

# Human Kinematics II (Actions)

Kris Kitani

Assistant Research Professor

**Carnegie Mellon University**  
The Robotics Institute

# What can FPV tell us about **my actions?**

What am I  
doing?

What can  
I do?

What are you  
doing to me?

# What can FPV tell us about **my actions?**

What am I  
doing?

What can  
I do?

What are you  
doing to me?

Understand interactions with **things**

# What can FPV tell us about **my actions?**

What am I  
doing?

What can  
I do?

What are you  
doing to me?

Understand interactions with **people**



## Understand interactions with **things**



1. Recognizing activities  
CVPR 2016

2. Learning scene functionality  
CVPR 2016

## Understand interactions with **people**



3. Recognizing social interactions  
CVPR 2016

## Understand interactions with **things**



**1. Recognizing activities**  
**CVPR 2016**

**2. Learning scene functionality**  
**CVPR 2016**

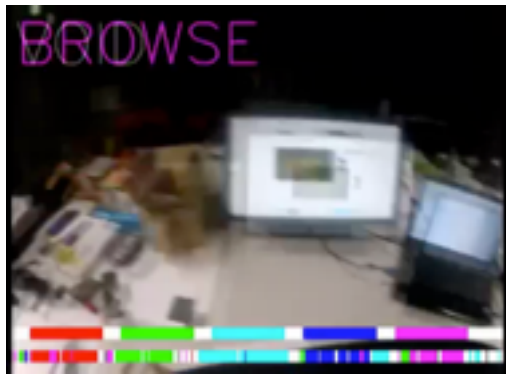
## Understand interactions with **people**



**3. Recognizing social interactions**  
**CVPR 2016**

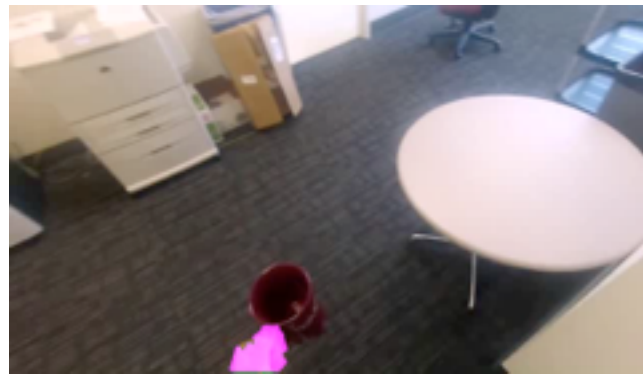
# Lessons Learned

Ego-motion is important



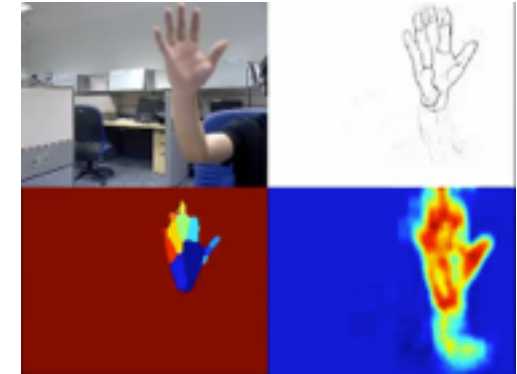
[**Kitani**, Okabe, Sato CVPR 2011]  
[Ogaki, **Kitani**, Sato EGOV 2012]

Hand detection is important



[Li, **Kitani** CVPR 2013]  
[Li, **Kitani** ICCV 2013]

Hand parts are important



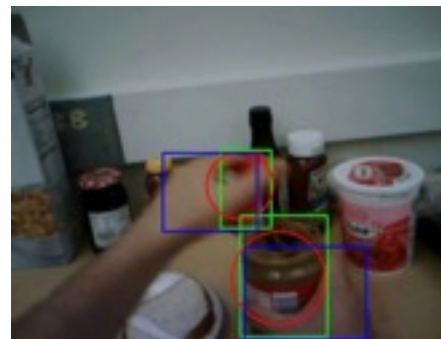
[Asaran, Teney, **Kitani** IROS 2015]  
[Cai, **Kitani**, Sato ICRA 2015]

Hand motion is important



[Ishihara, **Kitani**, Ma, Takagi, Asakawa ICIP 2015]  
[Cai, **Kitani**, Sato RSS 2016]

Object appearance is important



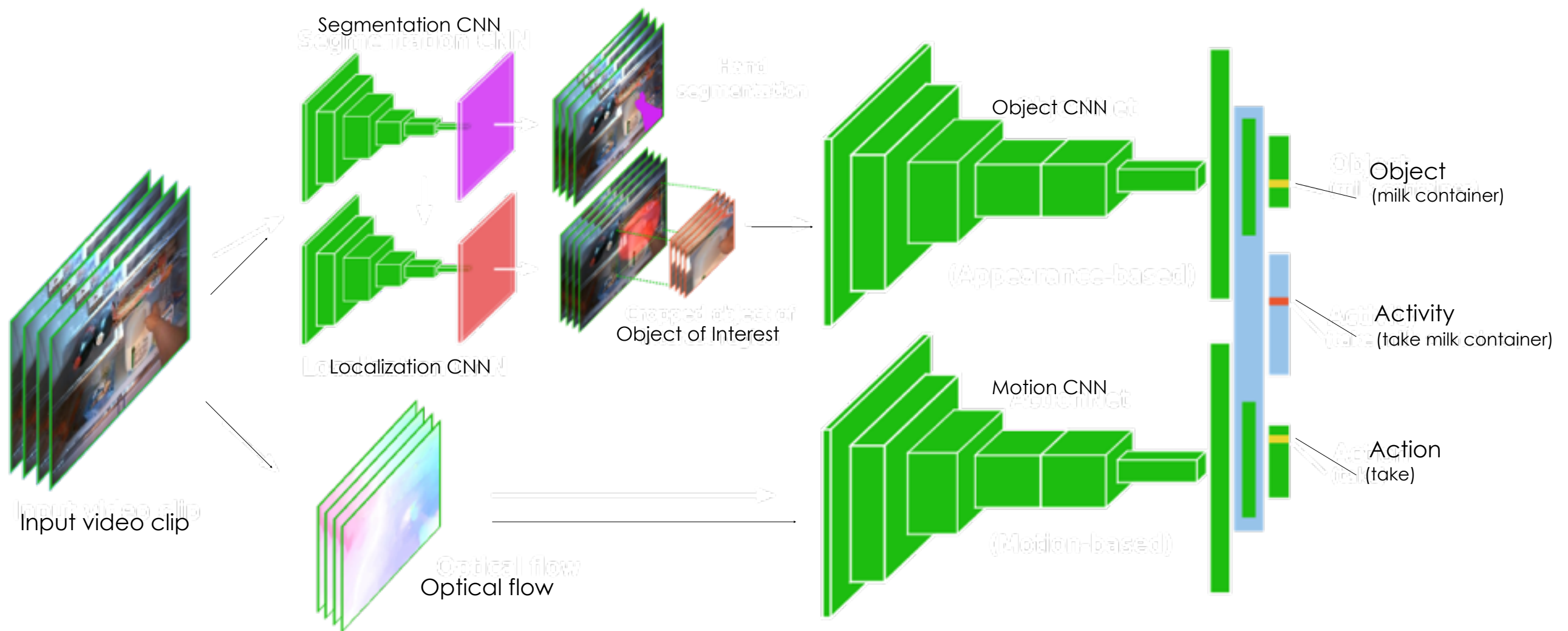
[**Kitani**, Okabe, Sugimoto, Sato ECCVW 2008]  
[Cai, **Kitani**, Sato RSS 2016]

Hand shape implies object region



[Huang, Ma, Ma **Kitani** CVPR 2015]  
[Cai, **Kitani**, Sato RSS 2016]

# Integrate lessons learned under one (deep) framework

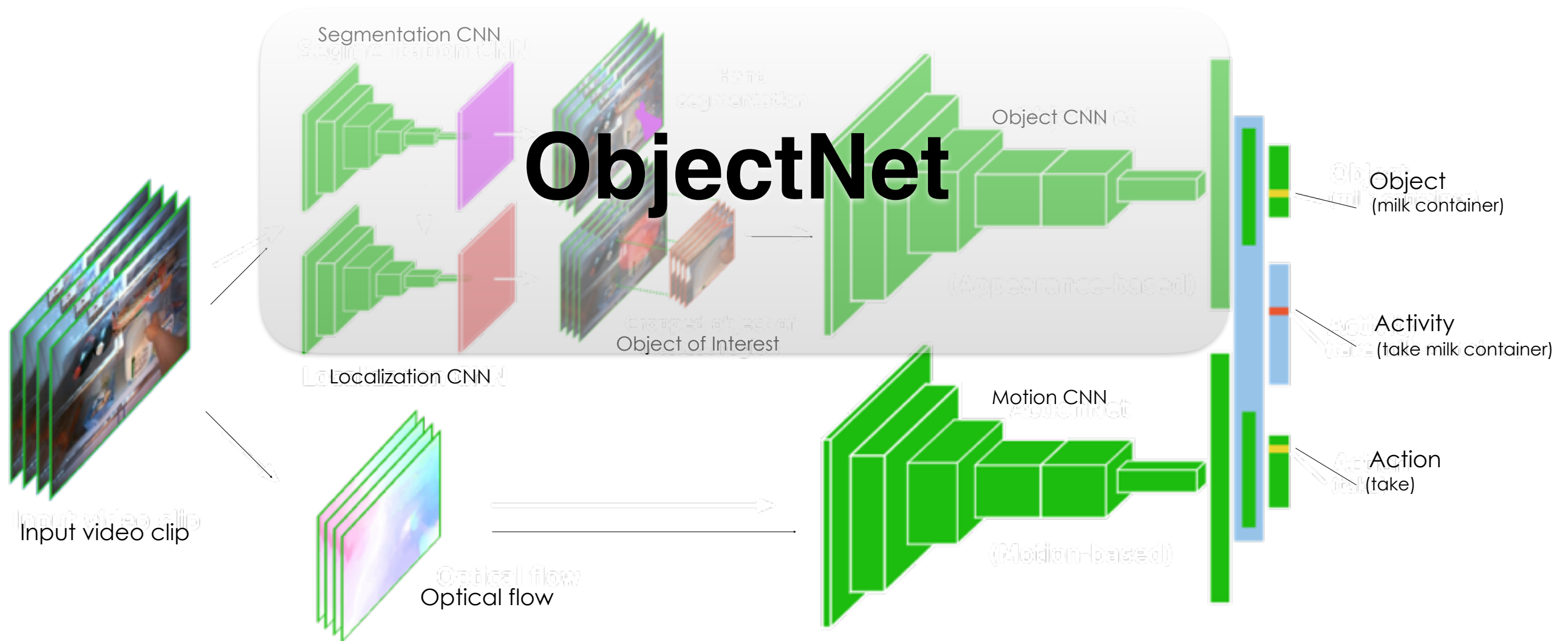


Minghuang Ma, Haoqi Fan, Kris M. Kitani.

Going Deeper into First-Person Activity Recognition.

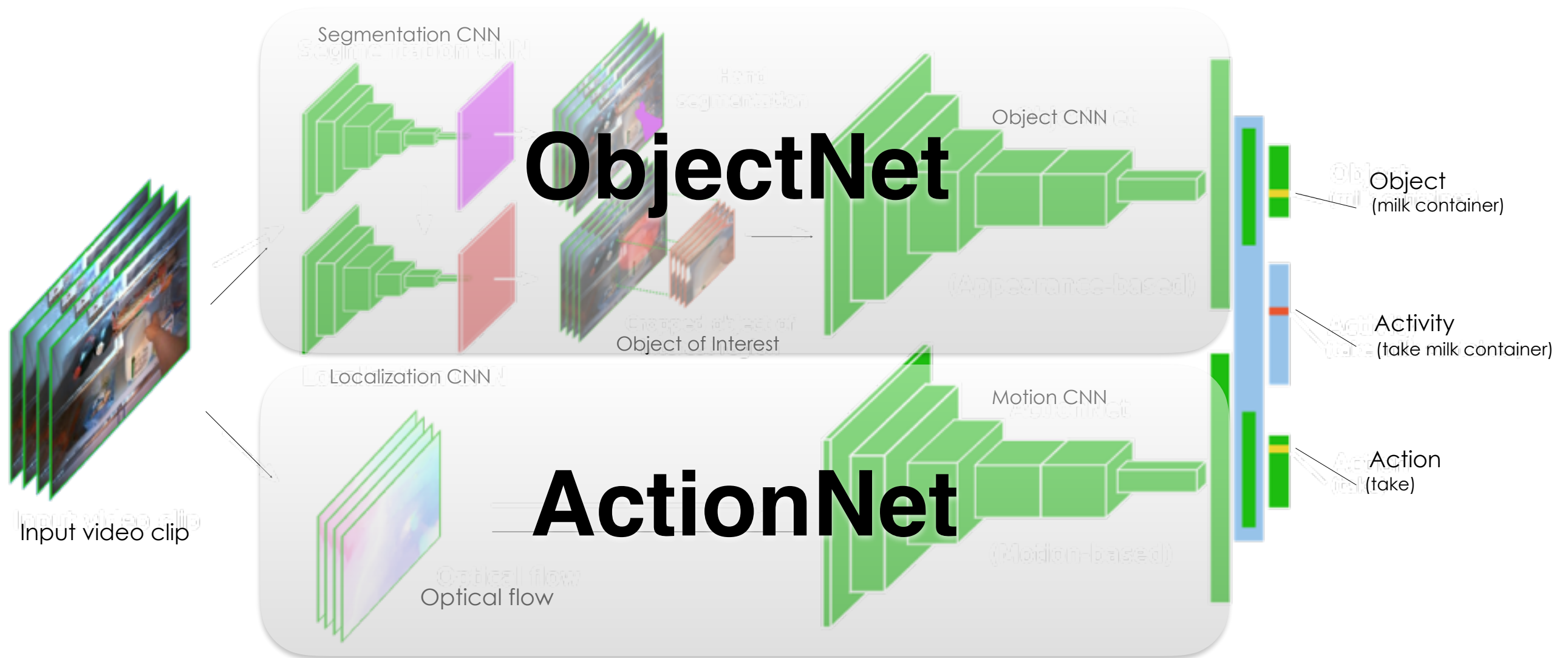
Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

