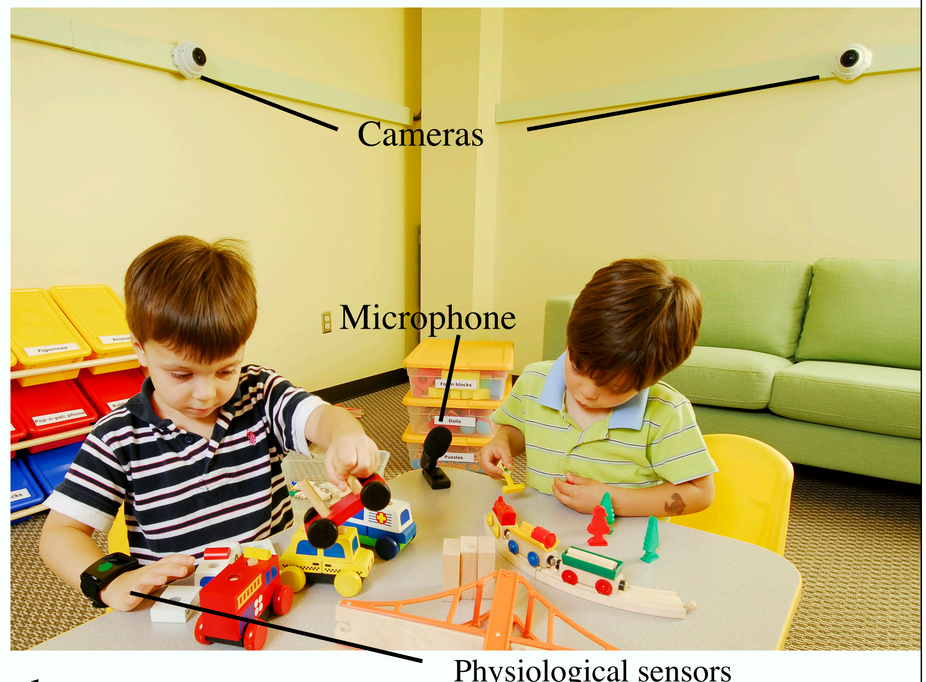
But what are they doing with their hands?

D.A. Forsyth, UIUC
with Daphne Tsatsoulis, UIUC

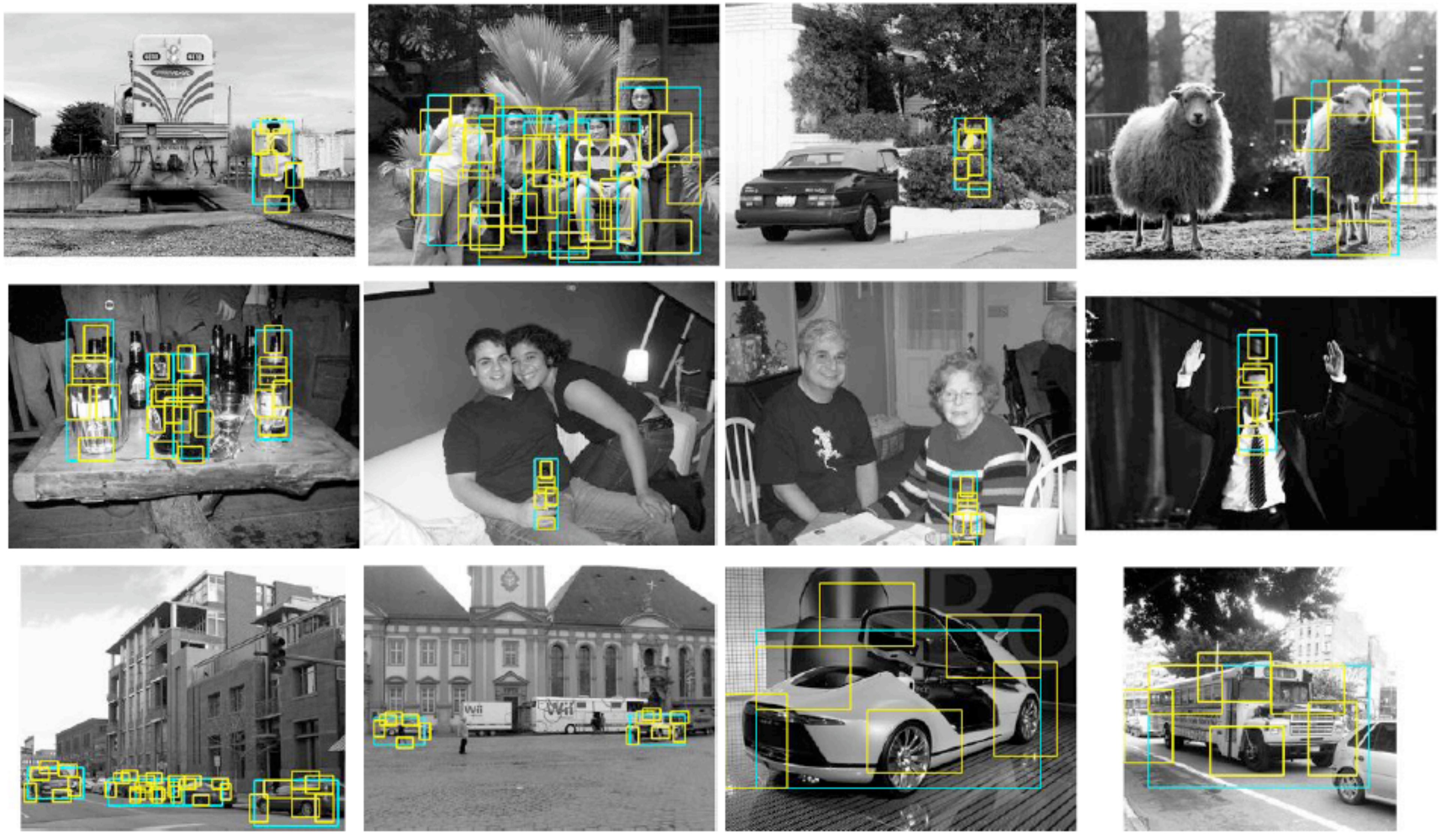
Computational Behavioural Science



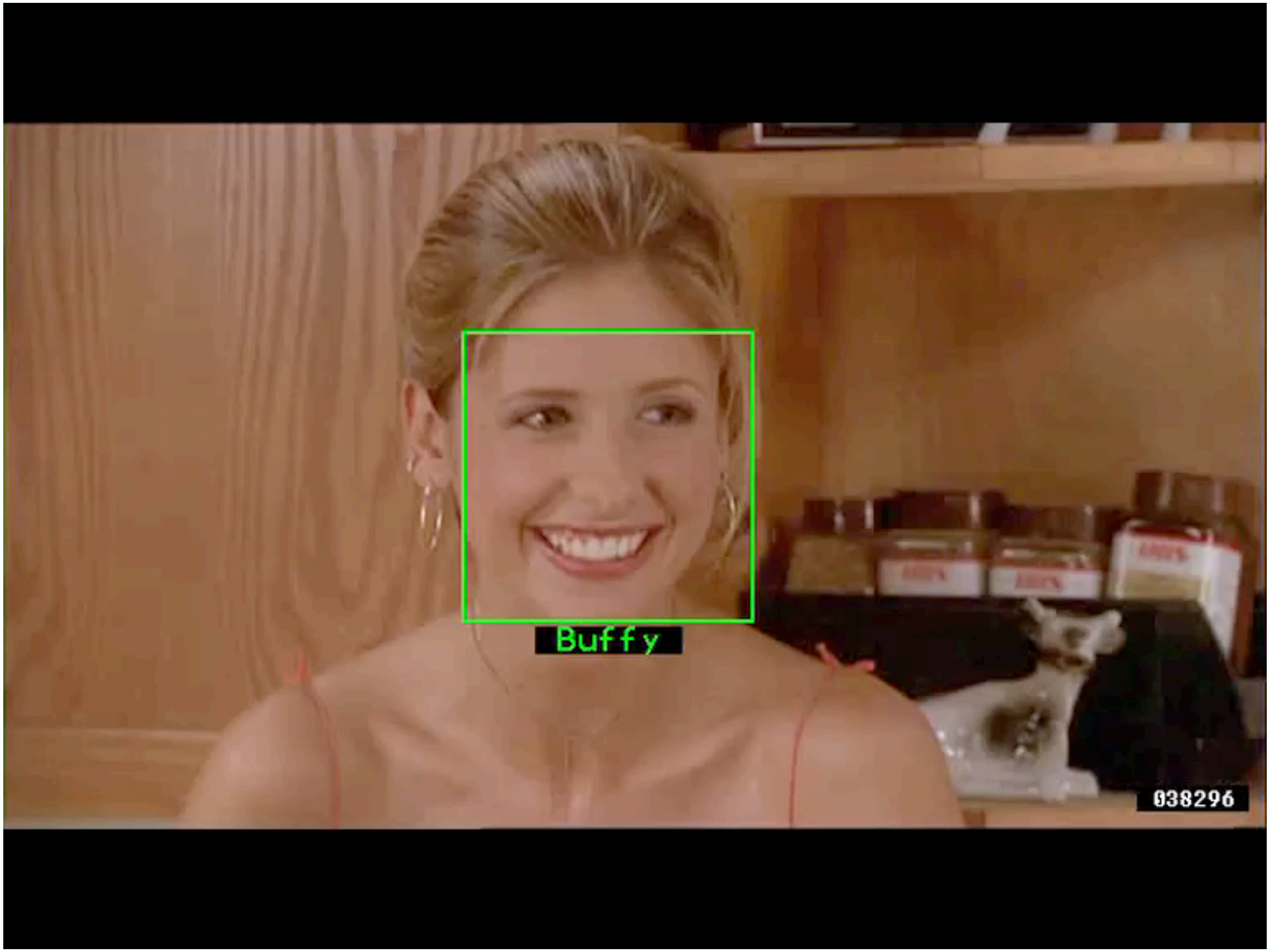
- Observe people
    - Using vision, physiological markers
  - Interacting, behaving naturally
    - In the wild
- drive feedback for therapy
    - Eg reward speech
- Applications
    - Model: screen for ASD
    - Other:
        - Any w here large scale observations help
            - Support in home care
            - Support care for demented patients
            - Support stroke recovery
            - Support design of efficient buildings
- 10M$, 5yr NSF award under Expeditions program
    - GaTech, UIUC(DAF, Karahalios), MIT, CMU, Pittsburgh, USC, Boston U

