "Social Interaction: A First-Person Perspective", CVPR'12



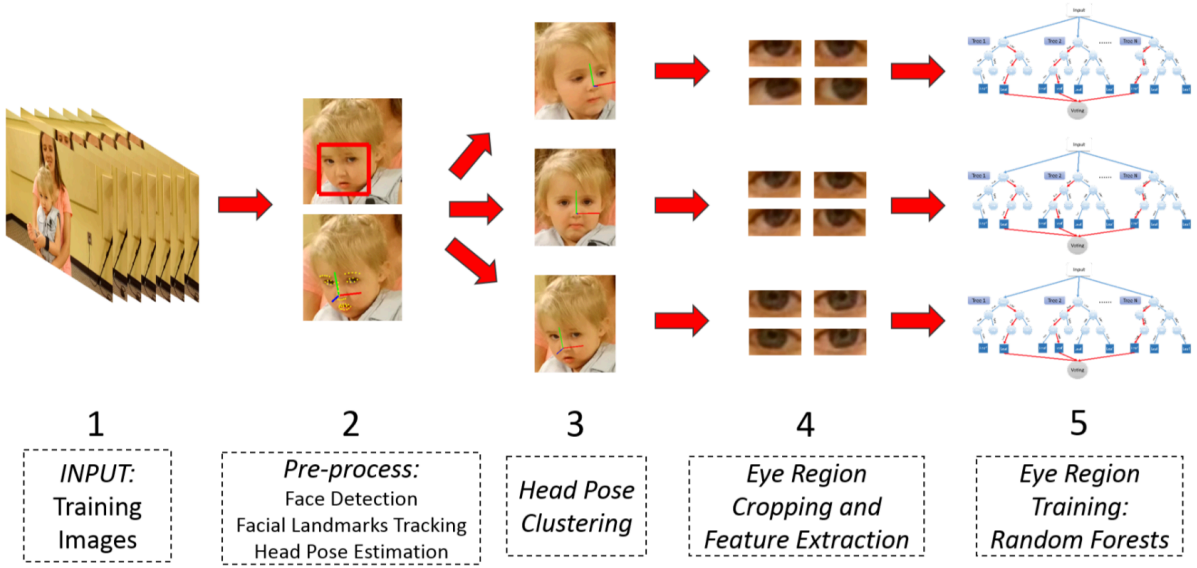
Attention-based features for a person A:

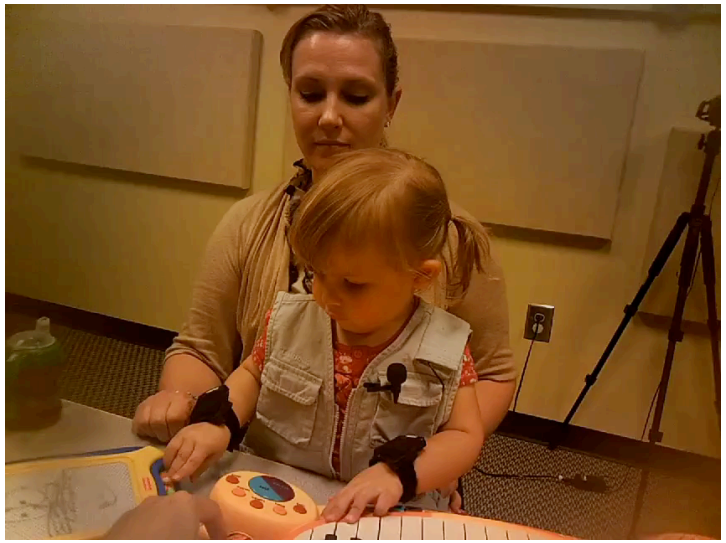
- Number of faces looking at A
- Whether first-person looks at A
- If there is mutual attention between A and first-person
- Number of faces looking at where A is attending

Classifying interaction types: dialog, discussion, monologue, walking…

Eye Contact Detection

Ye, Li, Liu, Bridges, Rozga, and Rehg, "Detecting Bids for Eye Contact Using a Wearable Camera", FG'15