Integrate lessons learned under one framework

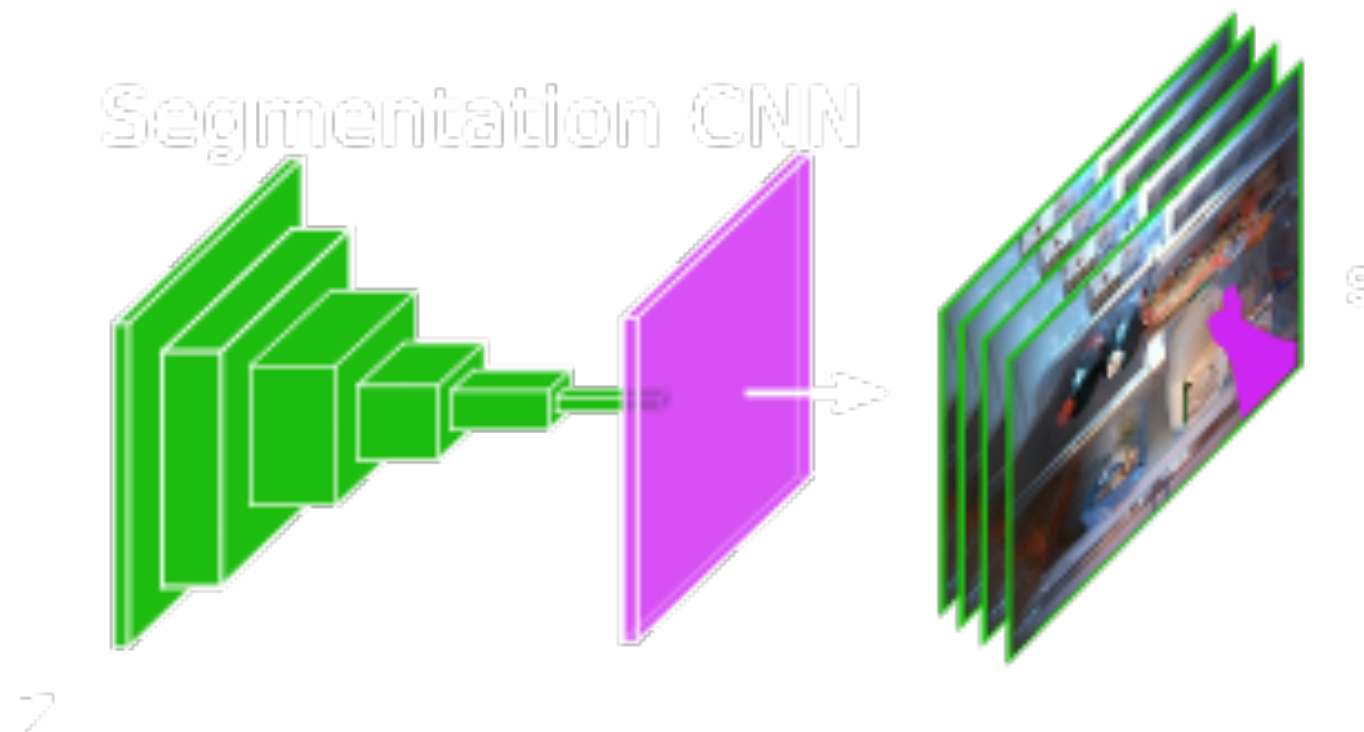




Integrate lessons learned under one framework

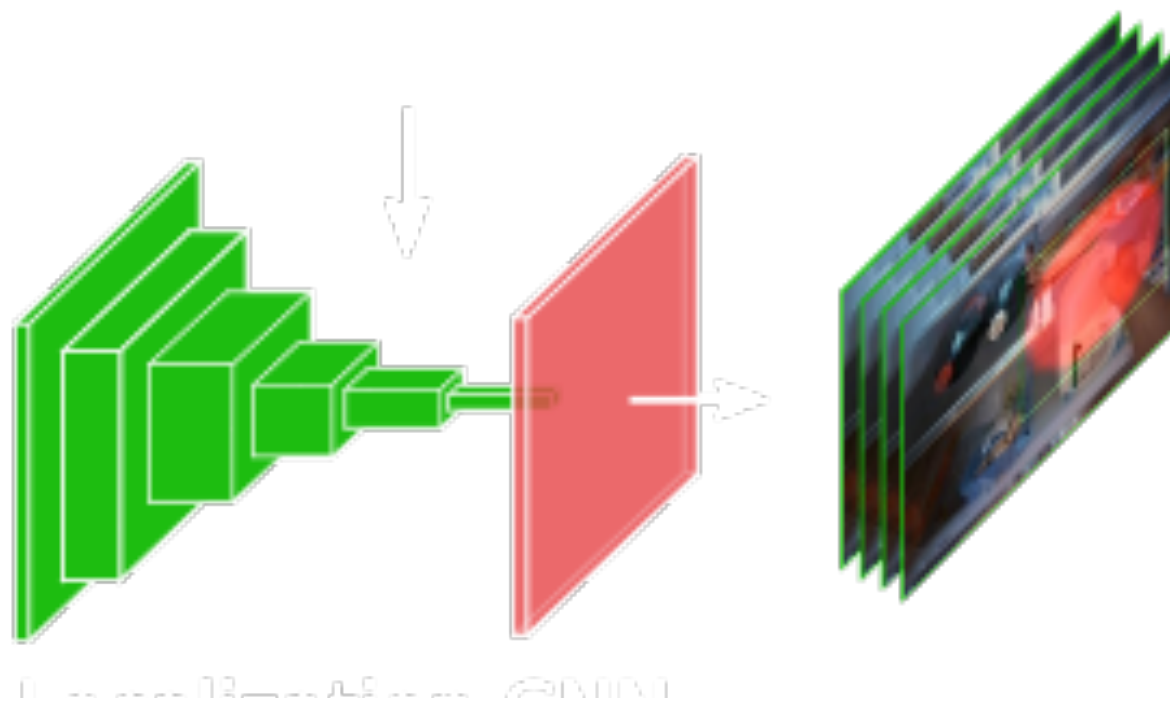


Hands should be used to identify important objects...



**Step 1:** Learn a hand region detector  
(use dataset of binary hand masks)

Fully Convolutional Network (FCN) [Long et al. 2014]