Powerful Technologies

- Structure from motion
  - reconstruct a 3D world and camera movement from video or pix
- Classification and detection
  - put in features, out comes a decision
  - sweep a window, classify - is it a face or not?
- Tracking
  - mark locations from frame to frame



P. Felzenszwalb, D. McAllester, D. Ramanan. “A Discriminatively Trained, Multiscale, Deformable Par
Model” CVPR 2008.

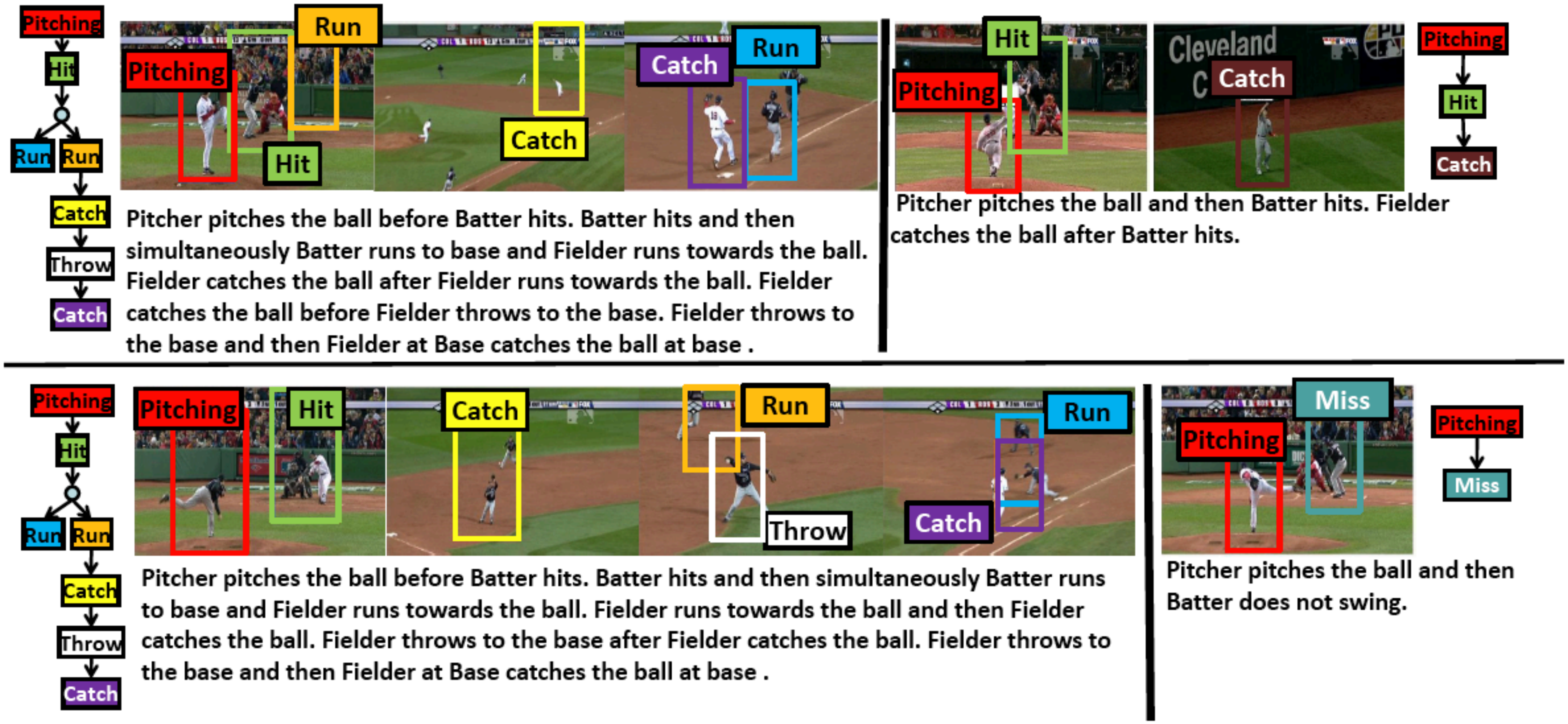


Everingham, M., Sivic, J. and Zisserman, A.
“Hello! My name is... Buffy” - Automatic naming of characters in TV video
BMVC 2006

Looking at people

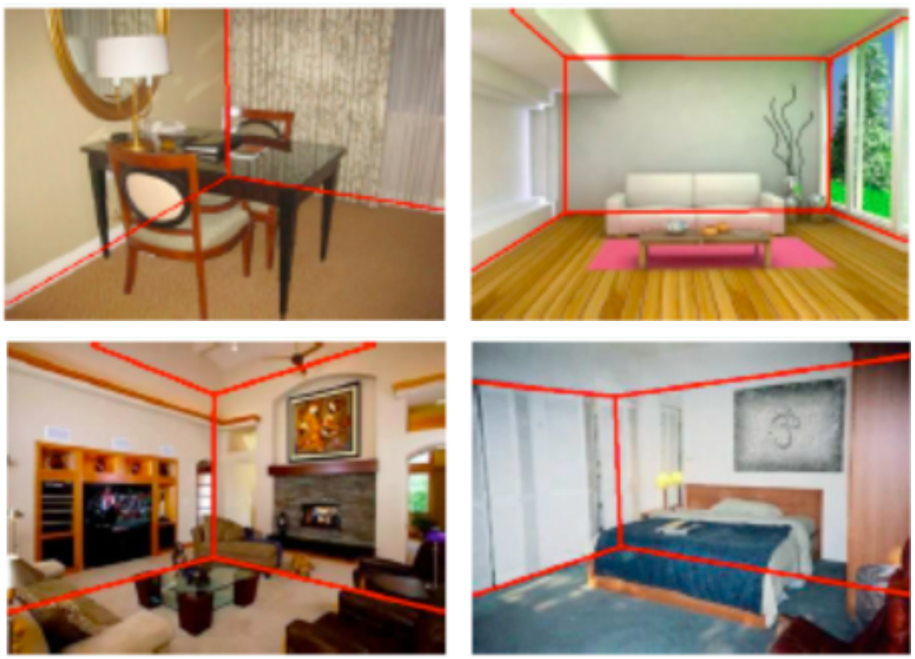
- Questions:
  - What are they doing?
  - Where are they doing it?
  - Why are they doing it?
  - What will happen?
- Problems:
  - Knowing what to measure is hard
    - faces; body configuration; hand positions; etc?
  - Practical difficulties in measurement
    - small fast body parts (eg hands); clothing
  - Knowing what to report is hard
    - much behavior is quite unusual
    - what should we say about behaviors?

Predicting stylized narrations



Gupta ea 09

Where People Act



Hedau et al 09

Hedau et al 12

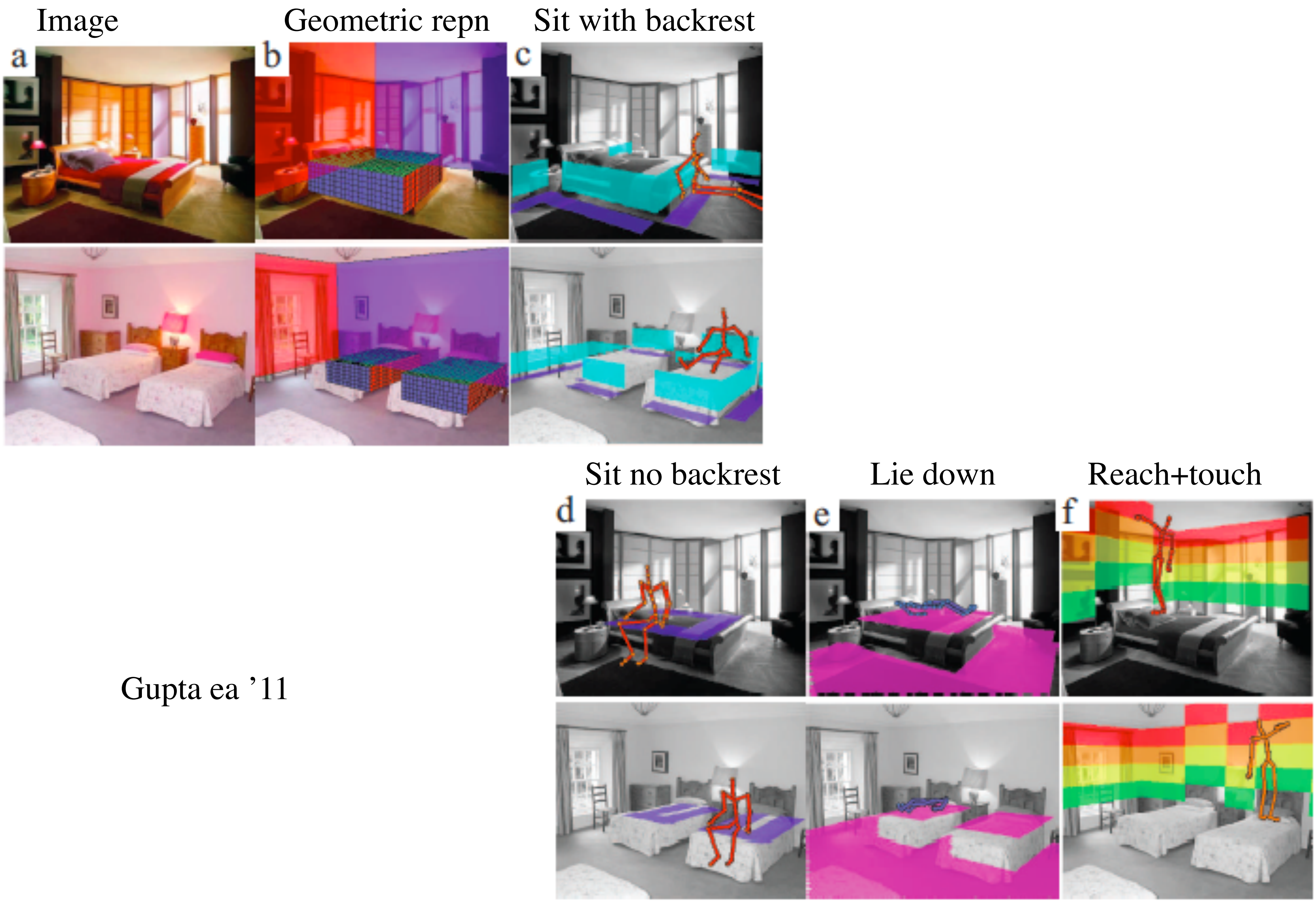
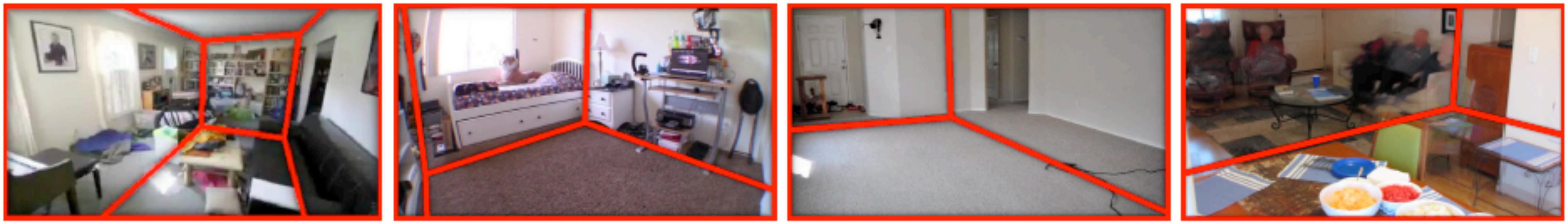