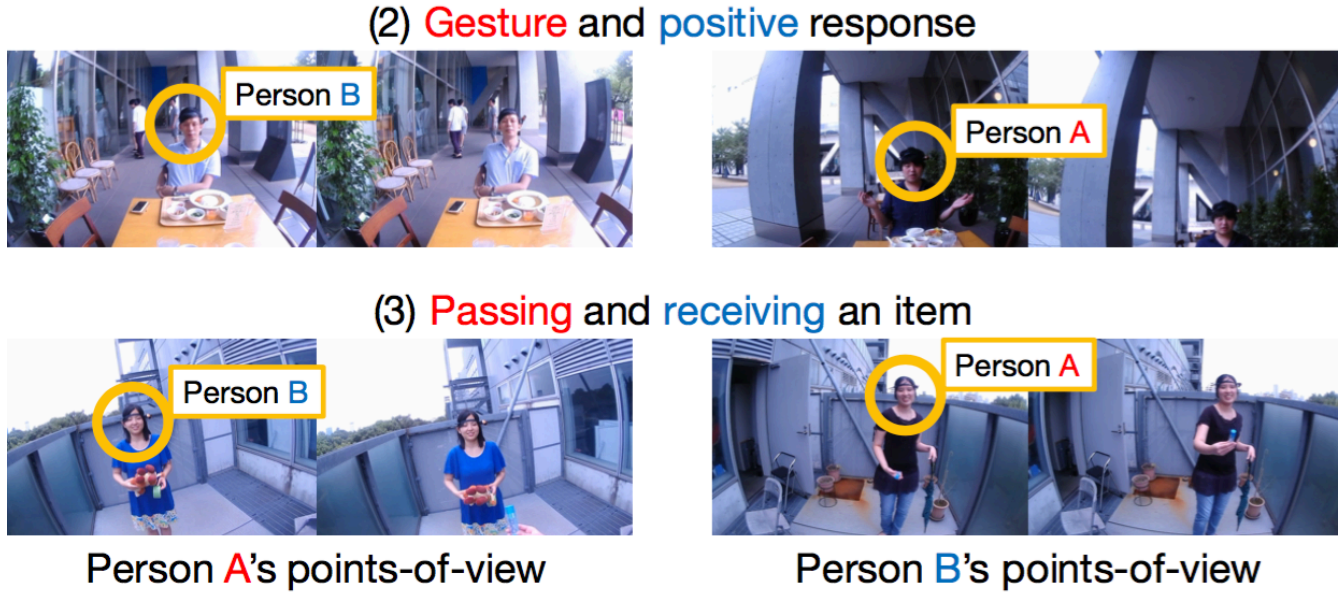




- Classifying if an interaction partner is looking at 'me'
- Learning eye-region appearances to classify eye-contact or not

Micro-Action Recognition

Yonetani, Kitani, and Sato, "Recognizing Micro-Actions and Reactions from Paired Egocentric Videos", CVPR'16



Head actions can appear as large global motion in first-person POV

Hand actions can be observed more clearly in second-person POV

- Recognizing various (re)actions where slight motion is only apparent in videos
- Combining first-person and second-person points-of-view features

Analyzing Interactions – Discussion

- Standard 3rd-person POV action recognition techniques can be used
  - Dense trajectories [Wang+, ICCV'13]
  - Two-stream CNN [Simonyan+, NIPS'14]
- People observed in first-person videos are:
  - Partially occluded
  - Suffer from large global motion, extreme camera poses