**Step 2:** Re-train output layer of the  
FCN to detect object region  
(use Gaussian heatmap centered  
on object)

By learning hand appearance first, we were able to detect object regions better

# Object Region Detection Results



GTEA dataset



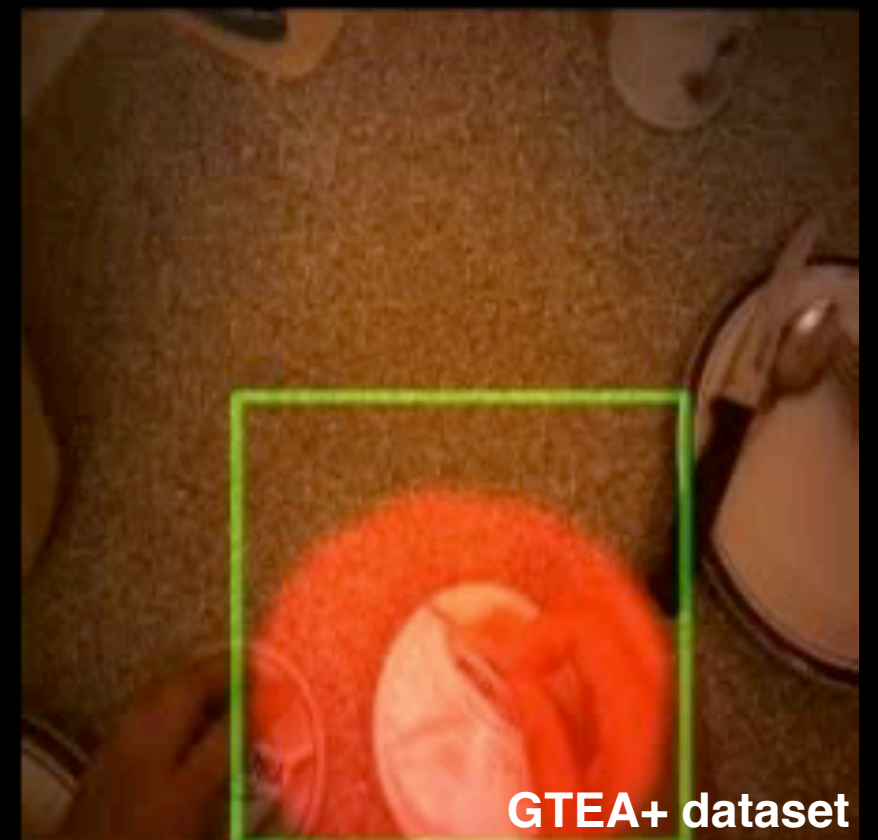
GTEA+ dataset

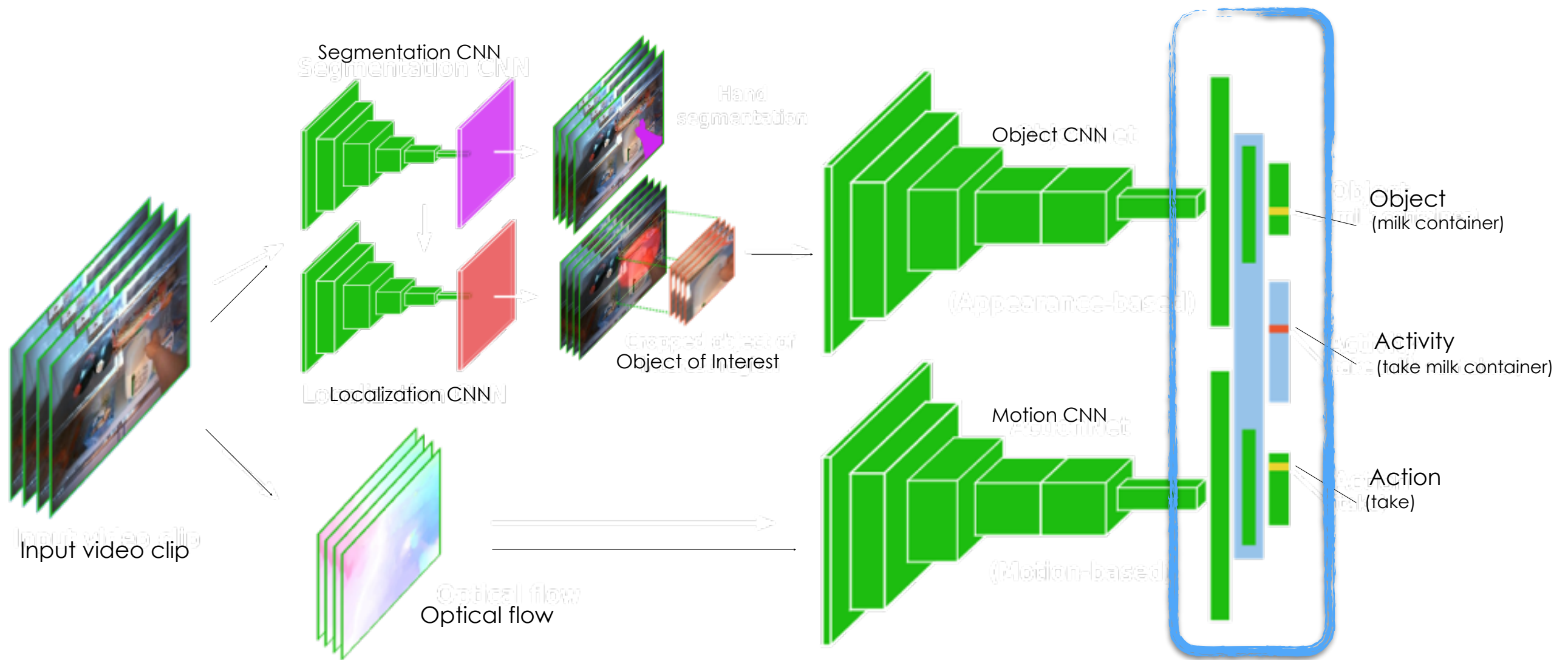


# Hand Region Detection



# Object Region Detection

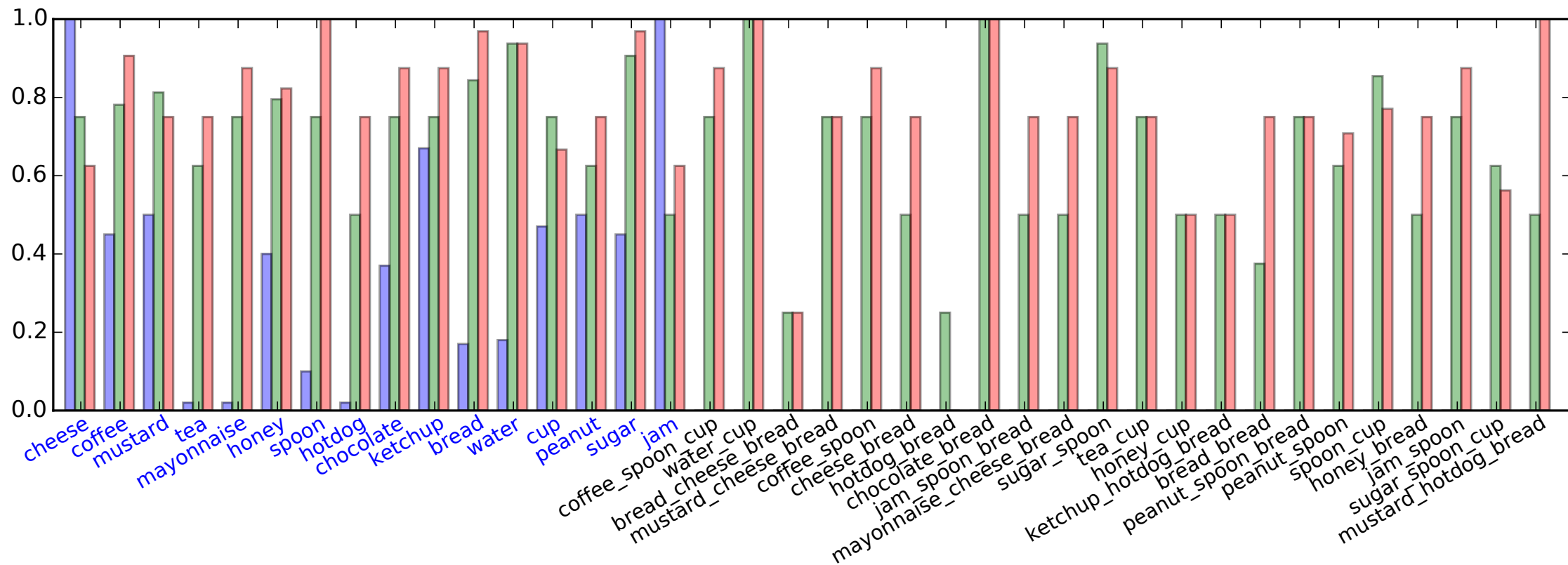




**Late fusion** with twin stream network,  
 fined-tuned for multi-task recognition  
 (object, action and activity)

Object Recognition	GTEA(71)	Gaze(40)	Gaze+(44)
Fathi <i>et al.</i> [9]	61.36	N/A	N/A
Object CNN	67.74	38.05	61.87
Joint training (Ours)	<b>76.15</b>	<b>55.55</b>	<b>74.34</b>

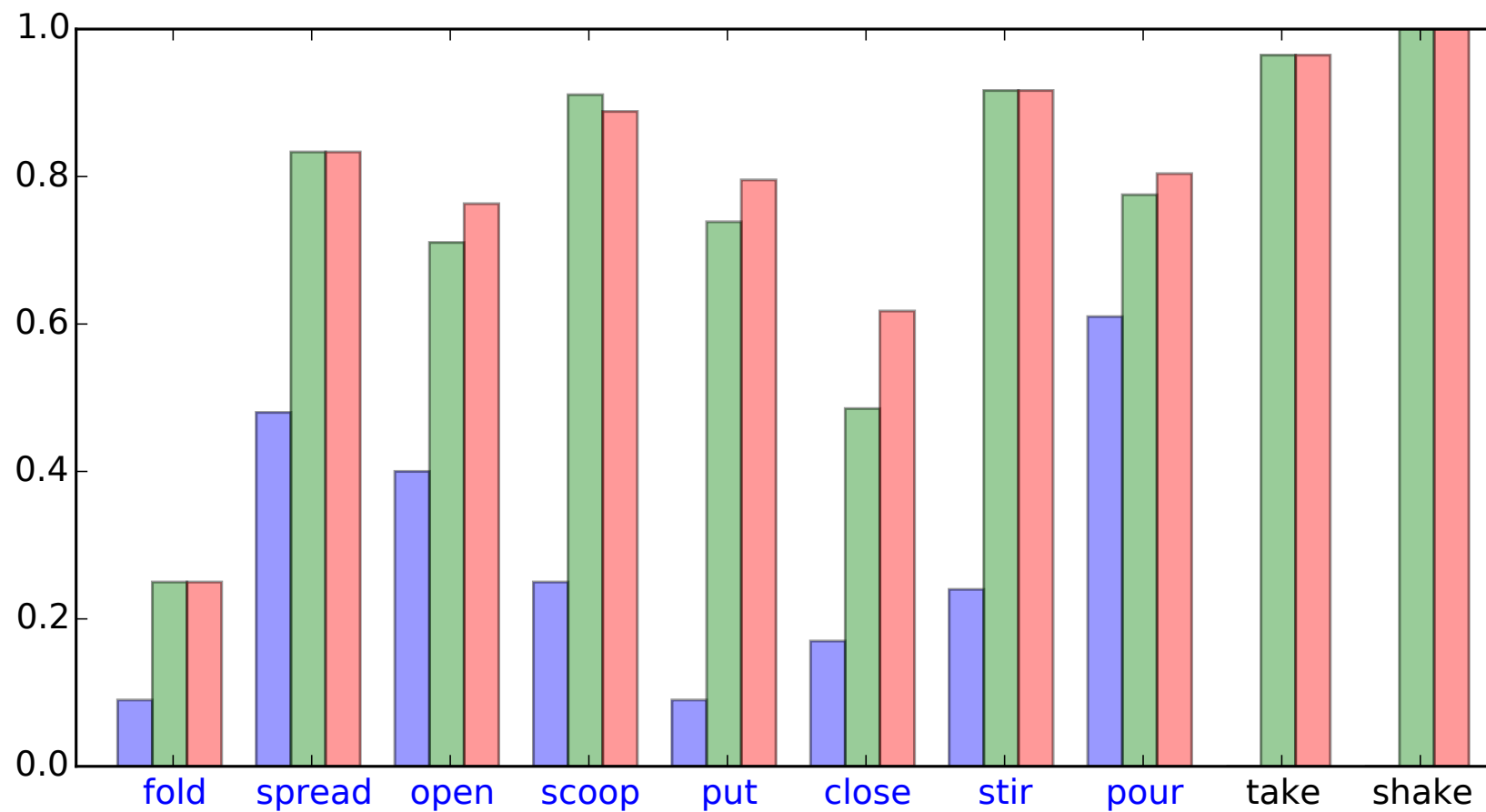
(a) Average **object recognition** accuracy.



(b) **Object recognition** accuracy for each class (GTEA).

Method & dataset	GTEA(71)	Gaze(40)	Gaze+(44)
Fathi <i>et al.</i> [5]	47.70	N/A	N/A
Motion CNN	75.85	33.65	62.62
Joint training	<b>78.33</b>	<b>36.27</b>	<b>65.05</b>

(a) Average **action recognition** accuracy.

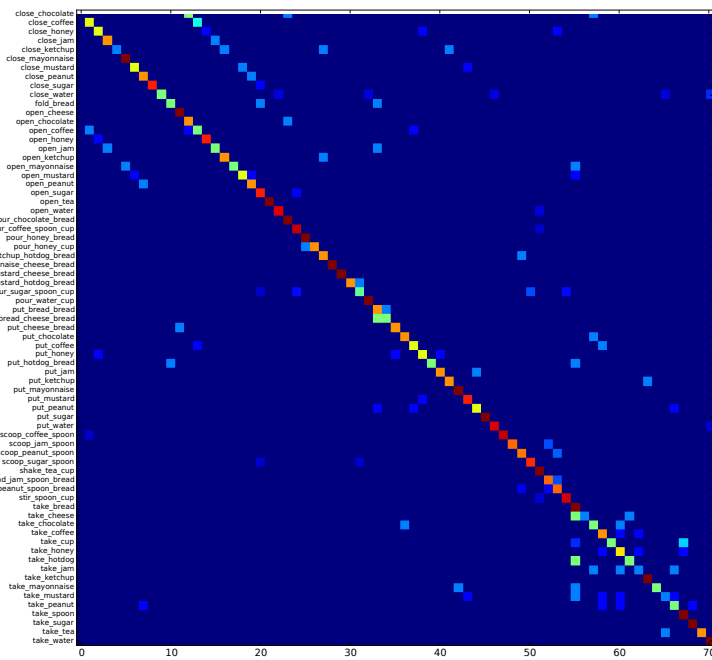


(b) **Action recognition** accuracy for each class (GTEA).

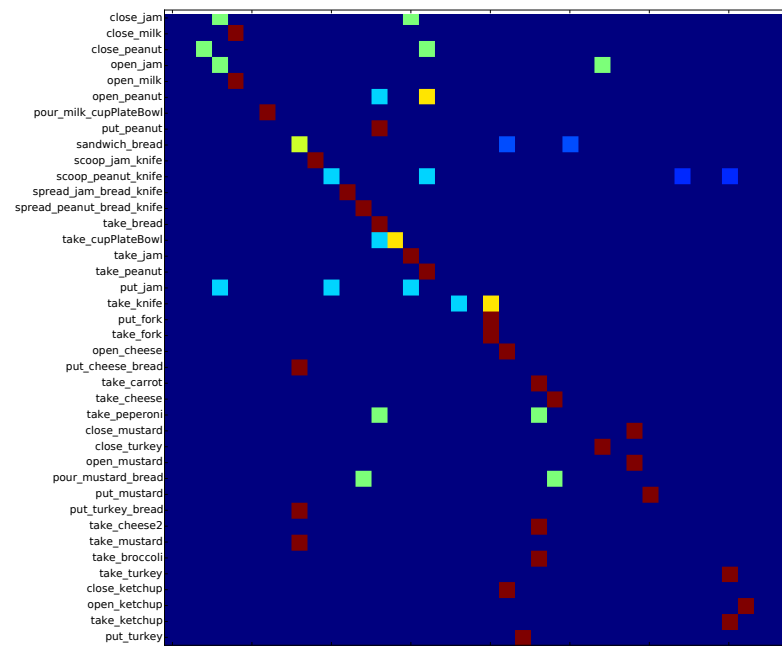


	Methods	GTEA(61)*	GTEA(61)	GTEA(71)	Gaze(25)*	Gaze(40)*	Gaze+(44)
Li <i>et al.</i> [20]	O+M+E+H	61.10	59.10	59.20	53.20	35.70	60.50
	O+M+E+G	N/A	N/A	N/A	60.90	39.60	60.30
	O+E+H	66.80	64.00	62.10	51.10	35.10	57.40
S. & Z.[30]	temporal-cnn	34.30	30.92	30.33	38.76	22.01	44.45
	spatial-cnn	53.77	41.13	40.16	30.84	18.46	45.97
	temporal+spatial-svm	46.51	35.69	35.81	25.94	22.18	43.23
	temporal+spatial-joint	57.64	51.58	49.65	44.29	34.70	58.77
Ours	object-cnn	60.02	56.49	50.35	47.09	35.56	46.38
	motion+object-svm	53.01	50.45	47.07	28.42	16.00	34.75
	motion+object-joint	<b>75.08</b>	<b>73.02</b>	<b>73.24</b>	<b>62.40</b>	<b>43.42</b>	<b>66.40</b>