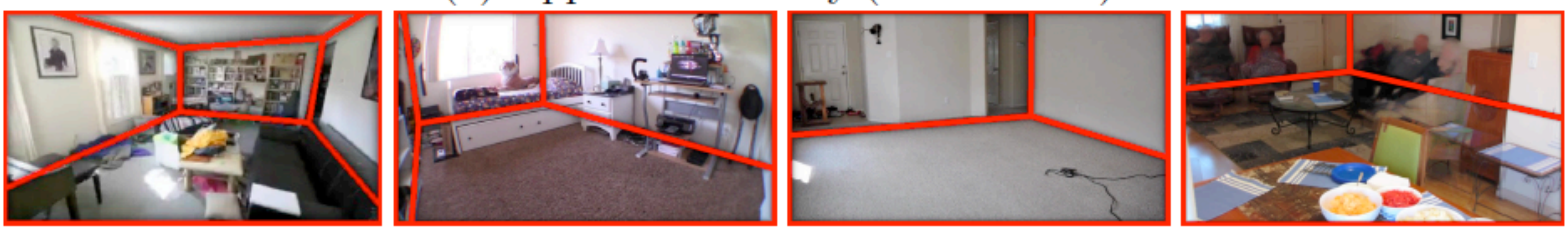
Image
Geometric repn
Sit with backrest
a
b
c
Sit no backrest
Lie down
Reach+touch
d
e
f

Gupta ea ’11

Human motion reveals space



(a) Appearances Only (Hedau *et al*).

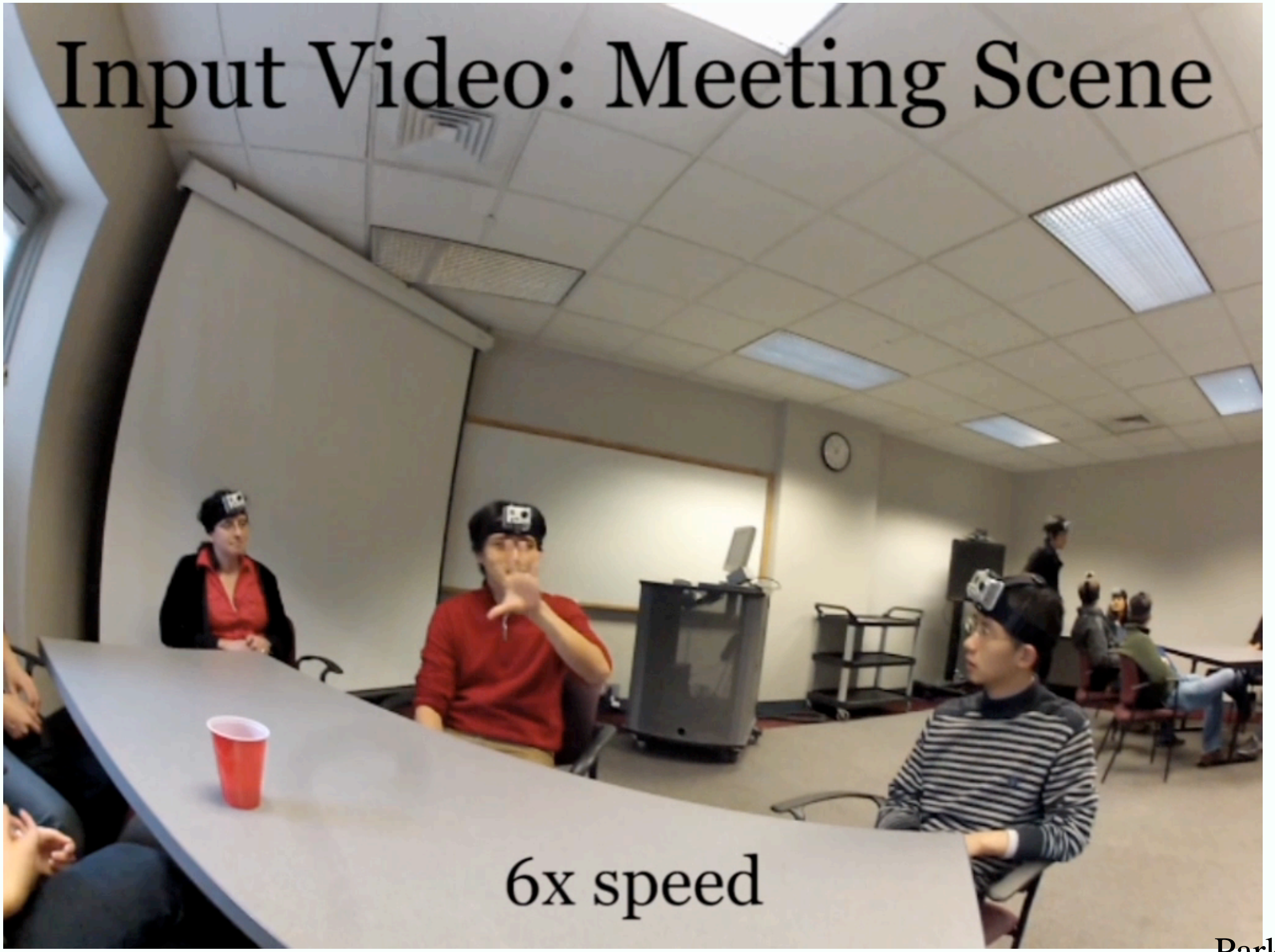


(b) Appearances + People (Our approach).

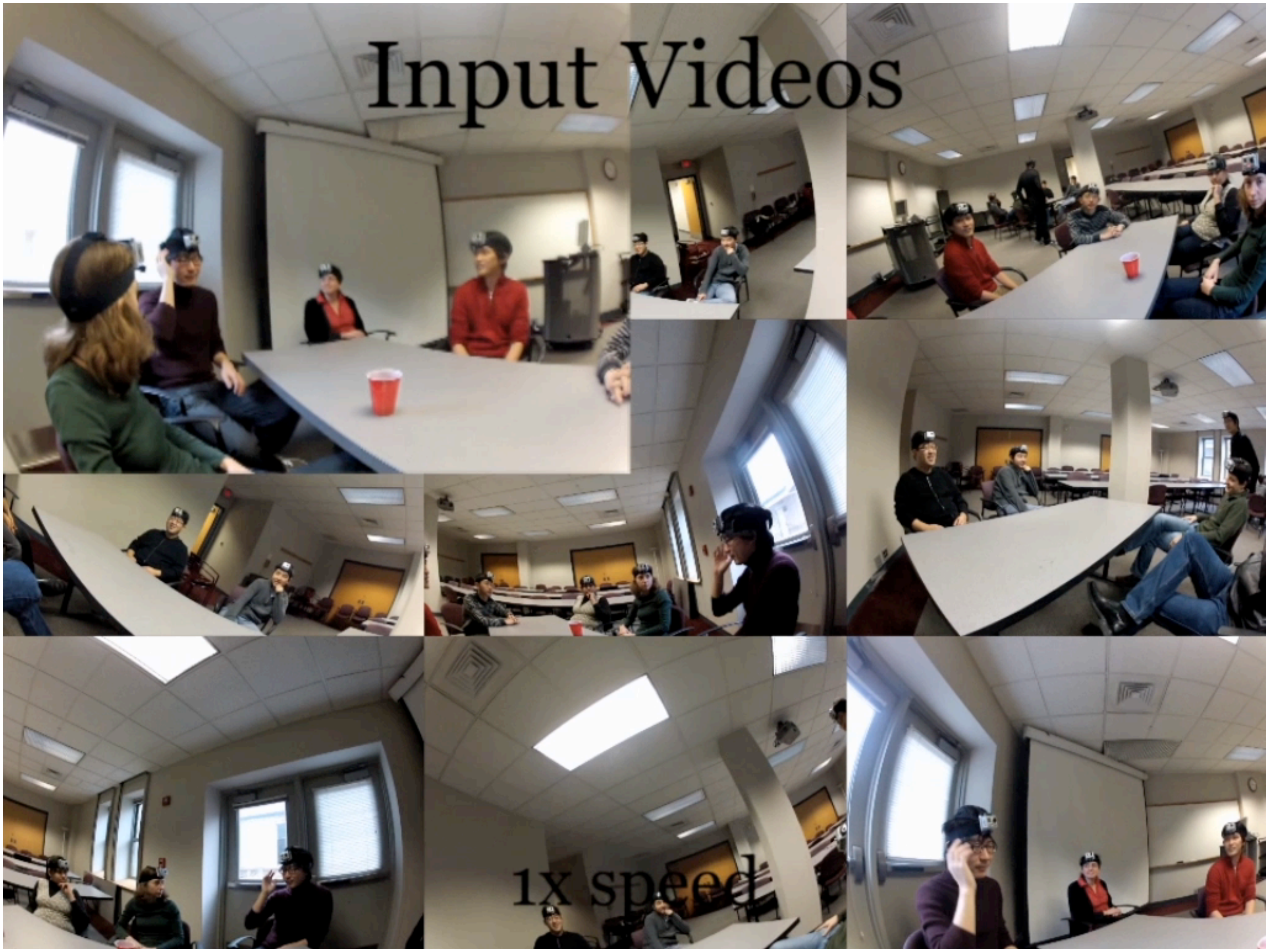
Fig. 6. Timelapse experiment: A comparison of (a) appearance only baseline [6] with (b) our improved room layout estimates. In many cases, the baseline system selects small rooms due to high clutter. On the right, even though the room is not precisely a cuboid, our approach is able to produce a significantly better interpretation of the scene.