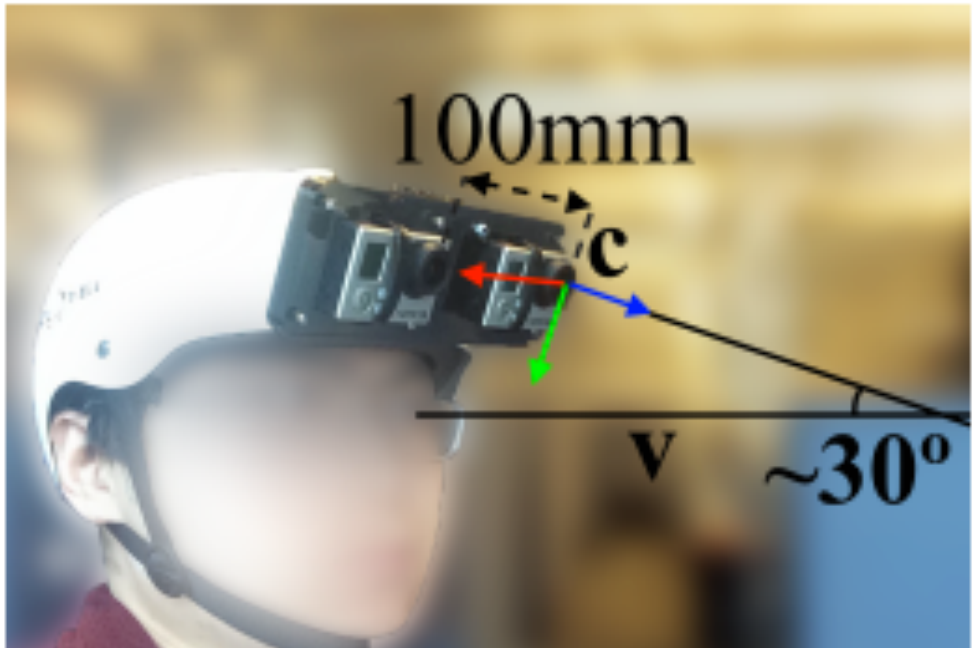
4. Leveraging Ego-Motion

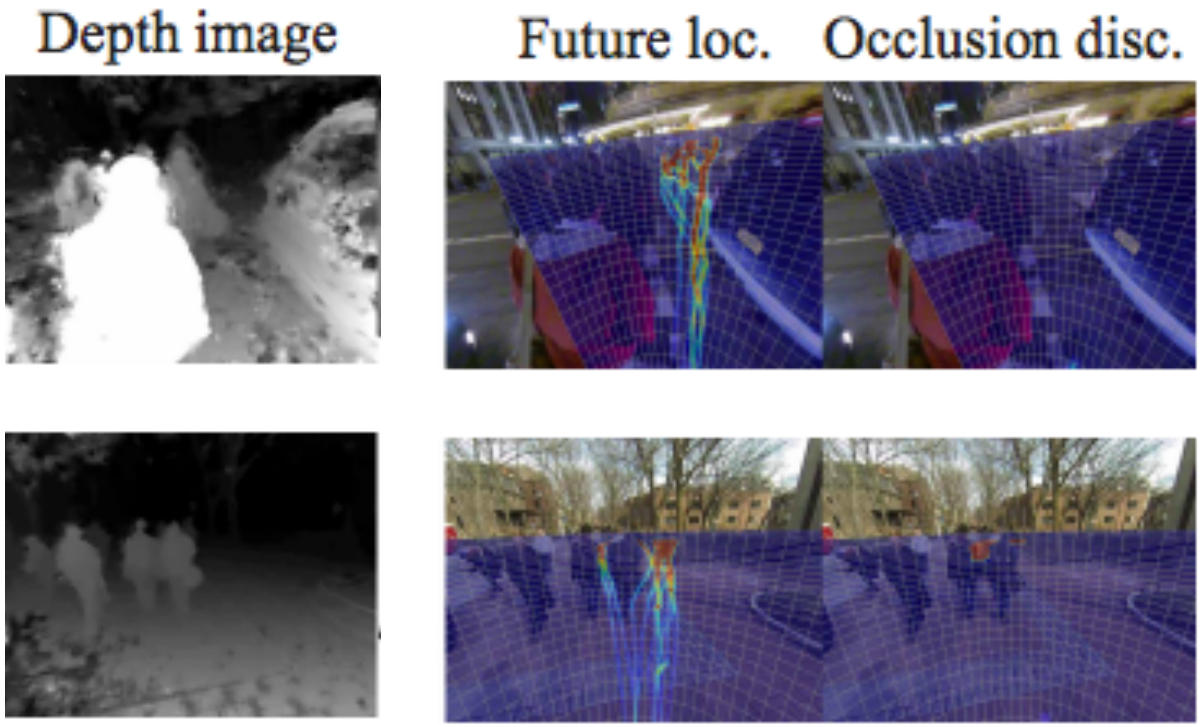
Future Localization

Park, Hwang, Niu, and Shi, "Egocentric Future Localization", CVPR'16

Future Localization

Park, Hwang, Niu, and Shi, "Egocentric Future Localization", CVPR'16

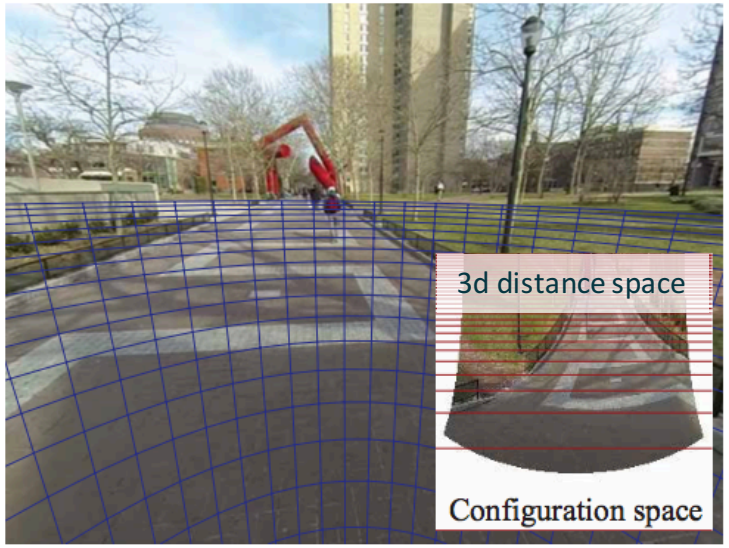




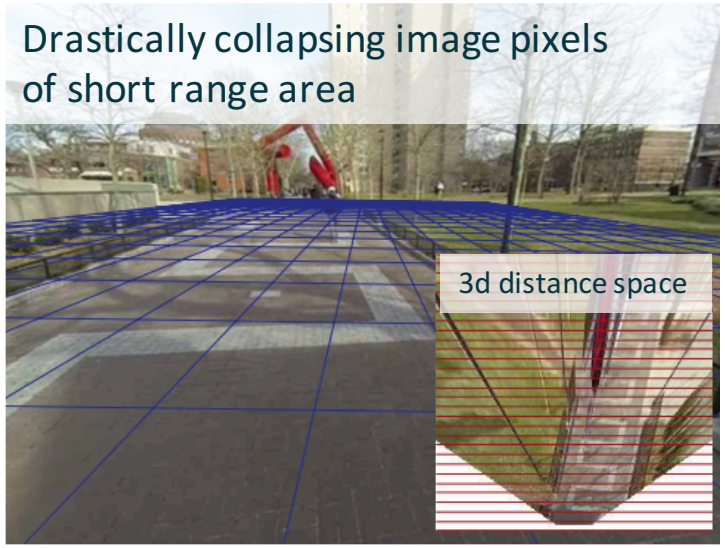
First-person stereo to capture the 3d layout and appearances of scenes