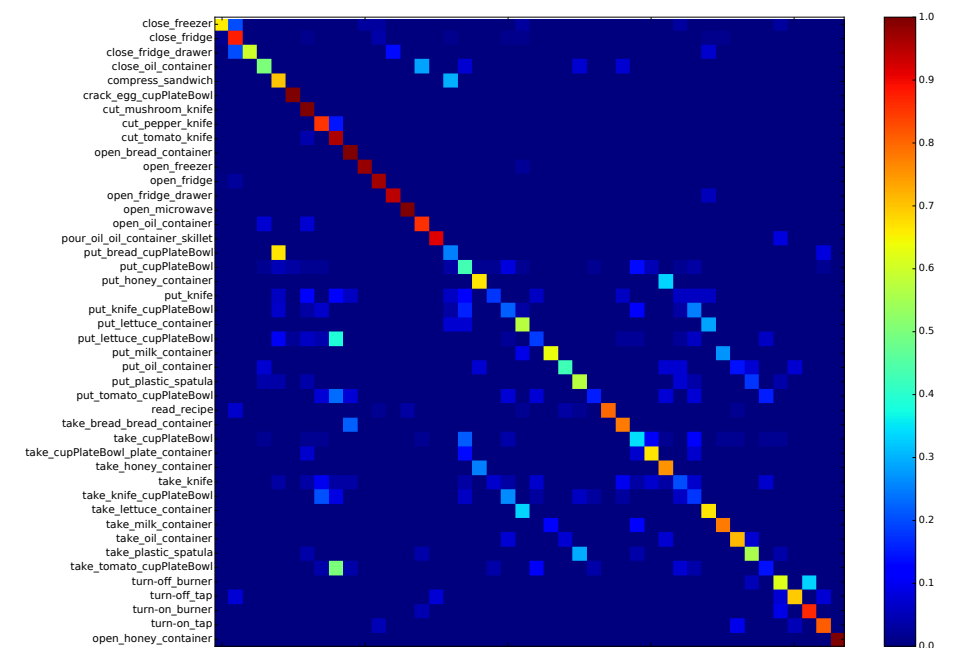
(a) **Activity recognition** results



(b) GTEA



(c) Gaze



(d) Gaze+

# What is the ObjectNet learning?

A neuron units in the conv-5 layer



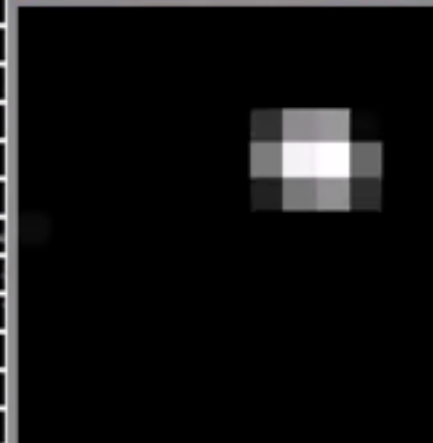
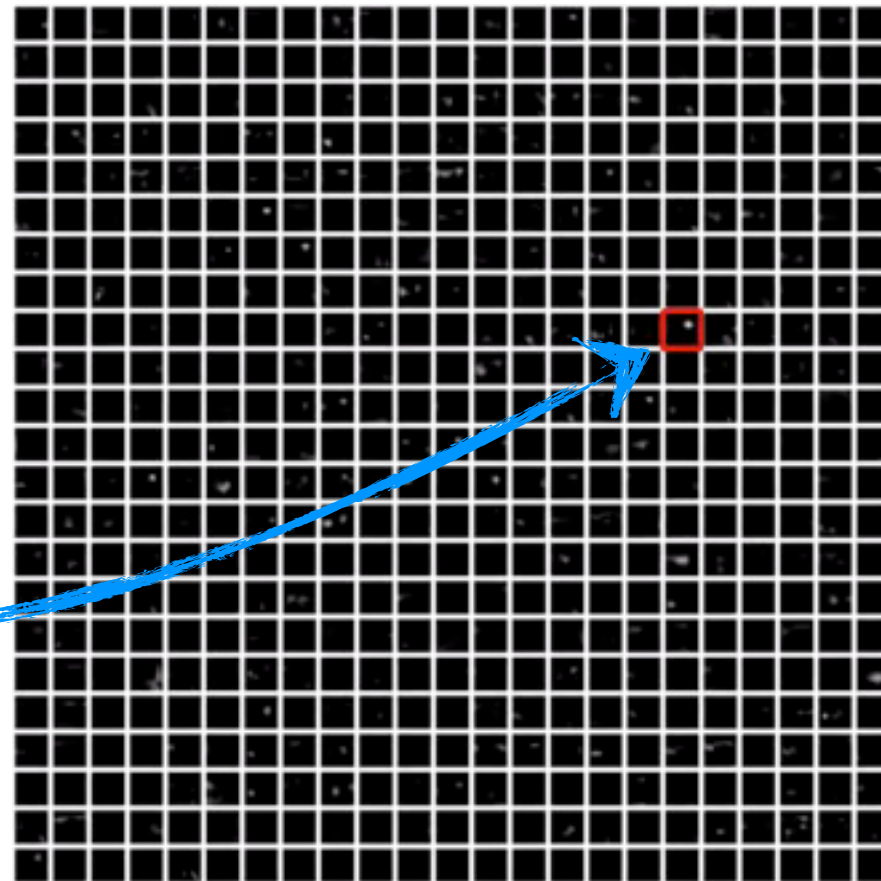
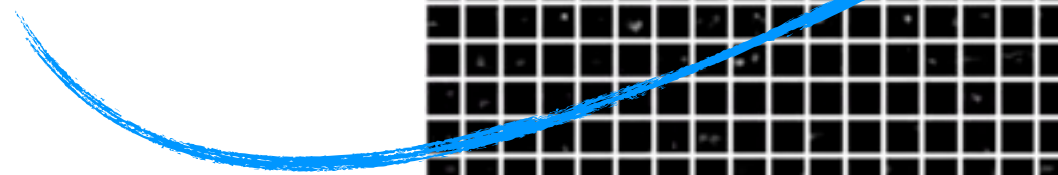
(a) blue

(b) water bottle

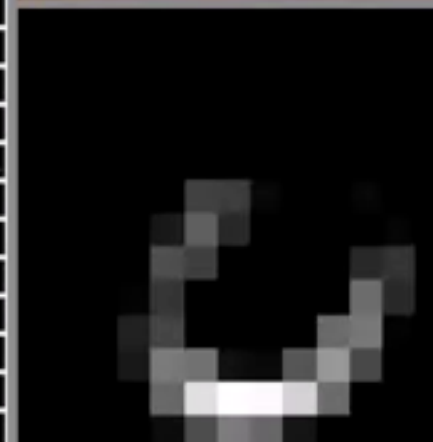
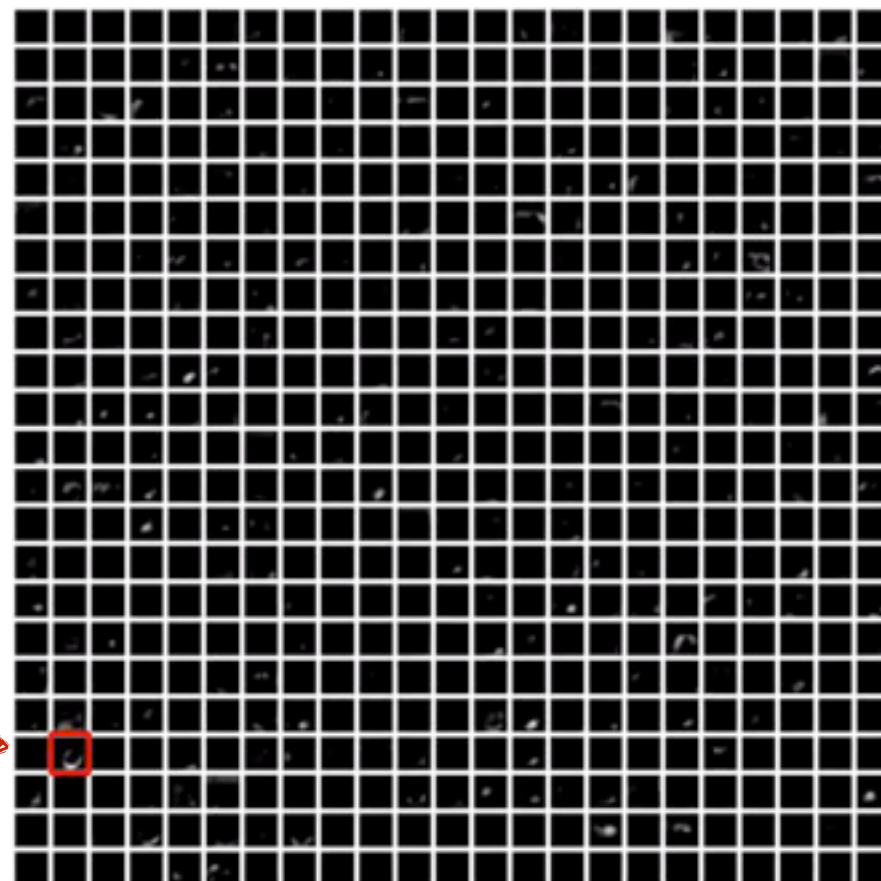
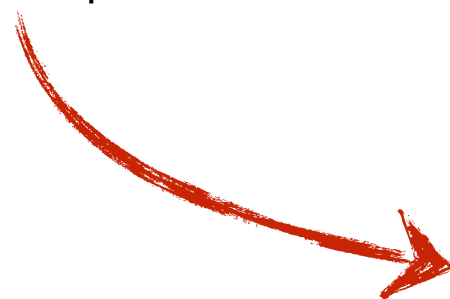
(c) round edge

(d) red cup

(a) blue color neuron



(b) red cup neuron



## Understand interactions with **things**



1. Recognizing activities  
CVPR 2016

2. Learning scene functionality  
CVPR 2016

## Understand interactions with **people**



3. Recognizing social interactions  
CVPR 2016



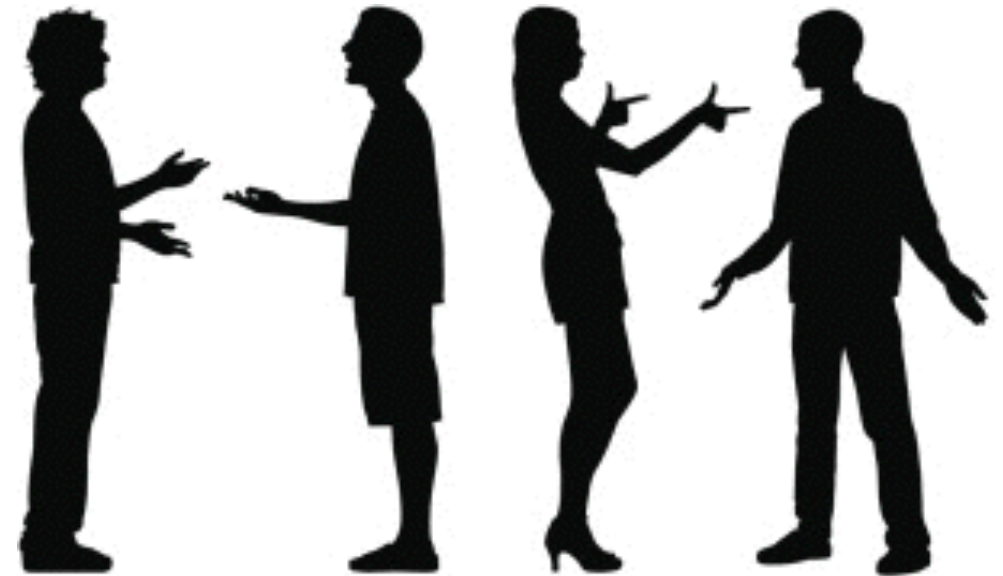
## Understand interactions with **things**



**1. Recognizing activities**  
**CVPR 2016**

**2. Learning scene functionality**  
**CVPR 2016**

## Understand interactions with **people**



**3. Recognizing social interactions**  
**CVPR 2016**

When we **observe** a scene...



... we know **how we can act** in that environment

When we **observe** a scene...



... we know **how we can act** in that environment



When we **observe** a scene...



... we know **how we can act** in that environment

When we **observe** a scene...



... we know **how we can act** in that environment

Can we teach a computer  
to  
**understand scene functionality?**

Nicholas Rhinehart, Kris M. Kitani.

Learning Action Maps of Large Environments via First-Person Vision.  
Conference on Computer Vision and Pattern Recognition (CVPR), 2016.