Fouhey et al 12

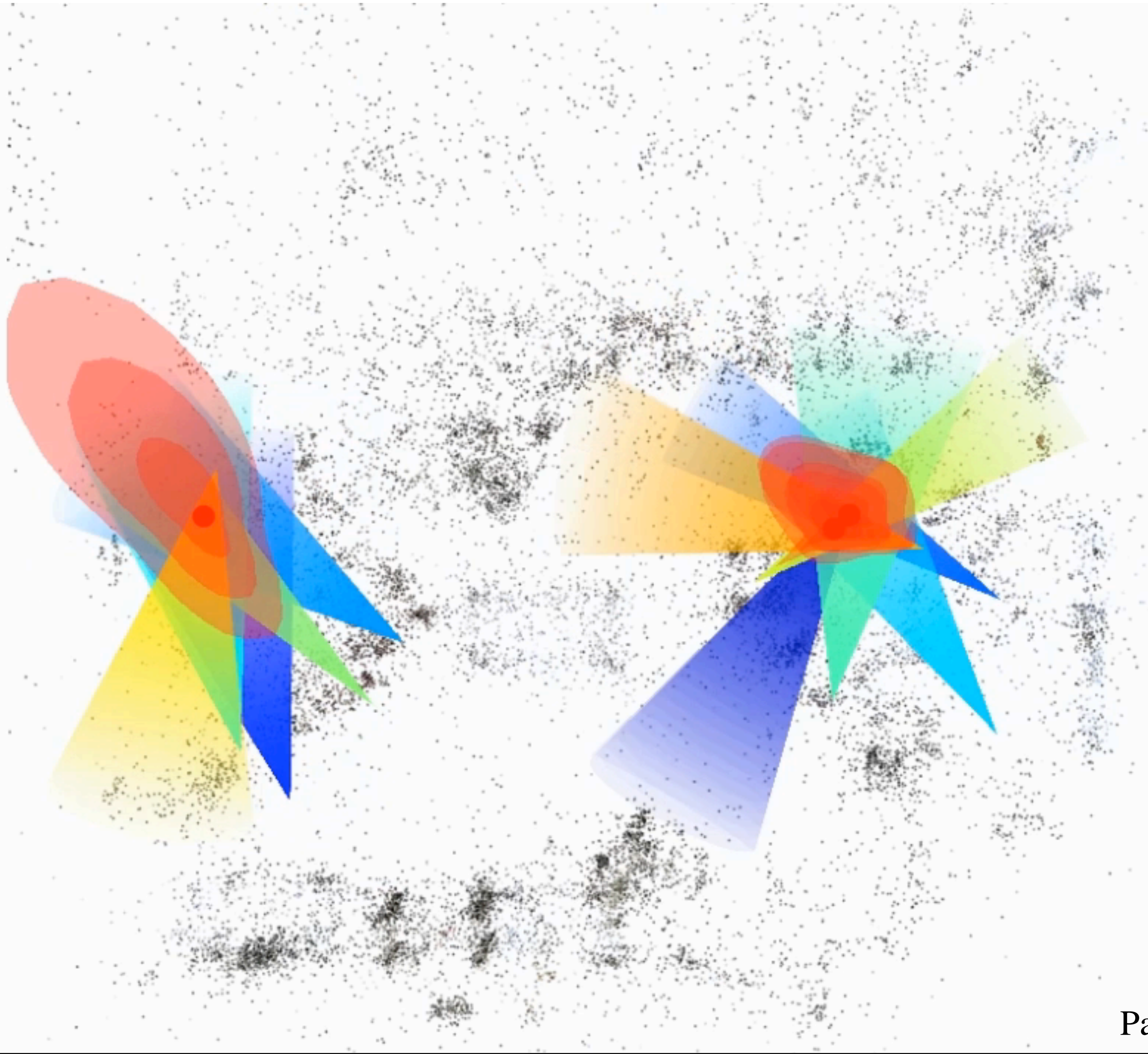
Where people look



Park et al 2011



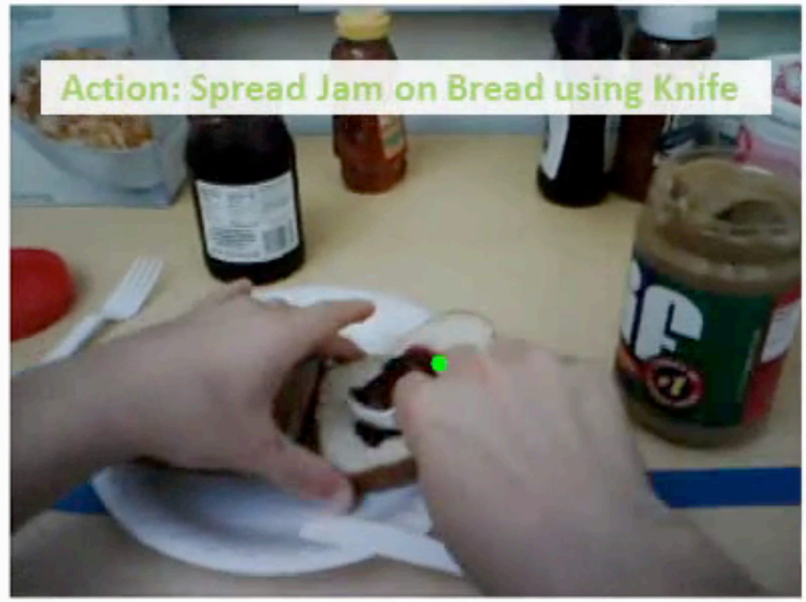
Park et al 2011



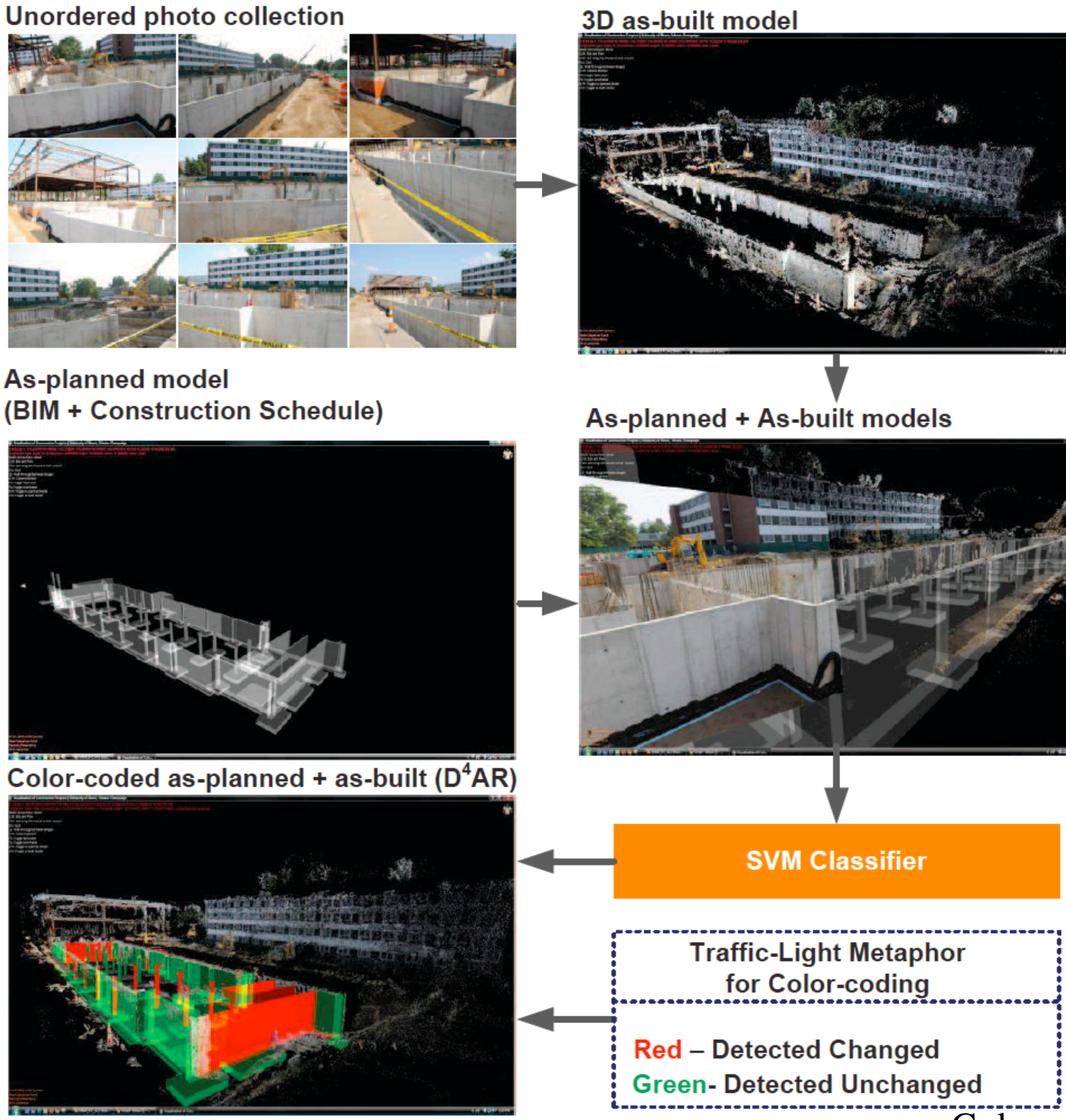
Park et al 2011

Test video w/ ground-truth gaze

Predicted gaze and action label



Fathi et al 2012



Golparvar-Fard et al 11

Parsing - where is the body?