Future Localization

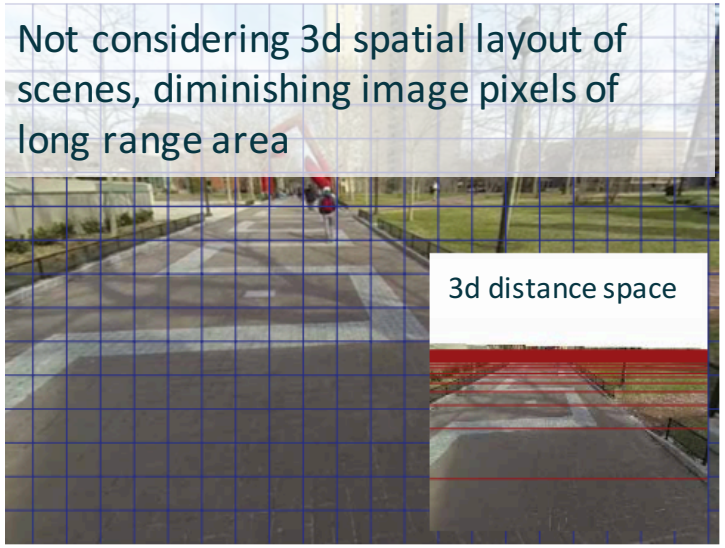
Park, Hwang, Niu, and Shi, "Egocentric Future Localization", CVPR'16



Log-polar coordinate system on ground plane



Cartesian coordinate on ground plane



Cartesian coordinate in image plane

Ego-retinal maps: projecting images onto log-polar coordinate on ground plane to describe the 3d spatial layout and appearance of scenes

Future Localization

Park, Hwang, Niu, and Shi, "Egocentric Future Localization", CVPR'16

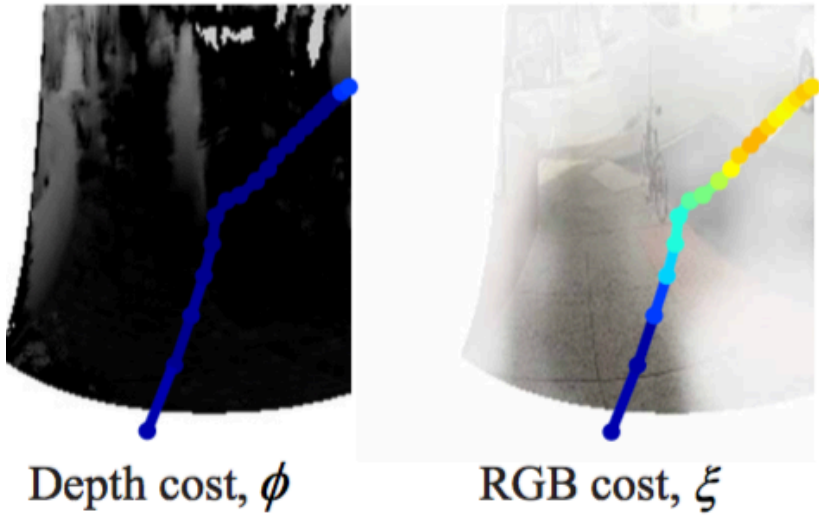
'Walking avoidance' based on RGB & Depth (trained from data)

Trajectory $X=(r_1, \theta_1, \ldots r_M, \theta_M)$

$$\underset{\mathbf{X}}{\text{minimize}} \sum_{i}^{F} \left(\phi\left(r_i, \theta_i\right) + \xi\left(r_i, \theta_i\right)\right) + \lambda \|\mathbf{X} - \mathbf{X}_D^*\|^2$$

$$\text{subject to } \mathbf{X}_D^* = \underset{\{\mathbf{X}_{D_j}\}_{j=1}^{m}}{\text{argmin}} \|\mathbf{X} - \mathbf{X}_{D_j}\|^2 \qquad (2)$$

Set of trajectories in similar scenes



Depth cost, ϕ RGB cost, ξ

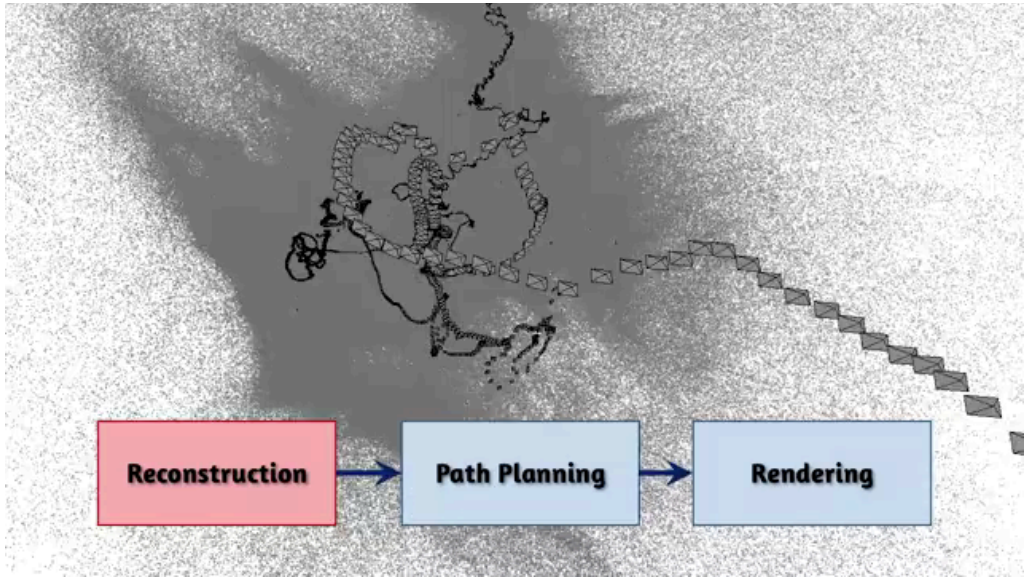
- Retrieve a set of trajectories of similar scenes from DB
- Find a trajectory walkable and similar to the retrieved trajectories