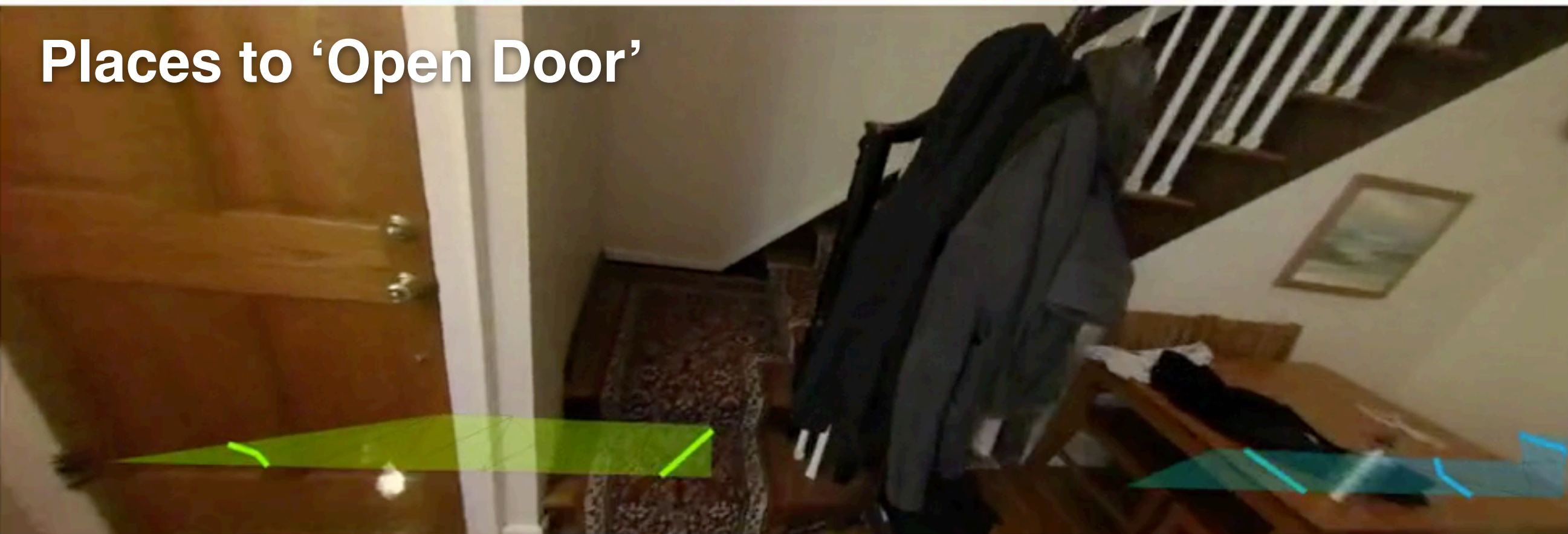


**Input: Captured visual experience**

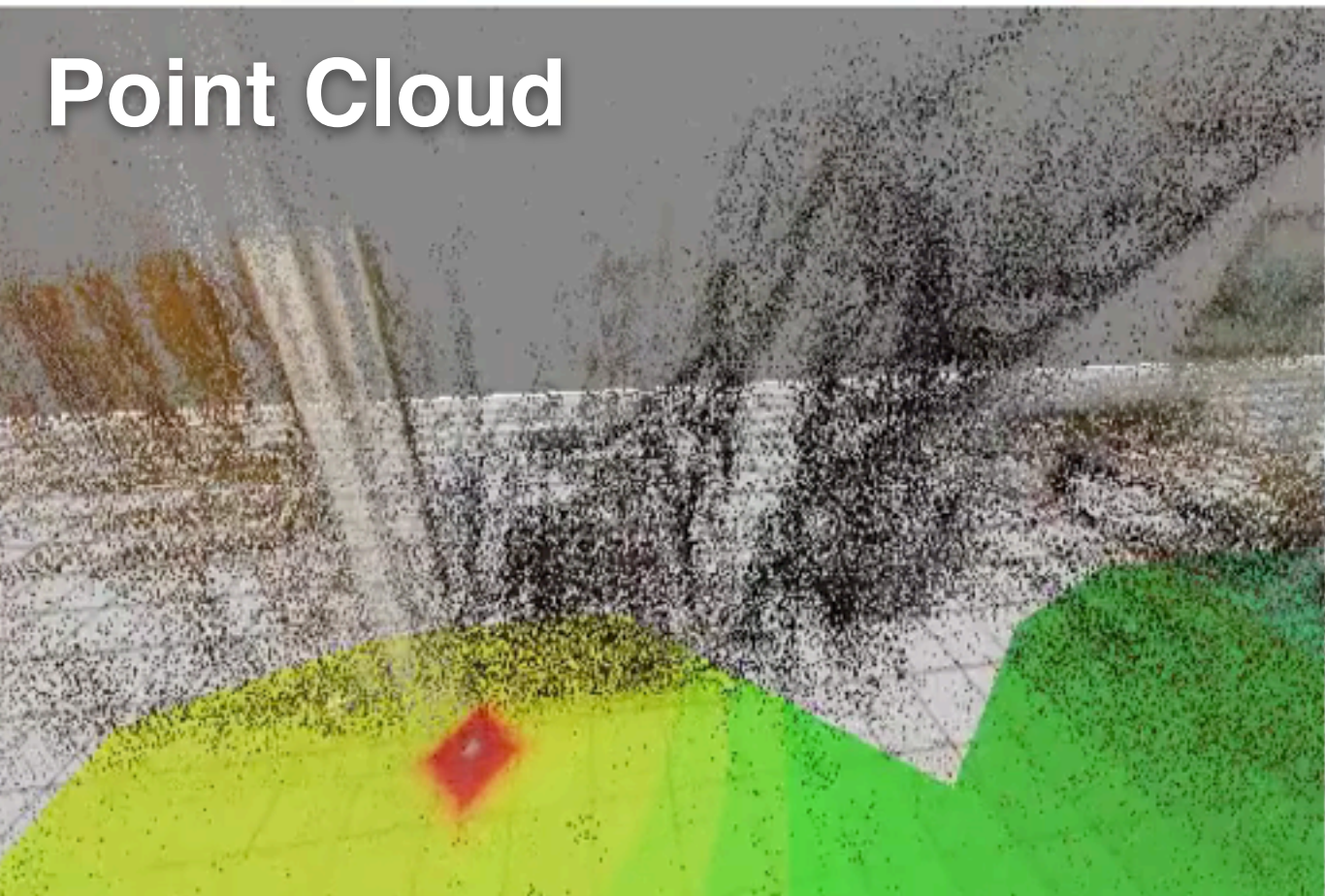




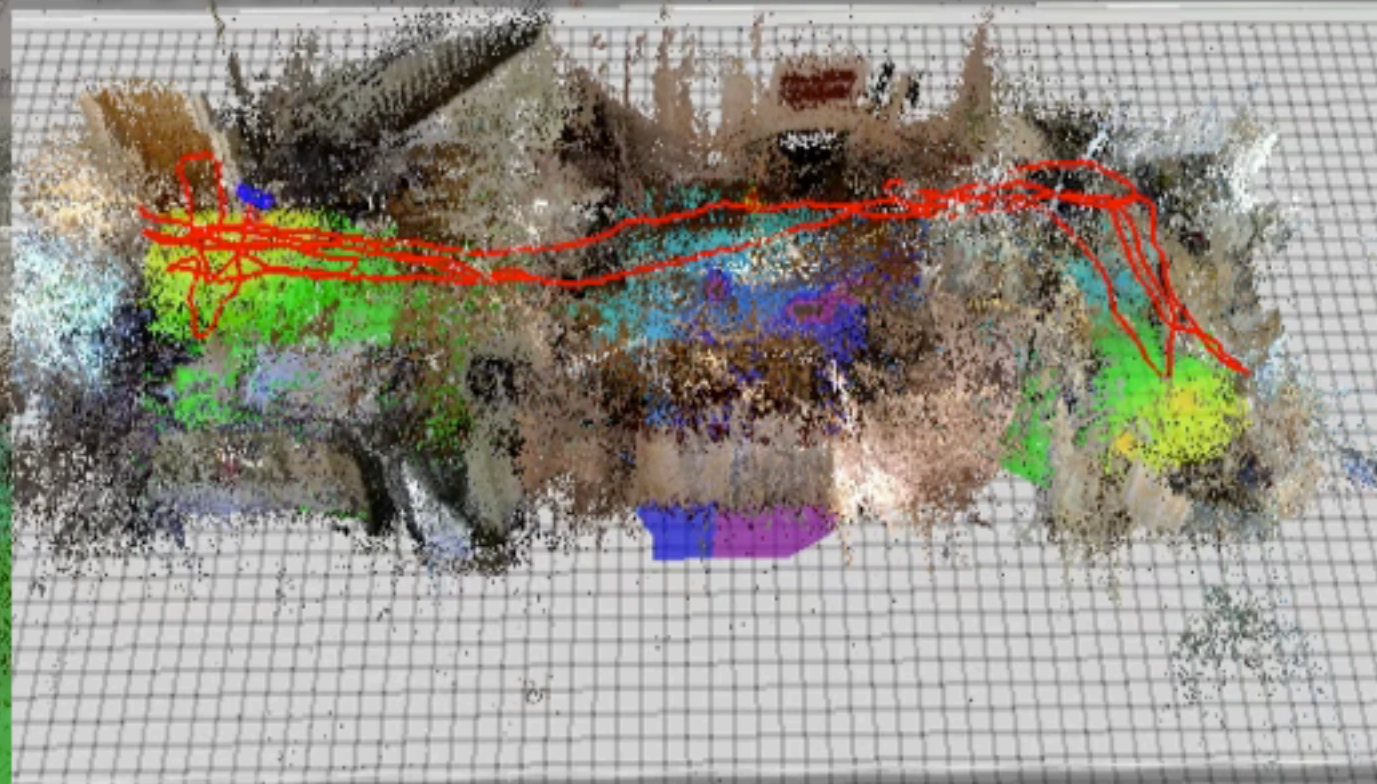
# Places to 'Open Door'