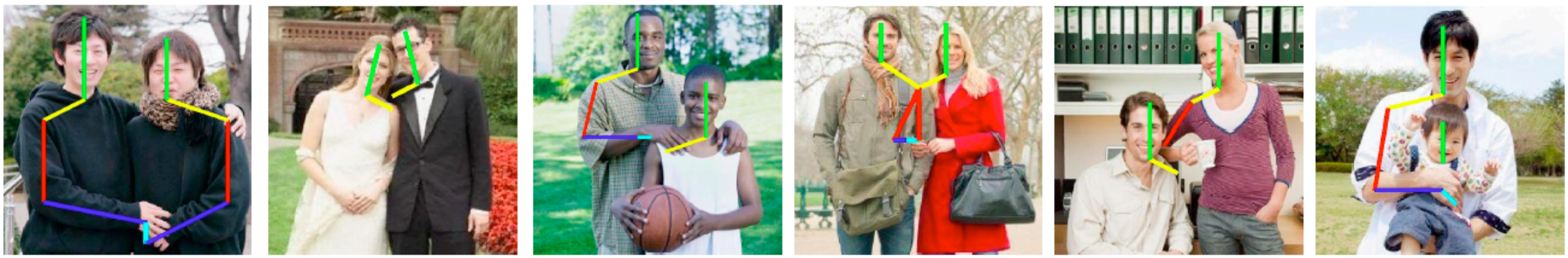
- Advances in human parsing
  - Appearance/layout interaction (Ramanan 06)
  - Improved appearance models (Ferrari et al 08; Eichner Ferrari 10)
  - Branch+bound (Tian Sclaroff 10)
  - Interactions with objects (Yao Fei-Fei 10; Desai et al 10)
  - Coverage and background (Buehler ea 08; Jiang 09)
  - Complex spatial models (Sapp ea 10a)
  - Cascade models (Sapp ea 10b)
  - Full relational models (Tran Forsyth 10)
  - Poselet style models (Bourdev ea 09; 10; 11; Wang ea 11)

Is the parse successful?

Fig. 9. Example evaluations. The pose estimates in the first two rows are correctly classified as successes by our pose evaluator. The last two rows are correctly classified as failures. The pose evaluator is learnt using the regime B and with a CPC threshold of 0.3. Poses in rows 1,3 are estimated by Eichner and Ferrari [5], and poses in rows 2,4 are estimated by Yang and Ramanan [22].

Jamalamadaka, 12

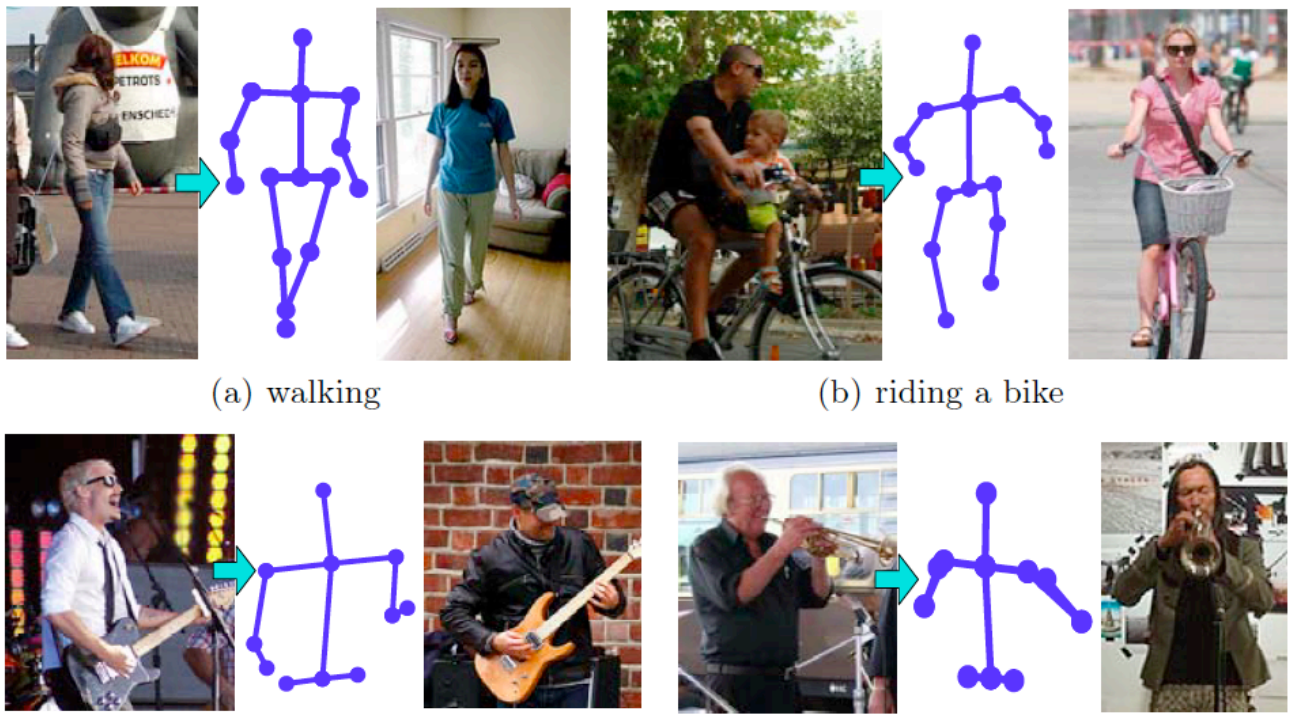
Proxemics - who's nearby?



(a) Hand hand (b) Shoulder shoulder (a) Hand shoulder (d) Hand elbow (c) Elbow shoulder (e) Hand torso

Yang et al, 2012

Actions reveal shape reveals action



(a) walking (b) riding a bike

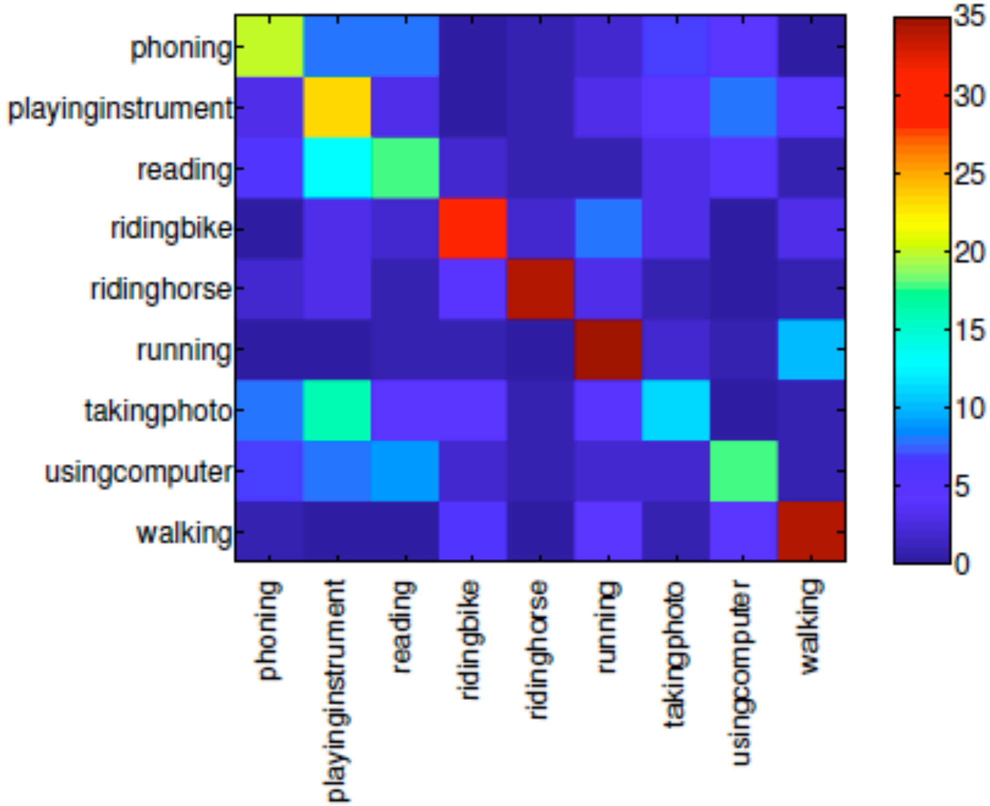
(c) playing guitar (d) playing trumpet

Fig. 4. The 3D representation of human body key-points allows us to rotate one image to the same view-point of the other image, and thus achieve view-independent similarity matching. In each subfigure, from left to right: human in profile view, its pose in frontal view, and the other human with the same action in the frontal view.