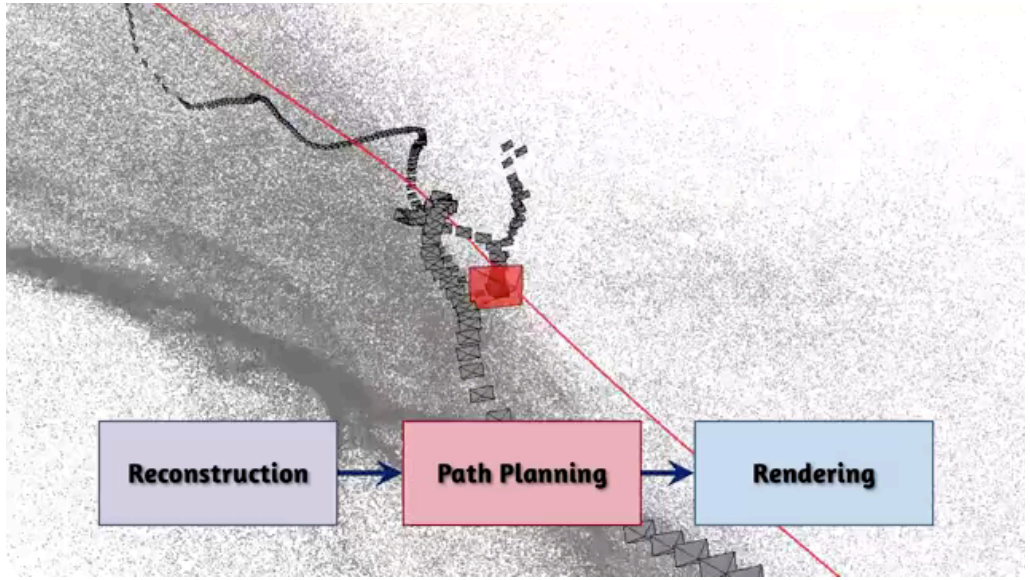
Hyperlapse

Kopf, Kohen, and Szeliski, "First-Person Hyperlapse Videos", SIGGRAPH'14

Hyperlapse

Kopf, Kohen, and Szeliski, "First-Person Hyperlapse Videos", SIGGRAPH'14

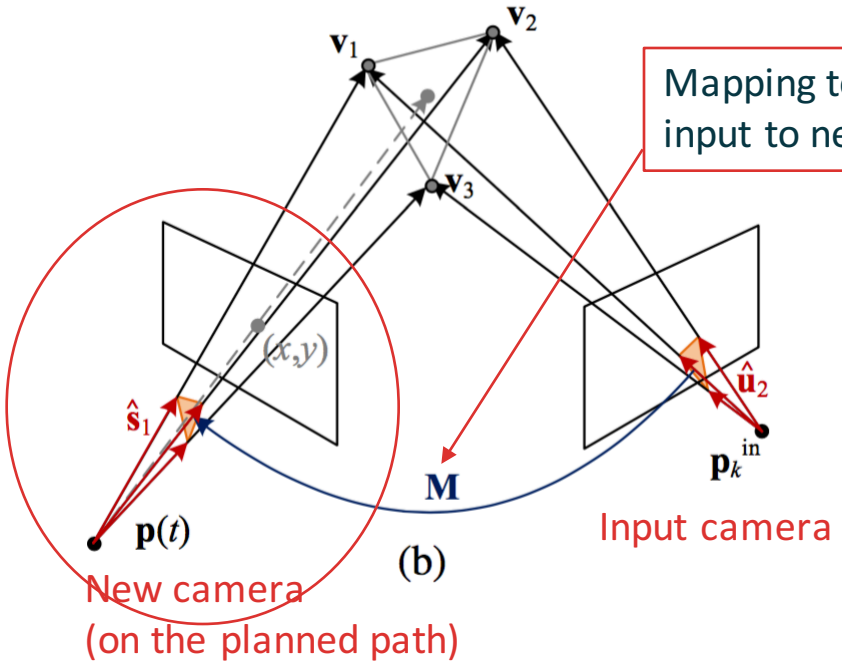




- Estimate camera trajectories via structure-from-motion
- Planning optimal paths (camera positions and orientations) which are 1) no longer than necessary; 2) smooth; 3) near the input cameras; 4) with good rendering quality