## Point Cloud



## Localization

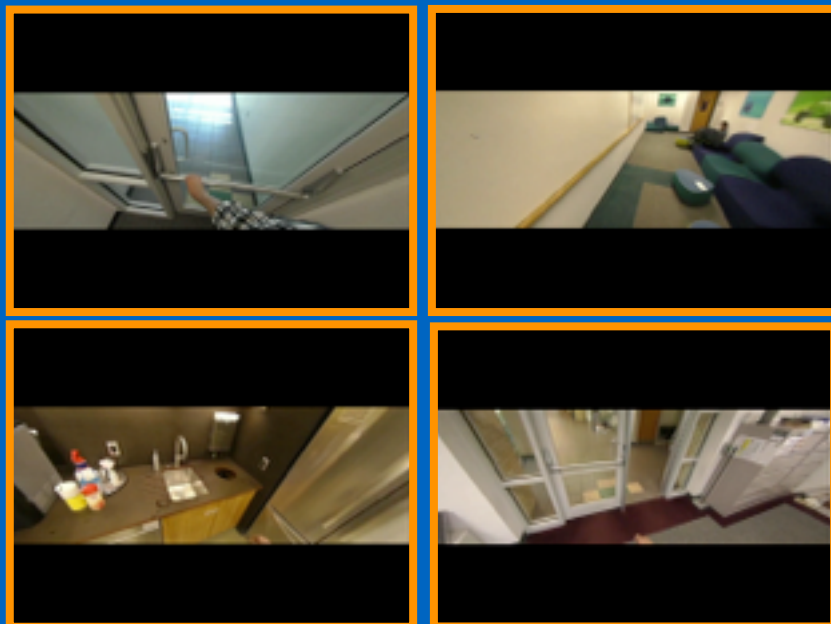




# Activity Detections

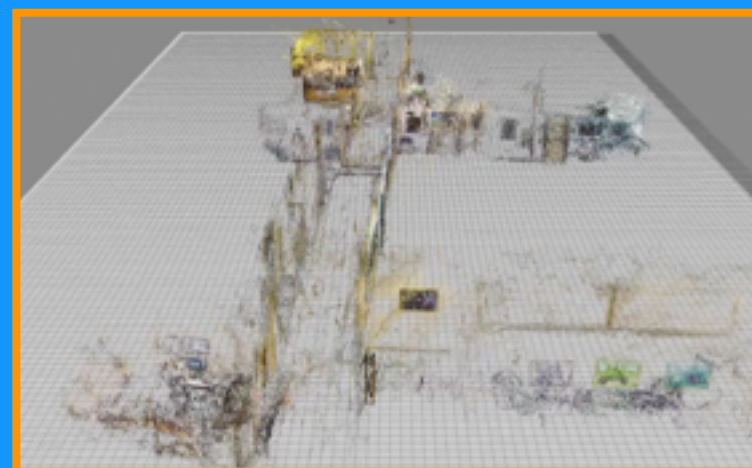
In **one or more** scenes

## Scene 1 Detections

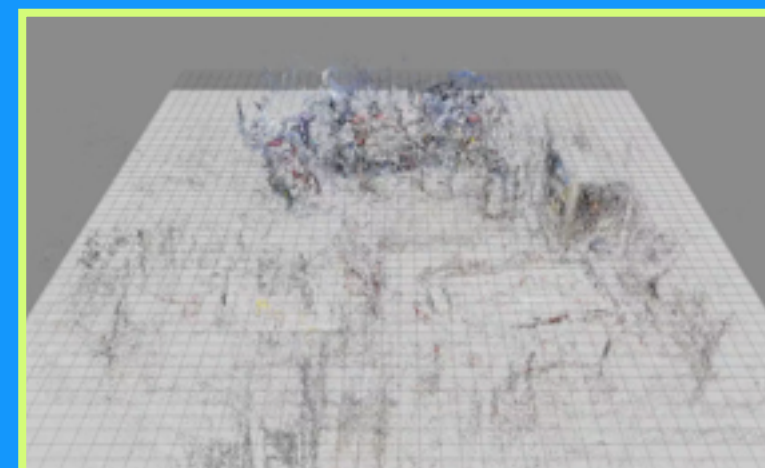


# Scenes

## Scene 1



## Scene 2



Regularized WNMF Modeling and Matrix Completion

Cross-location similarities

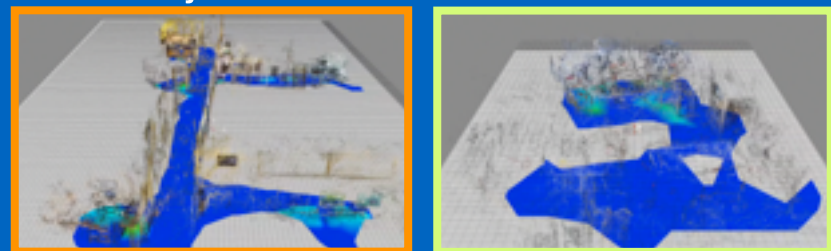
Actions



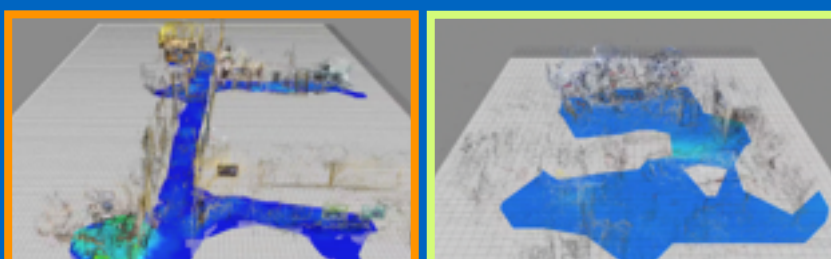
Locations

# Appearance Info

Object Detection Features

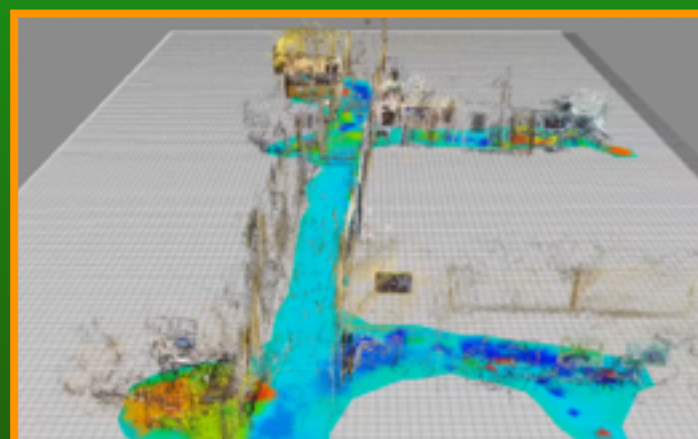


Scene Classification Features

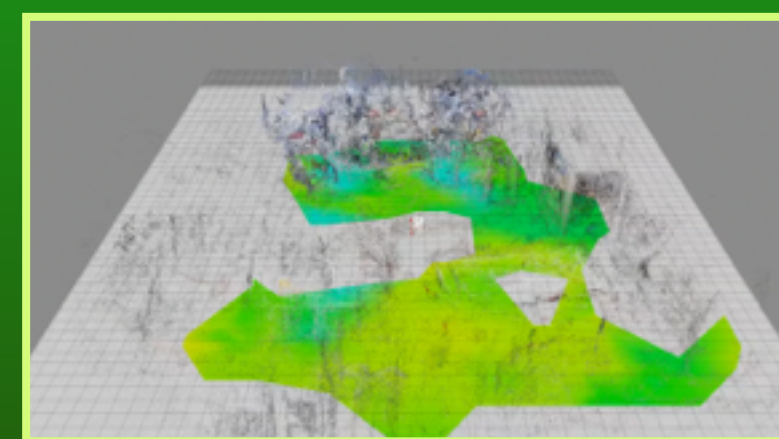


# Action Maps

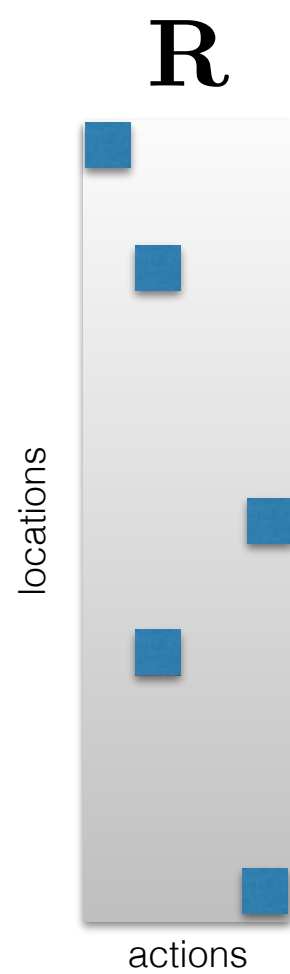
## Scene 1 'Sit' Action Map



## Scene 2 'Sit' Action Map



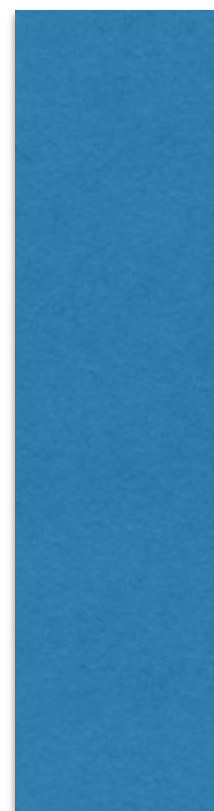
sparse  
action map



dense  
action map



$\tilde{\mathbf{R}}$



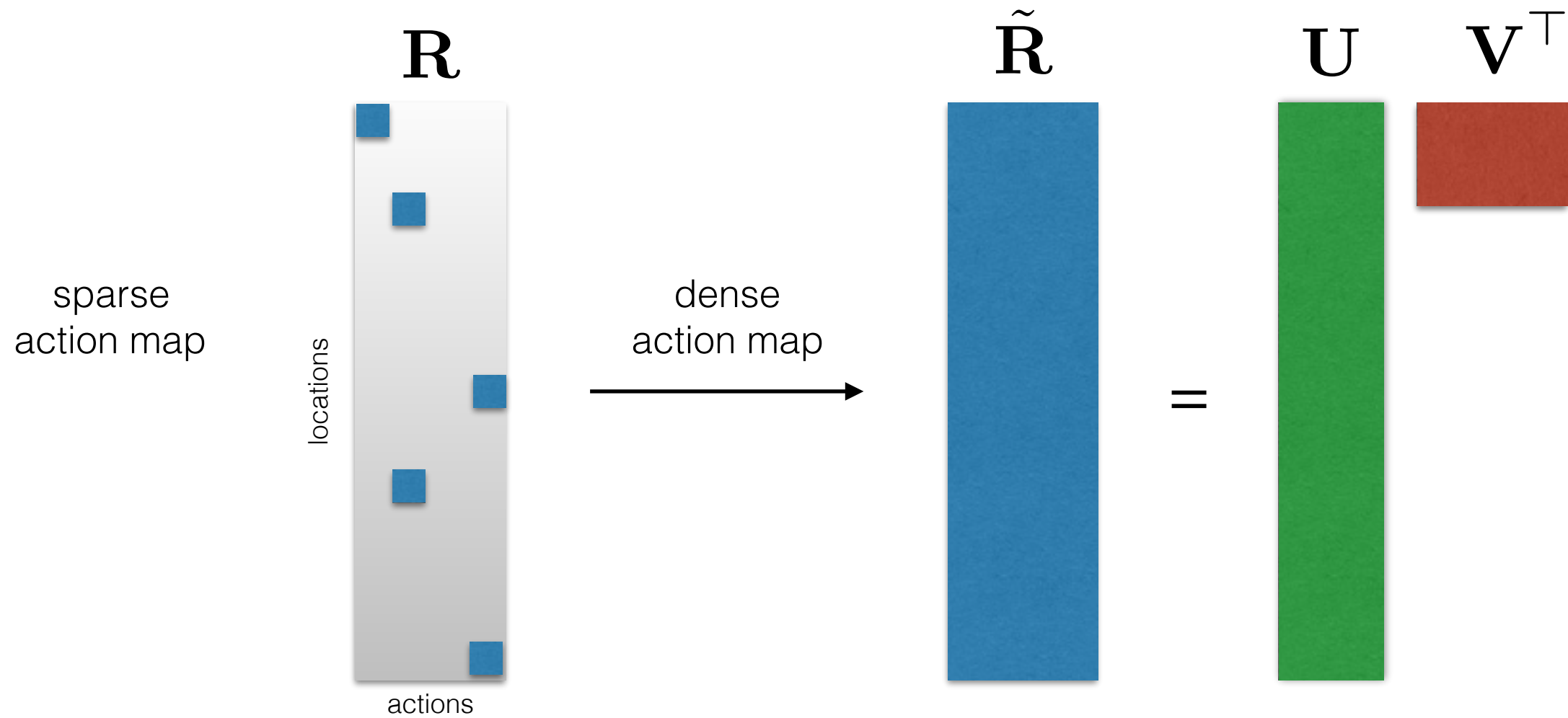
=

$\mathbf{U}$



$\mathbf{V}^\top$



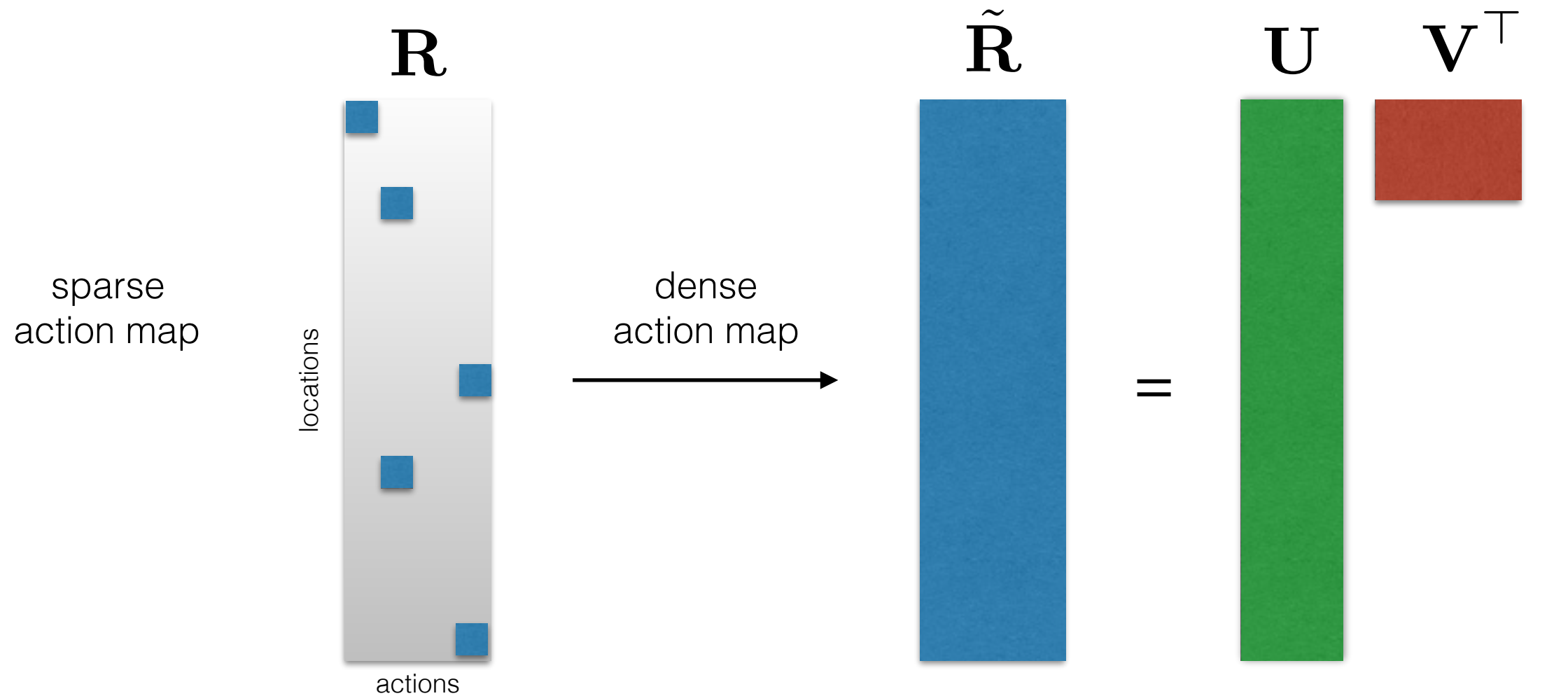


Binary mask  
(locations with actions)

Factorization error  
(non-negative)

$$\arg \max_{\mathbf{U}, \mathbf{V}} ||\mathbf{W} \circ (\mathbf{R} - \mathbf{UV}^\top)||_F^2$$

under-determined system



Binary mask  
(locations with actions)

Factorization error  
(non-negative)

Location similarity  
(scene, object, position)

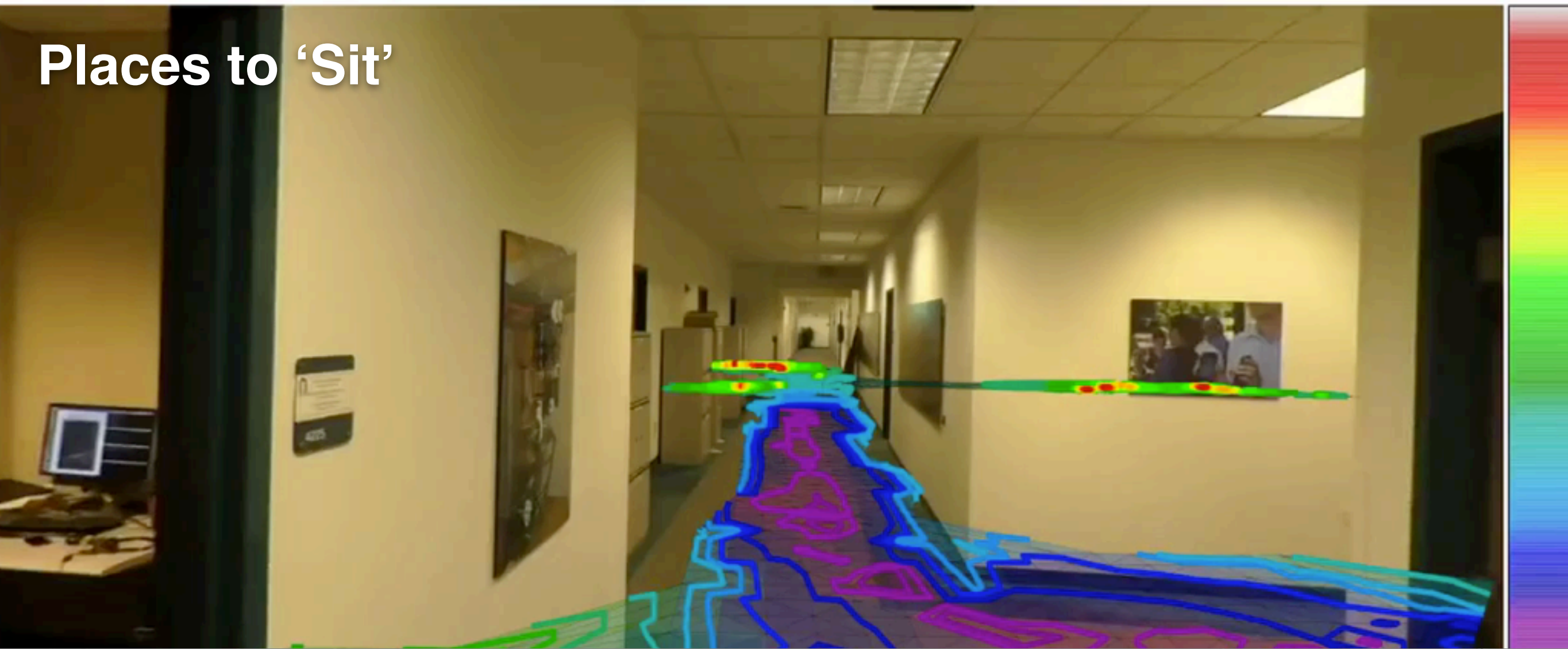
$$\arg \max_{\mathbf{U}, \mathbf{V}} ||\mathbf{W} \circ (\mathbf{R} - \mathbf{UV}^\top)||_F^2 + \lambda \sum_{ij} ||\mathbf{u}_i - \mathbf{u}_j||^2 k(\mathbf{u}_i, \mathbf{u}_j)$$