Yao et al 12

Desai+Ramanan, 2012

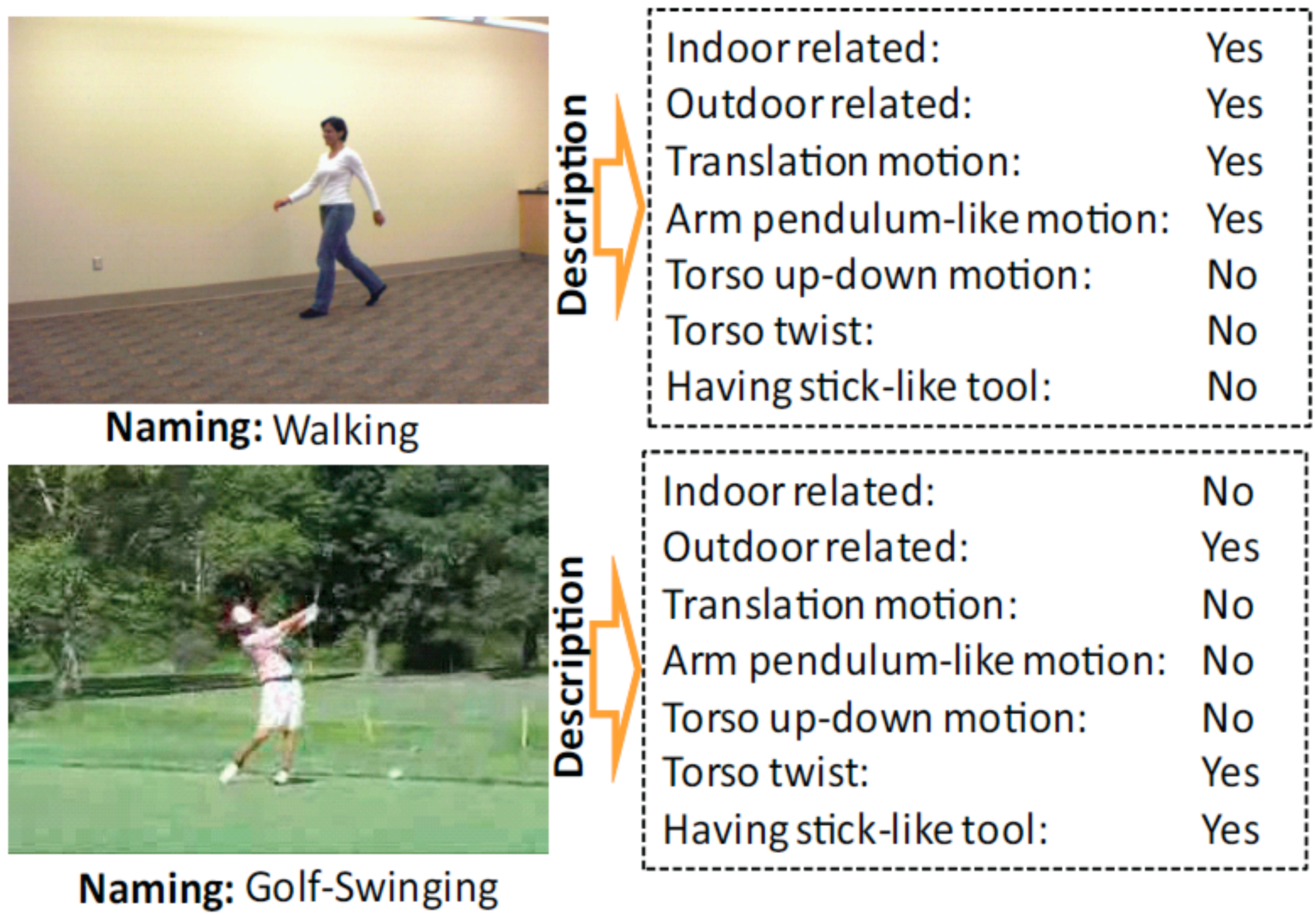
Poselets+context reveal actions



Maji et al., 11

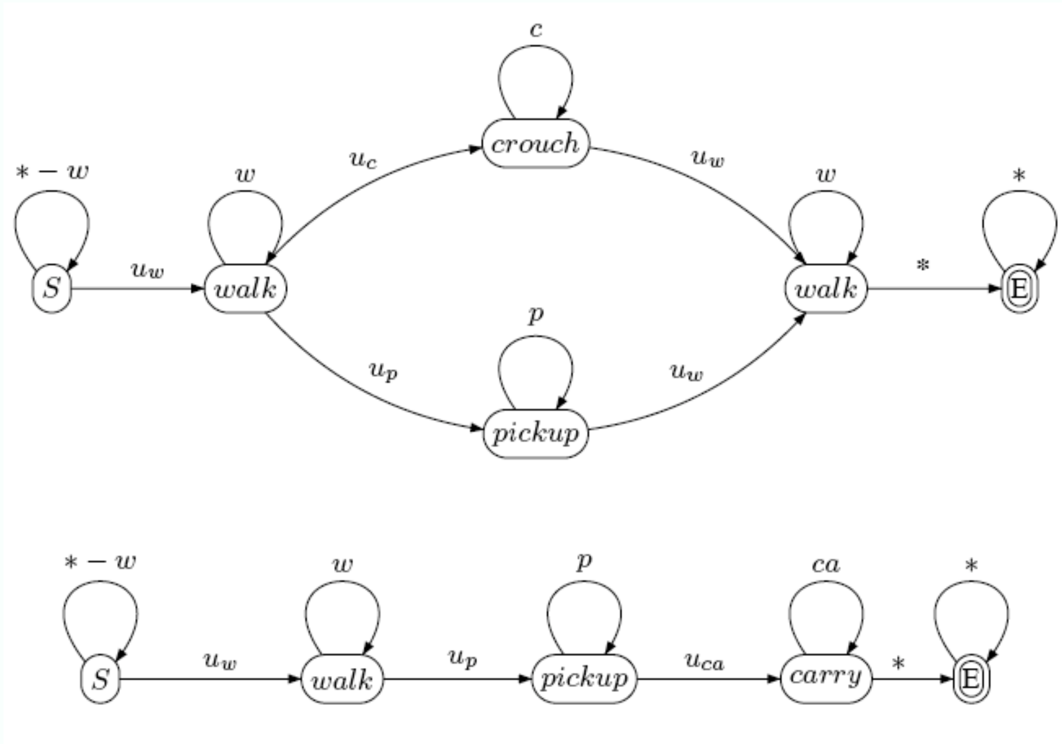
Figure 12. Confusion matrix for our action classifier. Each row shows the distribution of the true labels of the top 50 ranked examples for each action category on the validation subset of the images. Some high confusion pairs are {*reading, takingphoto*} $\rightarrow$ *playinginstrument* and *running* $\rightarrow$ *walking*.

Attributes - what is the body motion like?



Liu et al 2011

Composite reasoning is possible



Ikizler et al, 07, 08

with various architectures

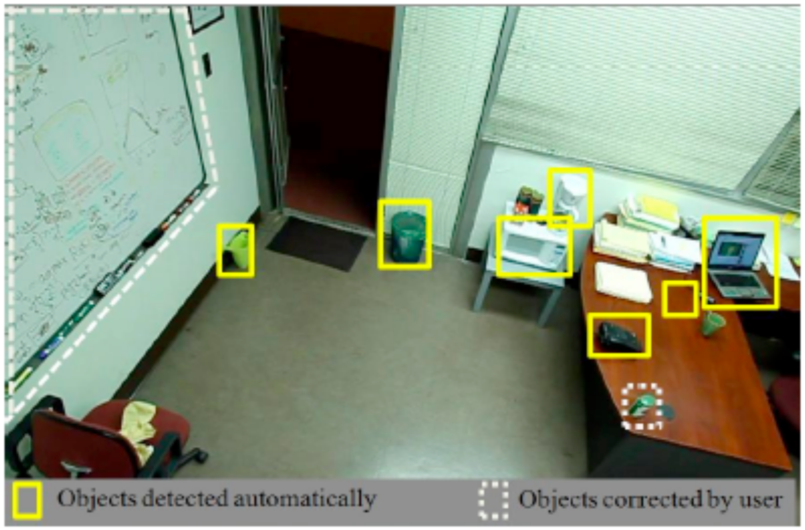


Figure 1. The Detection result of objects in an office scene, the objects of interest include cup, phone, laptop, trash-can, bucket, tea box, microwave, dispenser and white board. The tea box and the white board can not be detected automatically and are corrected by user.

Node Name	Semantic Name	Node Name	Semantic Name
a_1	arrive at phone	a_9	leave phone
a_2	arrive at trash-can	a_{10}	leave trash-can
a_3	arrive at basin	a_{11}	leave basin
a_4	arrive at dispenser	a_{12}	leave dispenser
a_5	arrive at tea box	a_{13}	leave tea box
a_6	arrive at board	a_{14}	leave board
a_7	arrive at laptop	a_{15}	leave laptop
a_8	arrive at microwave	a_{16}	leave microwave
a_{17}	use laptop	a_{18}	read paper
a_{19}	use tea box	a_{20}	use phone
a_{21}	use dispenser	a_{22}	use microwave
a_{23}	bend down	a_{24}	null
a_{25}	work	a_{26}	discuss
a_{27}	enter	a_{28}	exit

Table 1. The atomic action in the office scene which are the terminal nodes in AoG representation.

Pei et al, 11

We can detect simple named activities

BUT

Unfamiliar activities present no real problem to human observers

Unfamiliar activities present no real problem

Unfamiliar activities present no real problem to human observers

What outcome do we expect?

How are other people feeling?

What will they do?

What outcome do we expect?

How are other people feeling?

What will they do?

Narratives to explain away unfamiliar behavior

Narratives from unfamiliar behavior

What outcome do we expect?

How are other people feeling?

What will they do?

What outcome do we expect?

How are other people feeling?

What will they do?

How many adults were on the platform and what were they doing?