Hyperlapse

Kopf, Kohen, and Szeliski, "First-Person Hyperlapse Videos", SIGGRAPH'14



-> `stretchness' $\varphi_k^{TS3} = 1 - \frac{\min_i \sigma_i^{\mathbf{M}}}{\max_i \sigma_i^{\mathbf{M}}}$,

Penalty on rendering quality at frame t:
Stretchness given an optimal set of input cameras for rendering

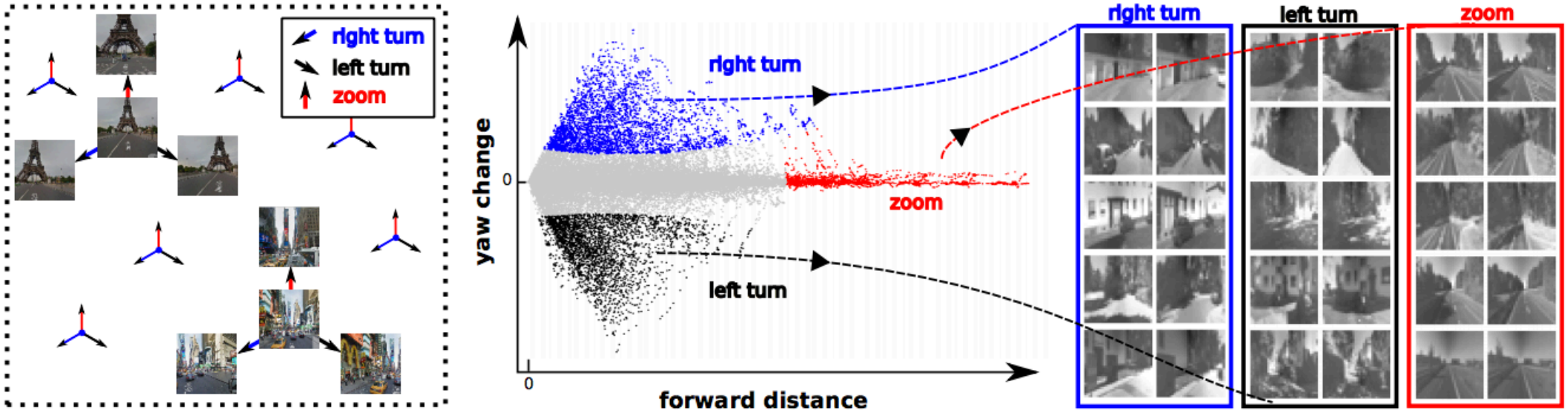
$$\Phi(t) = \iint \min_k \varphi_k(x,y,t)\, \mathrm{d}x\, \mathrm{d}y.$$

Stretchness minimized w.r.t input cameras

- Estimating rendering quality for a new camera based on texture stretches
- Using rendering qualities for optimizing new camera poses

Learning Image Representations of Ego-Motion

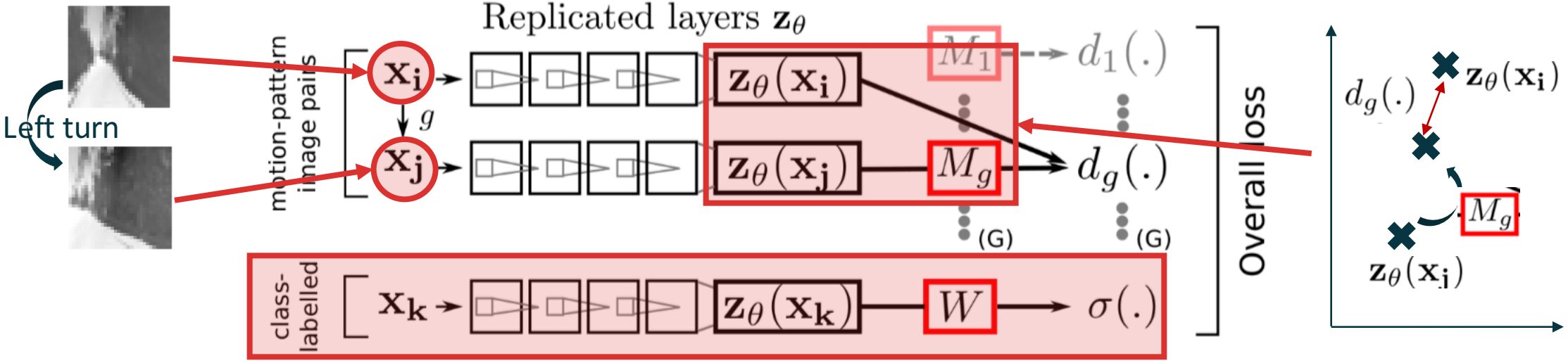
Jayaraman and Grauman, "Learning Image Representations Tied to Ego-Motion", ICCV'15



- Learning the connection between 'how I move' ⇔ 'how surroundings change'
- Mapping image pairs with similar ego-motion into feature pairs with similar distance