under-determined system

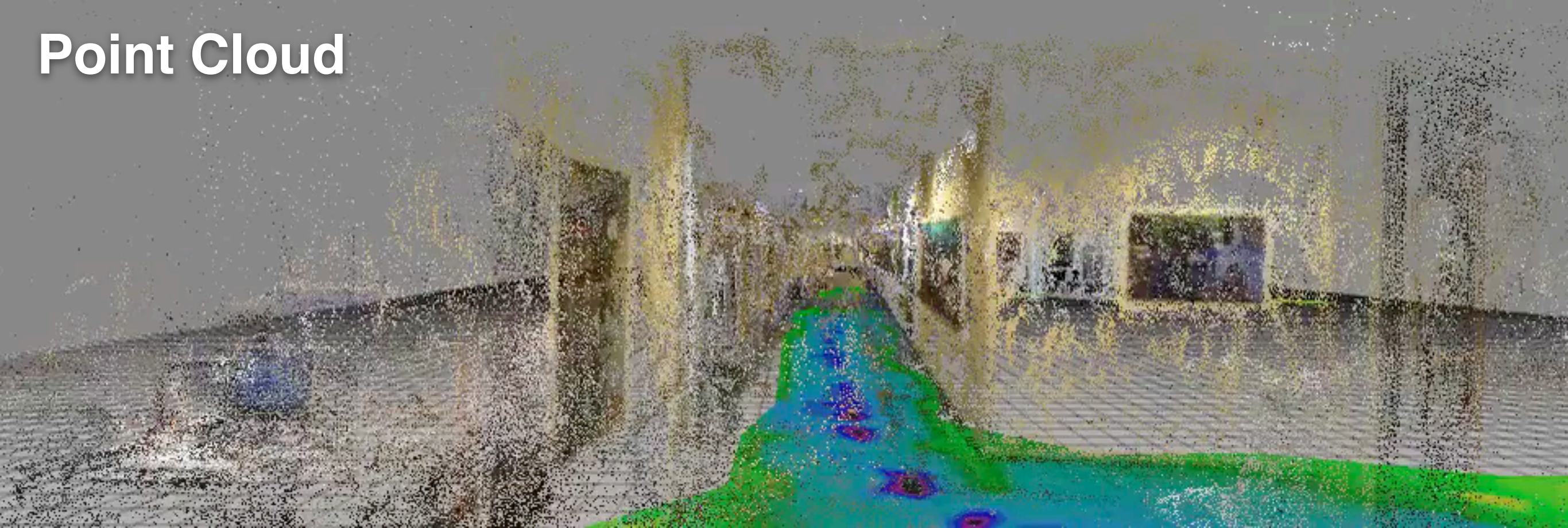
use regularization across rows  
(locations)



# Places to 'Sit'



# Point Cloud





## Understand interactions with **things**



1. Recognizing activities  
CVPR 2016

2. Learning scene functionality  
CVPR 2016

## Understand interactions with **people**



3. Recognizing social interactions  
CVPR 2016

## Understand interactions with **things**



**1. Recognizing activities**  
**CVPR 2016**

**2. Learning scene functionality**  
**CVPR 2016**

## Understand interactions with **people**



**3. Recognizing social interactions**  
**CVPR 2016**



Typical view from the first-person POV



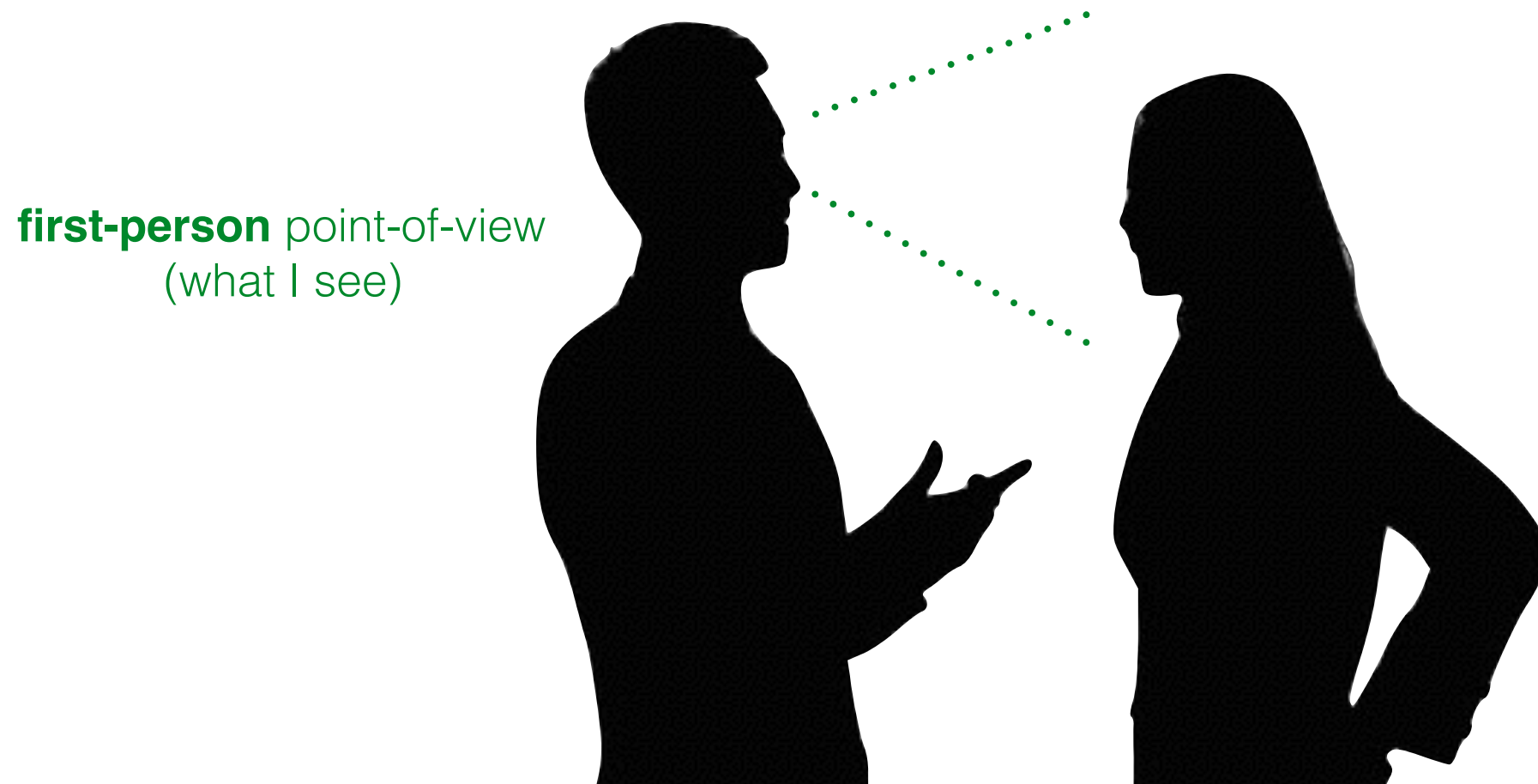
If **two** interacting people are wearing cameras...



...there are **two** points-of-view  
(for each person)

Ryo Yonetani, Kris M. Kitani, Yoichi Sato.  
Recognizing micro-actions and reactions from paired egocentric videos.  
Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

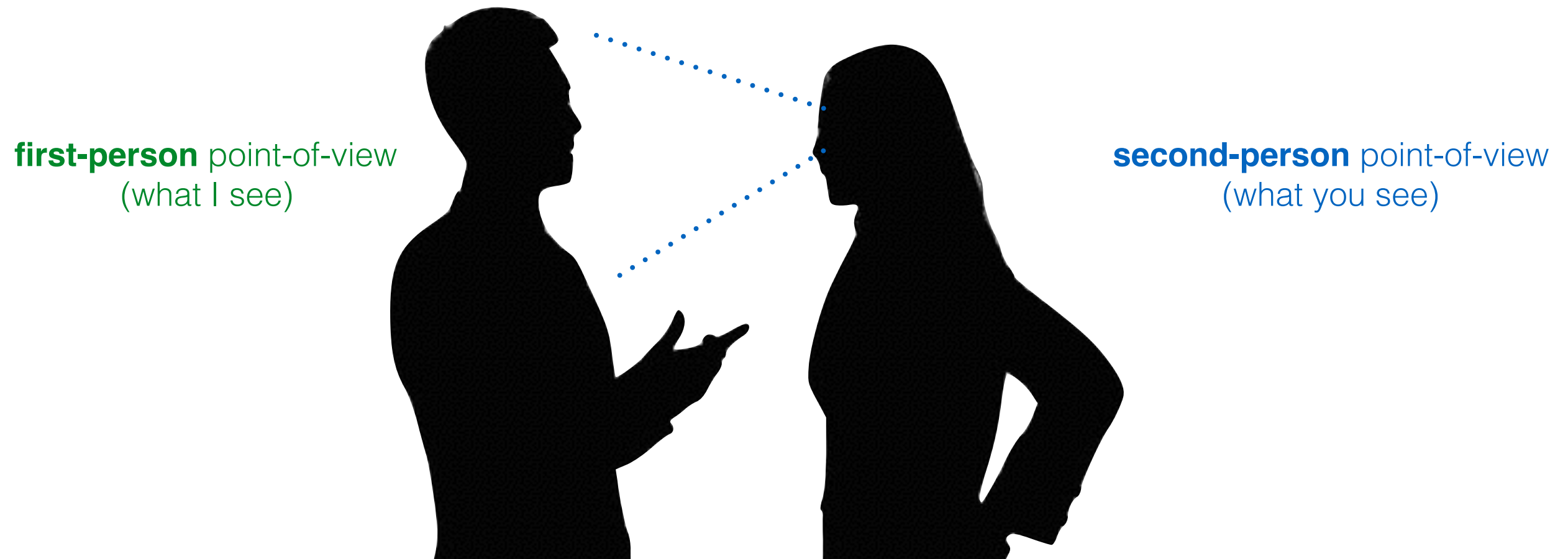
If **two** interacting people are wearing cameras...



...there are **two** points-of-view  
(for each person)

Ryo Yonetani, Kris M. Kitani, Yoichi Sato.  
Recognizing micro-actions and reactions from paired egocentric videos.  
Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