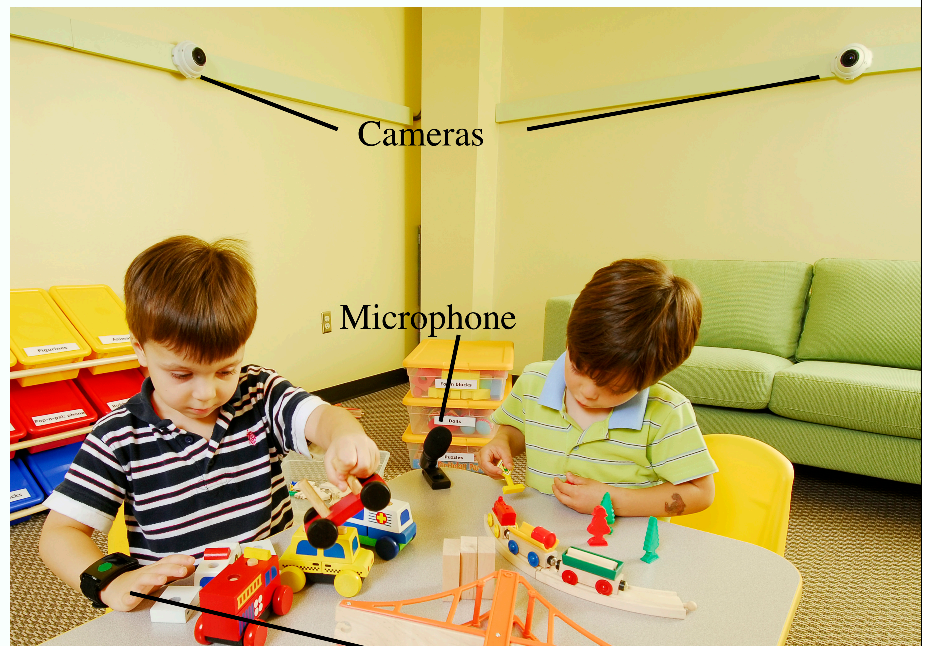
What’s going to happen to the baby?

What outcome do we expect?

How are other people feeling?

What will they do?

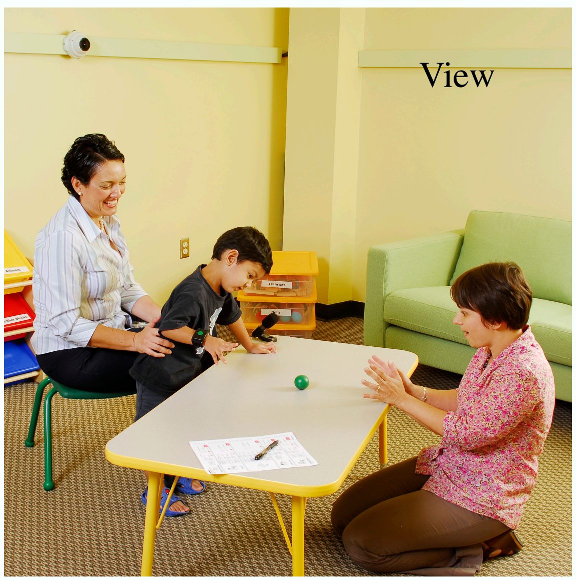
Computational Behavioural Science



- Observe people
  - Using vision, physiological markers
  - Interacting, behaving naturally
  - In the wild
- drive feedback for therapy
  - Eg reward speech
- Applications
  - Model: screen for ASD
  - Other:
    - Any where large scale observations help
      - Support in home care
      - Support care for demented patients
      - Support stroke recovery
      - Support design of efficient buildings
- 10M$, 5yr NSF award under Expeditions program
  - GaTech, UIUC(DAF, Karahalios), MIT, CMU, Pittsburgh, USC, Boston U

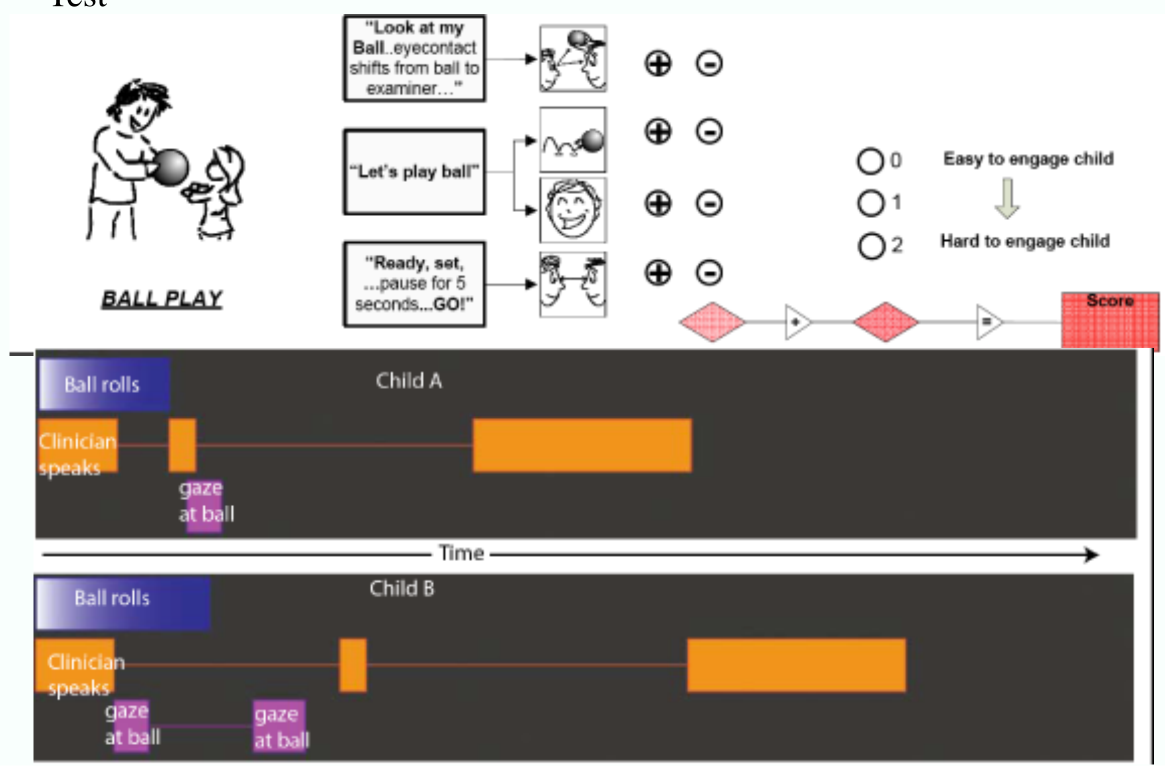
Rapid ABC

- Easily administered screening test
  - Challenge:
    - Automatic evaluation
    - To use unskilled screeners

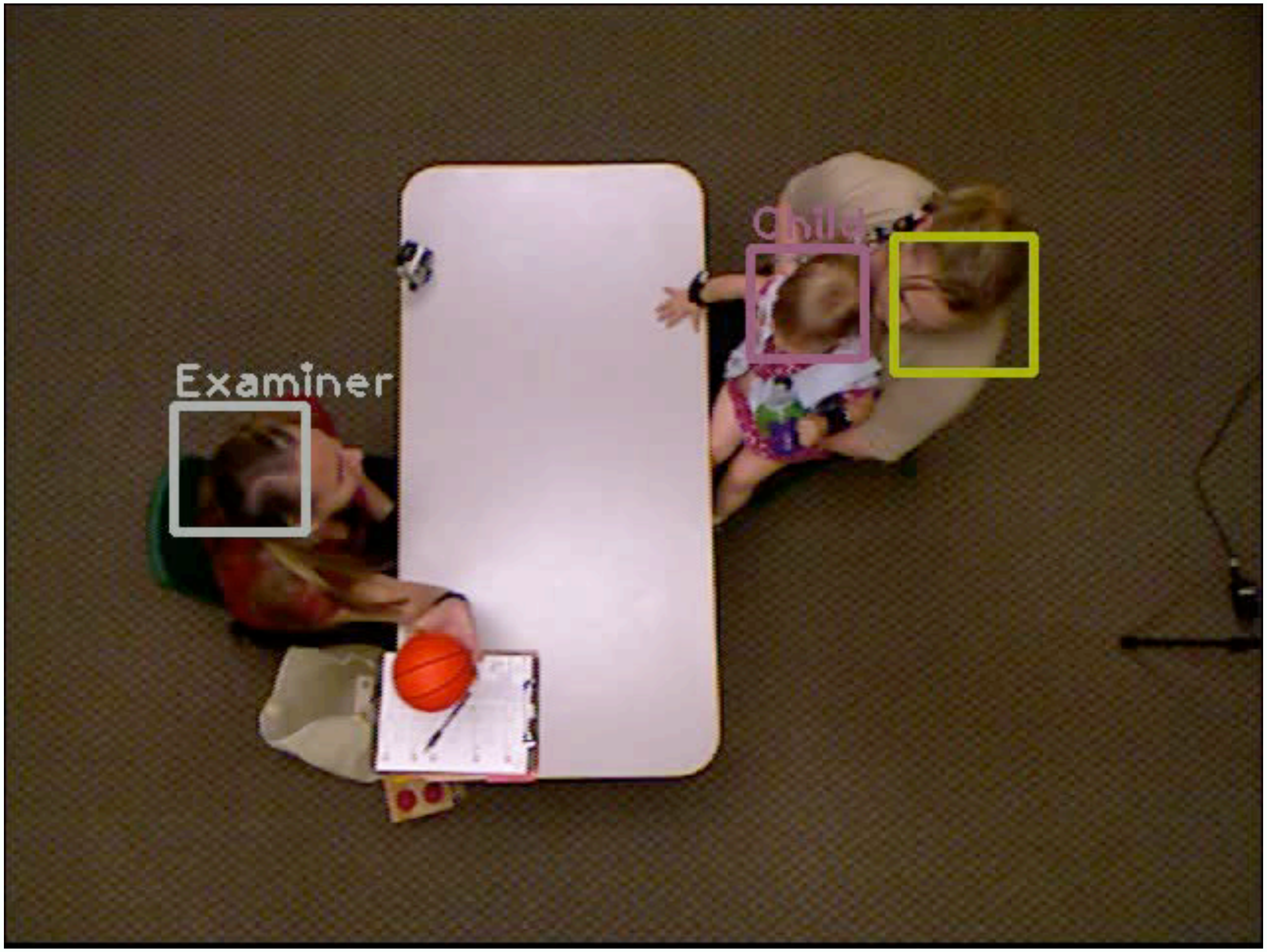


View

Test



Outcome s



LoPresti, ND

Challenge: Join up views

- Each actor has a model of what the other is up to
  - and uses it to structure what they do
- Imagine short clips of only child resp. interviewer
  - join up corresponding sides of interaction from a mess of clips
- Why
  - (a tiny bit of) Theory of mind, in concrete form

Where are their hands?