Learning Image Representations of Ego-Motion

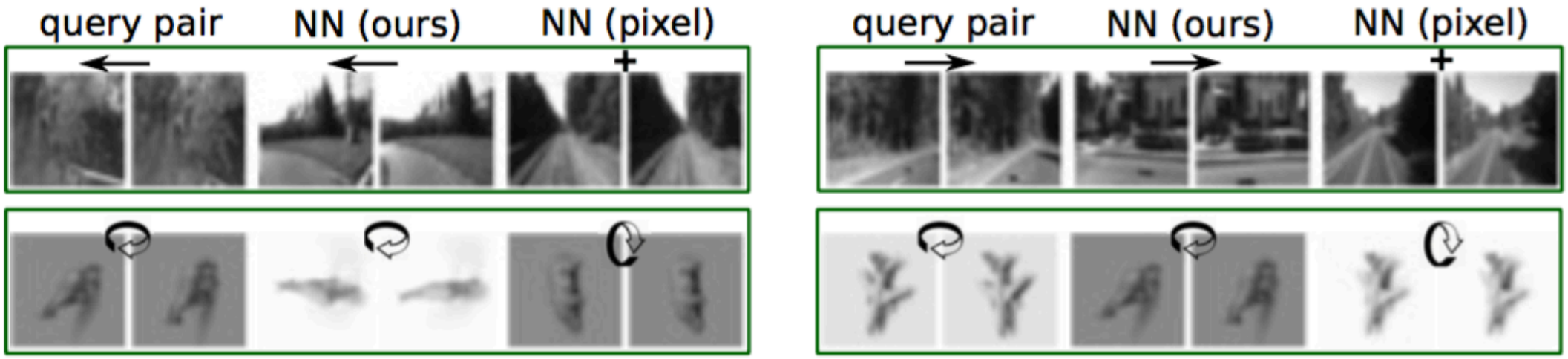
Jayaraman and Grauman, "Learning Image Representations Tied to Ego-Motion", ICCV'15



- Learning feature representation *z* and transformation M_g w.r.t. ego-motion *g*
- Regularizing another classifier for visual recognition (e.g., scene classification)

Learning Image Representations of Ego-Motion

Jayaraman and Grauman, "Learning Image Representations Tied to Ego-Motion", ICCV'15



- Nearest neighbors in the learned space have the same ego-motion
- Up to 30% accuracy increase over state of the art scene recognition

Leveraging Ego-Motion – Other Work

- Activity recognition using head motion [Kitani+, CVPR'11; Poleg+, CVPR'14]
  - Learning optical flows or global motion to recognize actions
- Wearer identification [Poleg+, ACCV'14; Yonetani+, CVPR'15]
  - Matching global motion of target videos and local motions in observer videos
- Wearer recognition [Hoshen+, CVPR'16]
  - Learning head motion as a gait for authenticating people
- Ego-motion based summarization [Poleg+, CVPR'15]
  - Subsampling frames so that FOE moves smoothly

Leveraging Ego-Motion – Discussion

- Relatively easier to conduct large-scale data collection
  - Not necessary to ask subjects to do complex tasks
- Beyond typical action recognition tasks
  - Much work has been done on ego-motion-based action recognition