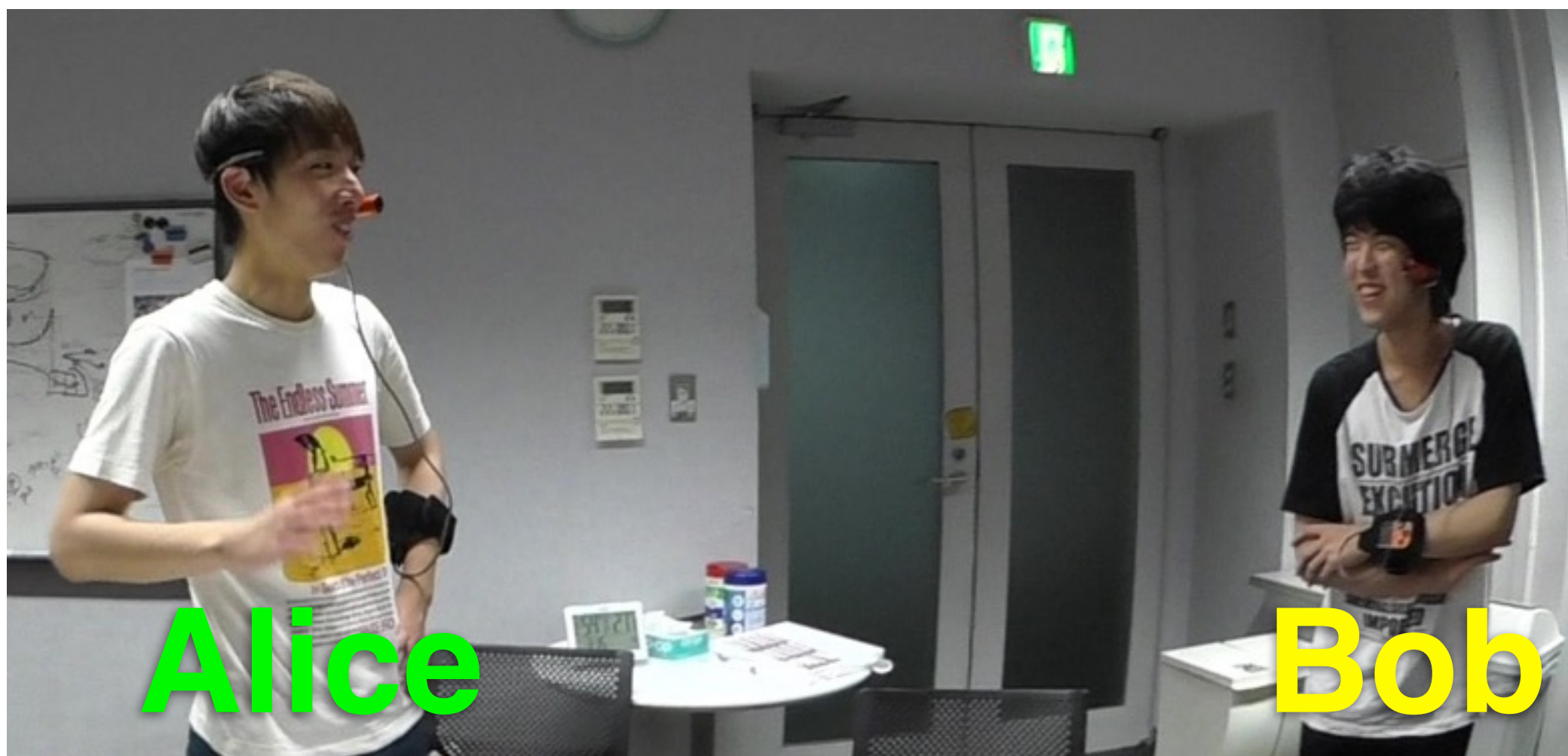
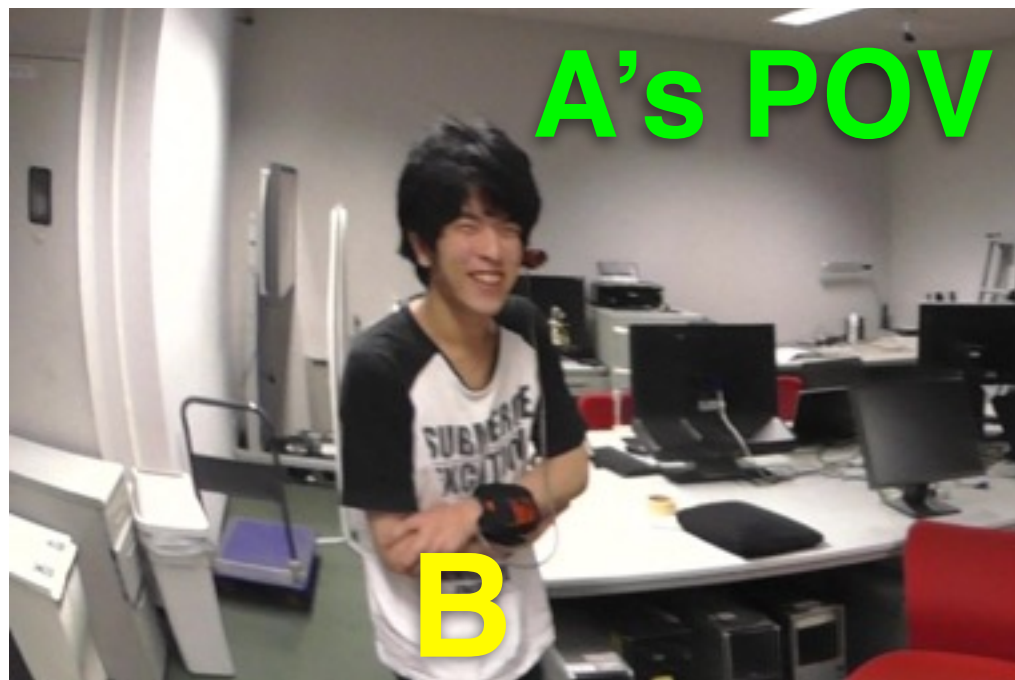
If **two** interacting people are wearing cameras...



...there are **two** points-of-view  
(for each person)

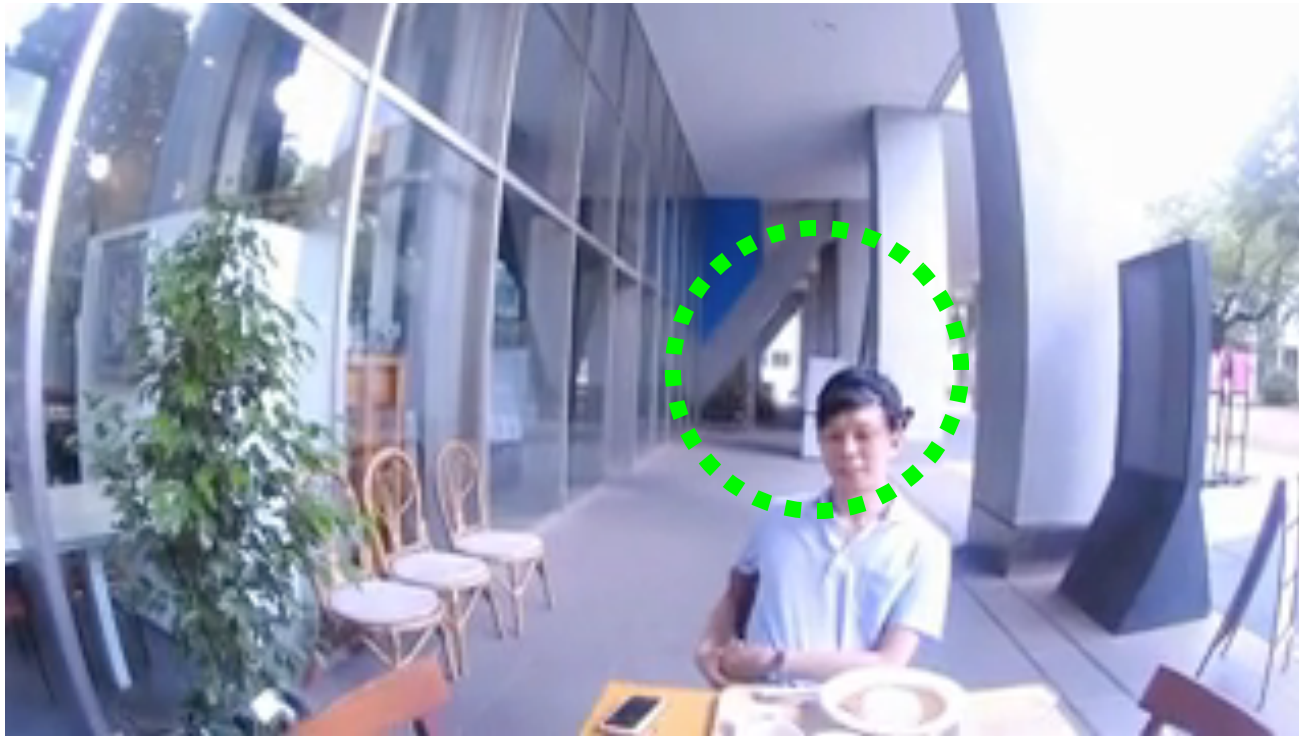
Ryo Yonetani, Kris M. Kitani, Yoichi Sato.  
Recognizing micro-actions and reactions from paired egocentric videos.  
Conference on Computer Vision and Pattern Recognition (CVPR), 2016.





Paired Egocentric Videos

# Daily social interactions contain subtle micro-actions



Slight head motion is hard to detect from  
**second-person POV**



Unextended hand motion is hard to detect  
**first-person POV**

**Micro-actions** are hard to detect...

...but can be resolved with **paired** egocentric videos





Slight head motion is hard to detect from



**Induces large global motion!**



Unextended hand motion is hard to detect



**Hand is clearly observed!**

Both views are complementary and essential for micro-action recognition



Action-reaction pairs are correlated



pointing gesture (**action**) induces a slight shift in attention (**reaction**)

these correlation can be used to recognize micro-actions

# What kinds of features can we use?

If **two** interacting people are wearing cameras...

- 1. ego-motion of self
- 2. appearance of partner



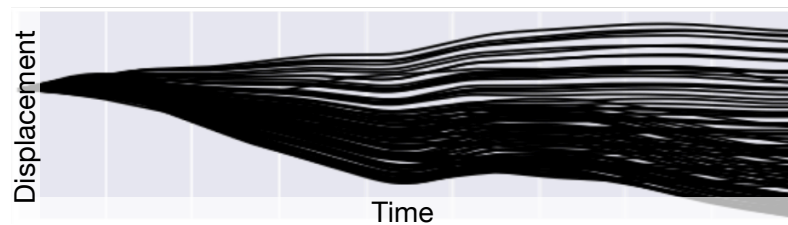
- 3. ego-motion of self
- 4. appearance of partner

...there are **four** feature types

**Input:** egocentric video *pair*



First-person feature of **A**



Second-person feature of **B**

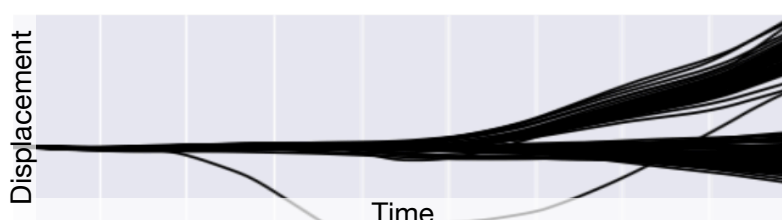


**First-person features**

- Egocentric + object features [Li+, CVPR15]
- Cumulative displacement patterns [Poleg+, CVPR14]
- Pooled time-series encoding [Ryoo+, CVPR15]



Second-person feature of **A**



First-person feature of **B**

**Second-person features**

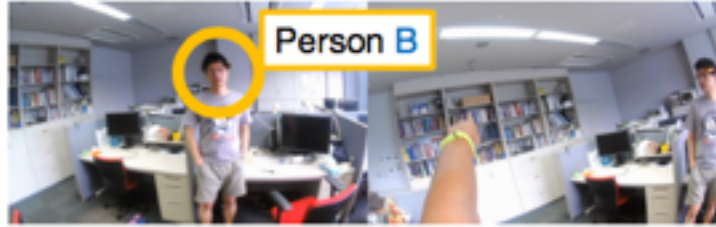
- Dense trajectory [Wang+, ICCV13]
- Two-stream CNN [Simonyan+, NIPS14]
- Trajectory-pooled convolutional descriptor [Wang+, CVPR15]



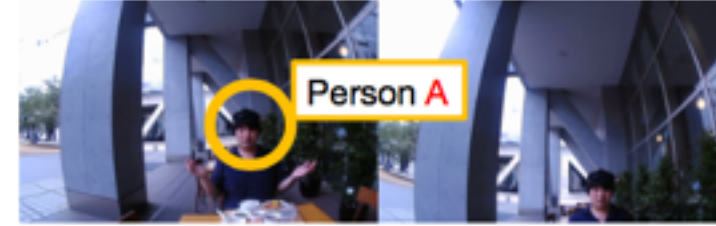
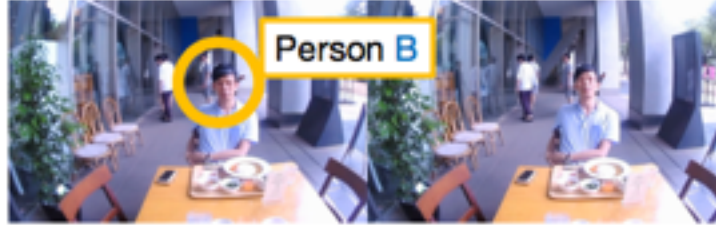
Recognize micro-action  
(point, shift in attention, positive gesture, etc. )



(1) **Pointing** and shift in **attention**



(2) **Gesture** and **positive** response



(3) **Passing** and **receiving** an item




Person **A**'s points-of-view

Person **B**'s points-of-view

		Pointing	Attention	Positive	Negative	Passing	Receiving	Gesture	Average
(1) First-person POV features of <i>A</i>	<b>E</b> [18]	0.65	0.77	0.91	0.88	0.64	0.78	0.73	0.76
	<b>E+O</b> [18]	0.74	0.77	0.94	0.73	0.71	0.85	0.69	0.77
	<b>CD</b> [27]	0.64	0.62	0.58	0.56	0.71	0.71	0.56	0.63
	<b>PoTCD</b> [27, 30]	0.70	0.66	0.94	0.84	0.69	0.74	0.63	0.74
(2) Second-person POV features of <i>A</i>	<b>IDT</b> [39]	0.74	0.71	0.67	0.59	0.81	0.93	0.78	0.75
	<b>TCNN</b> [32]	0.59	0.58	0.55	0.58	0.54	0.67	0.60	0.59
	<b>TDD</b> [40]	0.63	0.70	0.61	0.51	0.68	0.79	0.63	0.65
(3) Multiple POV features of <i>A</i>	<b>E+IDT</b>	0.77	0.73	0.86	0.81	0.82	0.92	0.79	0.81
	<b>E+O+IDT</b>	0.80	0.78	0.95	0.77	0.83	0.95	0.78	0.84
	<b>PoTCD+IDT</b>	0.79	0.78	<b>0.96</b>	0.89	0.84	0.93	0.80	0.86
(4) Multiple POV features of <i>A</i> and <i>B</i>	<b>Degraded-A</b>	0.82	0.76	<b>0.96</b>	0.86	0.56	0.95	0.69	0.84
	<b>Degraded-B</b>	0.73	0.72	0.67	0.61	0.82	0.94	0.78	0.75
	<b>Proposed</b>	<b>0.85</b>	<b>0.83</b>	<b>0.96</b>	<b>0.91</b>	<b>0.89</b>	<b>0.97</b>	<b>0.82</b>	<b>0.89</b>



# Paired egocentric video dataset

- 
- **> 1,000 pairs** of egocentric videos
  - Recorded during 28 different two-persons interactions
  - 7 different micro-actions and reactions



## Understand interactions with **things**



1. Recognizing activities  
CVPR 2016

2. Learning scene functionality  
CVPR 2016

## Understand interactions with **people**



3. Recognizing social interactions  
CVPR 2016