- Hands are hard
  - because they're at the far end of lower arms, and we're not good at them

Fig. 1. Sample high ranked results of person layout detection task for the VOC 2010 test dataset. The blue rectangle represents the provided bounding box of the person, green rectangles are the detected hands and red rectangle is the detected head respectively. Our method yields the best results despite high variation in pose and occlusion.

Mittal ea 12

Great hand results

Method	Head	Hand	Mean
SVM linear	73.92±3.15	20.29±1.76	47.1±1.87
Rank linear	79.32±2.77	27.88±1.75	53.6±1.28
Rank RBF	79.55±2.88	28.22±2.25	**53.9±1.29**

(a)

Method	Head	Hand	Mean
Our Method	72.85	26.7	**49.8**
BCNPCL	74.4	3.3	38.8
OXFORD	52.7	10.4	31.5

(b)

Table 3. (a) AP scores resulting from different learning techniques. The last two rows are different variants of the proposed method. They differ only slightly, but improve substantially over the SVM. The dataset used for the experiments was train-val set of the VOC 2011 layout dataset. **(b) AP for the VOC 2010 person layout test dataset.** We train our method on the train-val portion of the VOC 2010 layout dataset. The evaluation was computed on the competition server. The results for the other methods are as reported on the competition website [27]. Our result for hand detection is even better than [26], which reports AP of 23.18 for the same dataset.

Mittal ea 12

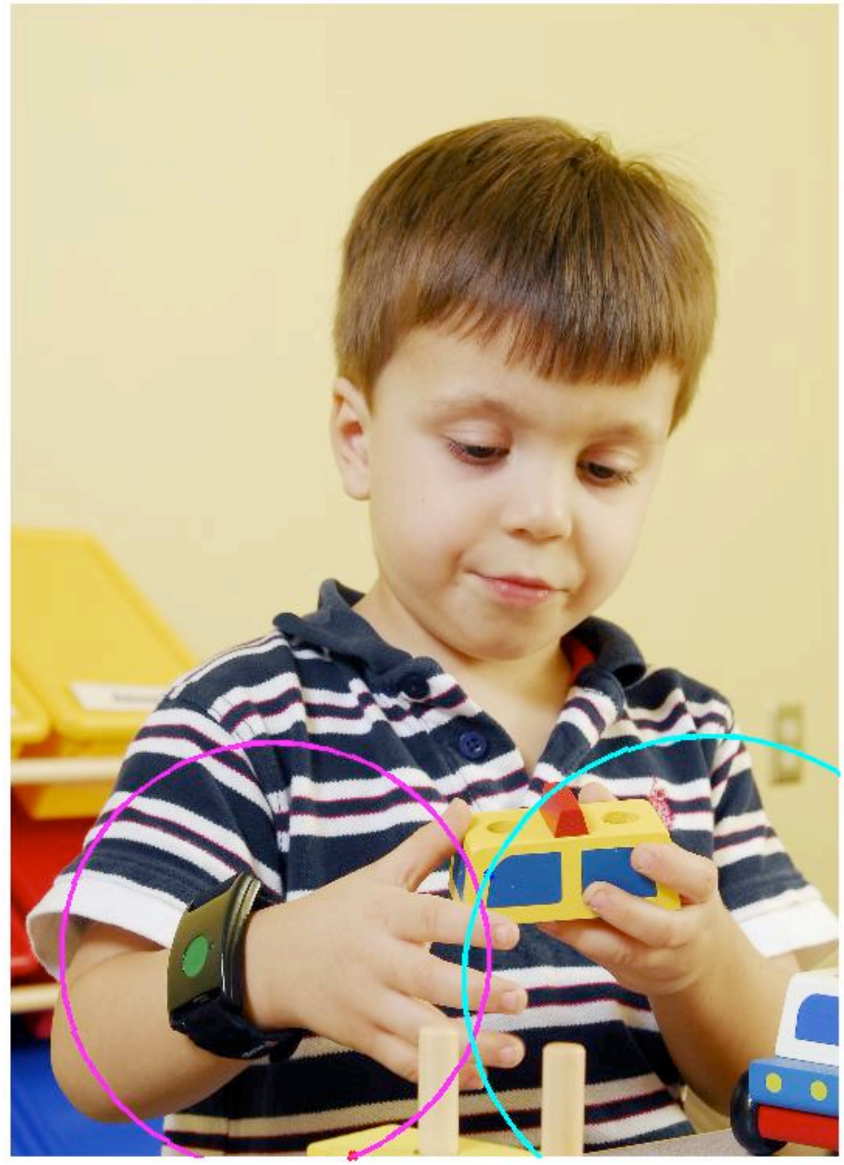
Alternate strategy

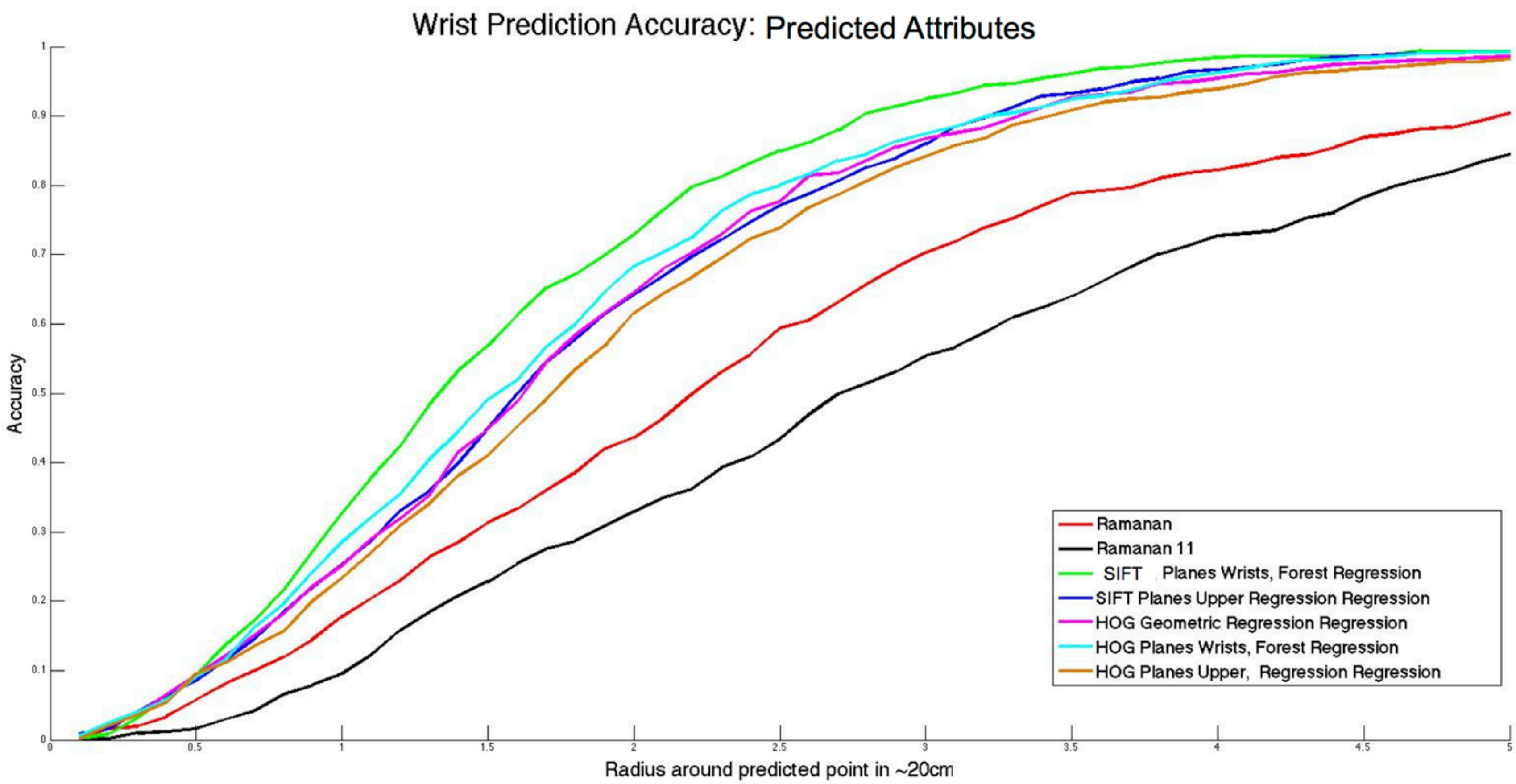
- Work up a set of body configuration attributes
  - L hand in front of plane of R arm, etc.
- Use poselets, Bourdev's data to learn predictors
- Regress hand position against attributes
  - in 3D relative to body
- Identify torso, rectify regressed hand to image

- (Coming) clean up w/ prior from Motion Capture pose

What are they doing with their hands?

Tsatsoulis, ND





Tsatsoulis, ND

Take Home

- Computer vision has extremely powerful tools
  - 3D reconstruction
  - detection
  - tracking
- But...
  - they’re not yet super reliable
  - very hard for non-specialists to use and adopt
    - problem is on the collective agenda, but unresolved
- Huge open problems on the vision agenda
  - that are problems of representation or semantics