5. Recording Everyday Life

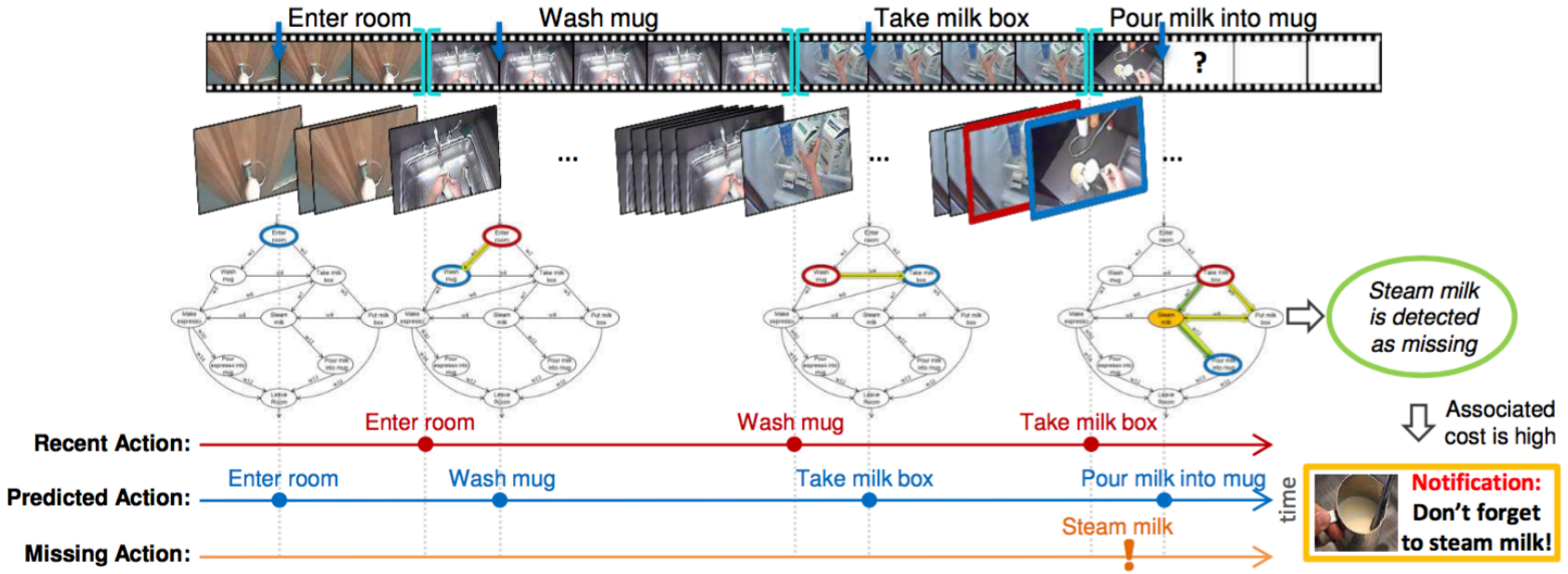
Learning Action Maps

Rhinehart and Kitani, "Learning Action Maps of Large Environments via First-Person Vision", CVPR'16

- Learning to predict what a wearer can do at a certain place
- Estimating dense 'action-maps' from sparse observations of actions

Detecting Missing Actions

Soran, Farhadi, and Shapiro, "Generating Notifications for Missing Actions", ICCV'15



- Notifying of actions that are significant but not performed
- Learning a series of actions during a certain task (e.g., making coffee)

Large-Scale FPV

Singh, Fatahalian, and Efros, "KrishnaCam: Using a Longitudinal, Single-Person, Egocentric Dataset for Scene Understanding Tasks", WACV'16

Walking in urban/campus/ residential areas, waiting at intersections and for bus

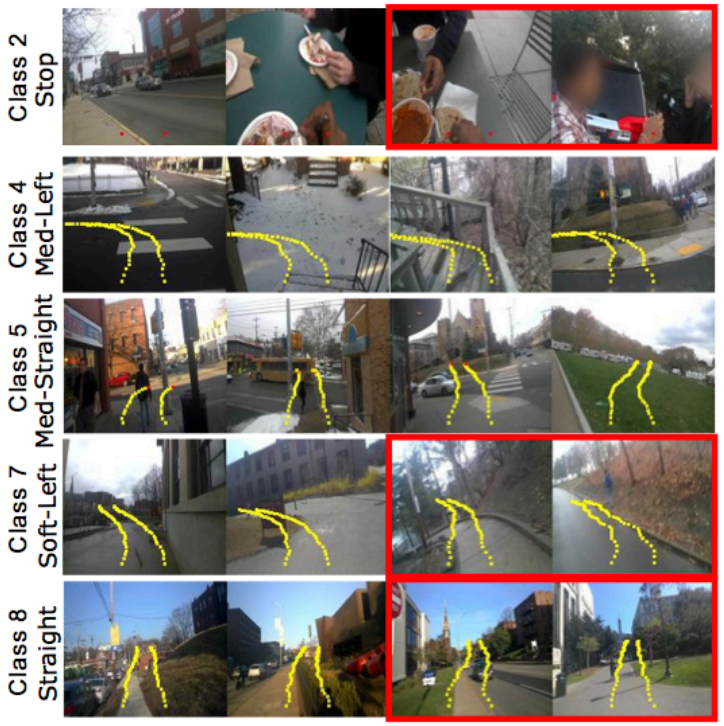
Shopping, eating

Evening and night recording

Activities in parks, at events

Seasonal change

Socializing with friends



9 months, 70 hours (7.6 million frame) video of a single person

Summary

Summary

- First person vision can be used for …
  1. Measuring attention
  2. Observing hands
  3. Analyzing interactions
  4. Leveraging ego-motion
  5. Recording Everyday Life

Future Directions

- Large scale FPV
  - How can we scale-up the area of first-person sensing?
- Privacy preserving FPV
  - How can we preserve the privacy of wearers/subjects?
- Strong applications
  - Which domains should FPV be applied to?

Large-Scale FPV

Dataset	# videos	hours	Scenario
GT Disney [Fathi+, CVPR'12]	113	50	Amusement park
CMU MMAC [De la Torre+, FPV'09]	171	17	Kitchen
UT Life [Lee+, CVPR'12]	4	17	Life Log
GTEA Gaze+ [Fathi+, ECCV'12]	30	5	Kitchen
UTokyo Paired [Yonetani+, CVPR'16]	1226	0.5	Conversation

Privacy-Preserving FPV

- Screen detection [Koreyam+, CHI'16]
- Sensitive space detection [Templeman+, NDSS'14]
- Privacy-preserving action recognition [Ryoo+, arXiv]

Downsizing videos to anonymize people while keeping high recognition accuracy

Applications

- Blind navigation [Leung+, CVPRW'14; Tang+, ISWC'14]
- Cooperative work [Fussell+,CHI'03;Kasahara+,CHI'16;Higuchi+,CHI'16]
- Sport analytics, rapid diagnosis for autism, life-logging

- We need very strong applications that make people willing to use wearable cameras!

Thank you!

- Acknowledgments:
  - JST CREST 「人間と調和した創造的協働を実現する知的情報処理システムの構築: 集合視による注視・行動解析に基づくライフイノベーション創出」

Human-Machine Harmonious Collaboration

JST CREST Project
COLLECTIVE VISUAL SENSING

- Contact information:
  - 米谷　竜（yonetani@iis.u-tokyo.ac.jp）