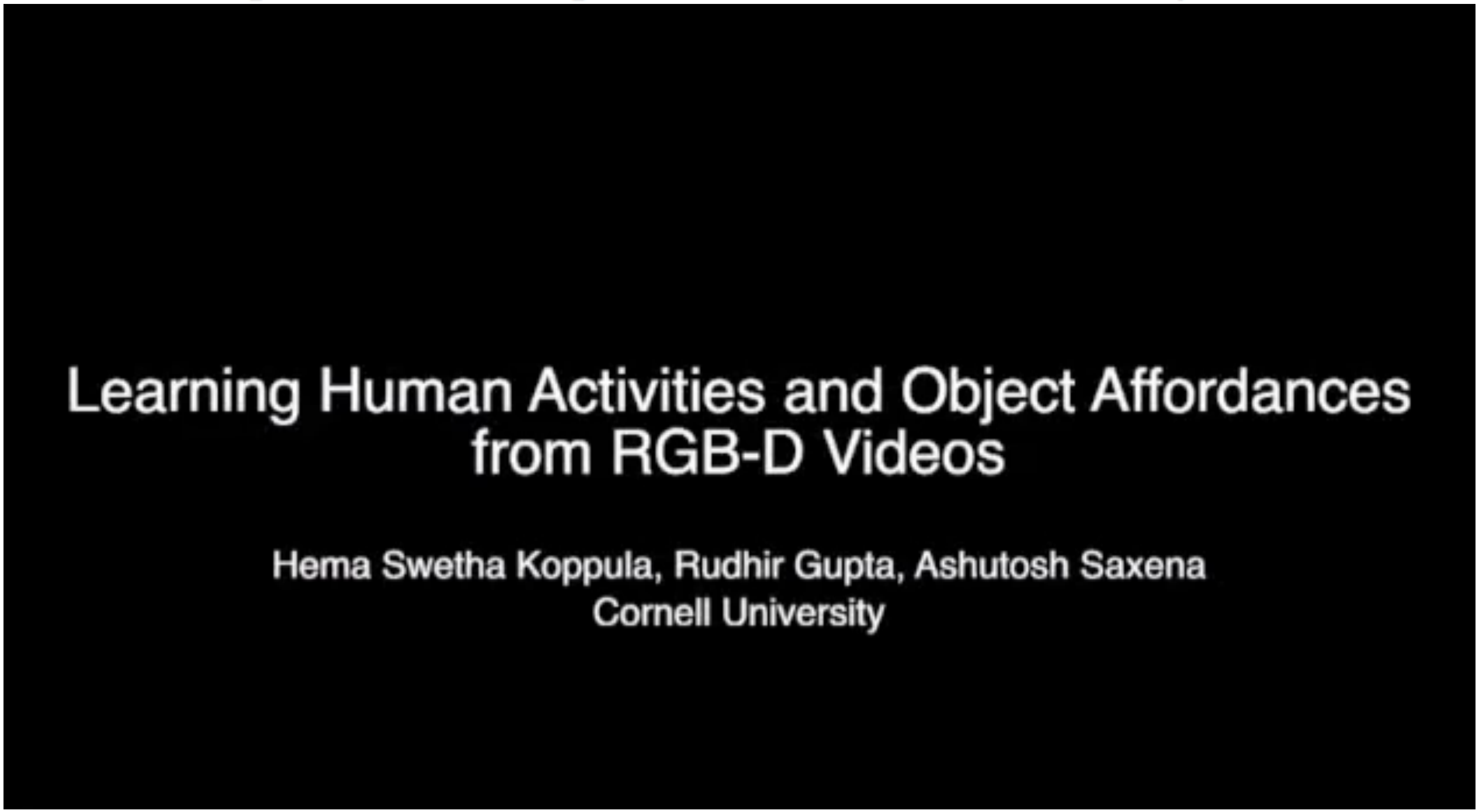
Learning Human Activities and Object Affordances from RGB-D Videos

Hema Swetha Koppula, Rudhir Gupta, Ashutosh Saxena
Department of Computer Science, Cornell University, USA.



CIRL 4/24/2013

Overview

Jointly model:

Object affordances

e.g. cup: ‘pourable’, ‘drinkable

sofa: ‘sittable’

Dynamic:

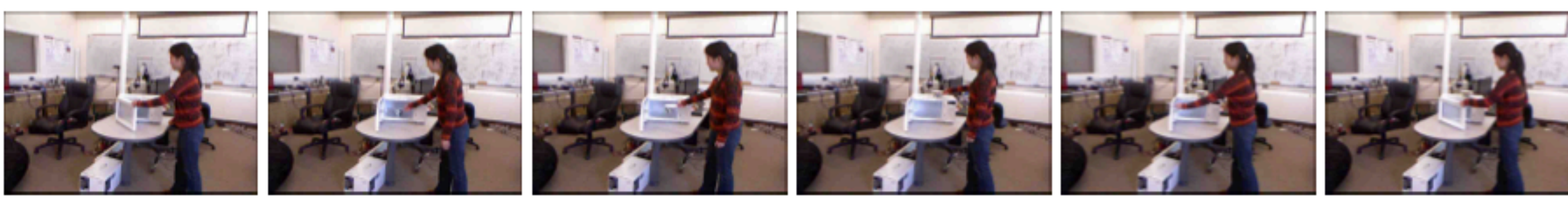
pitcher: ‘reachable’ then ‘movable’ then ‘pourable’

Sub-activities

e.g. reaching for pitcher, moving pitcher to bowl,

pouring milk into bowl

Add temporal segmentation as latent variable



Subject *opening openable* object1 | Subject *reaching reachable* object2 | Subject *moving movable* object2 | Subject *placing placable* object2 | Subject *reaching reachable* object1 | Subject *closing closable* object1

Model

Markov Random Field **[Whiteboard drawing]**

Nodes: Sub-actions + Objects

Edges: Interactions

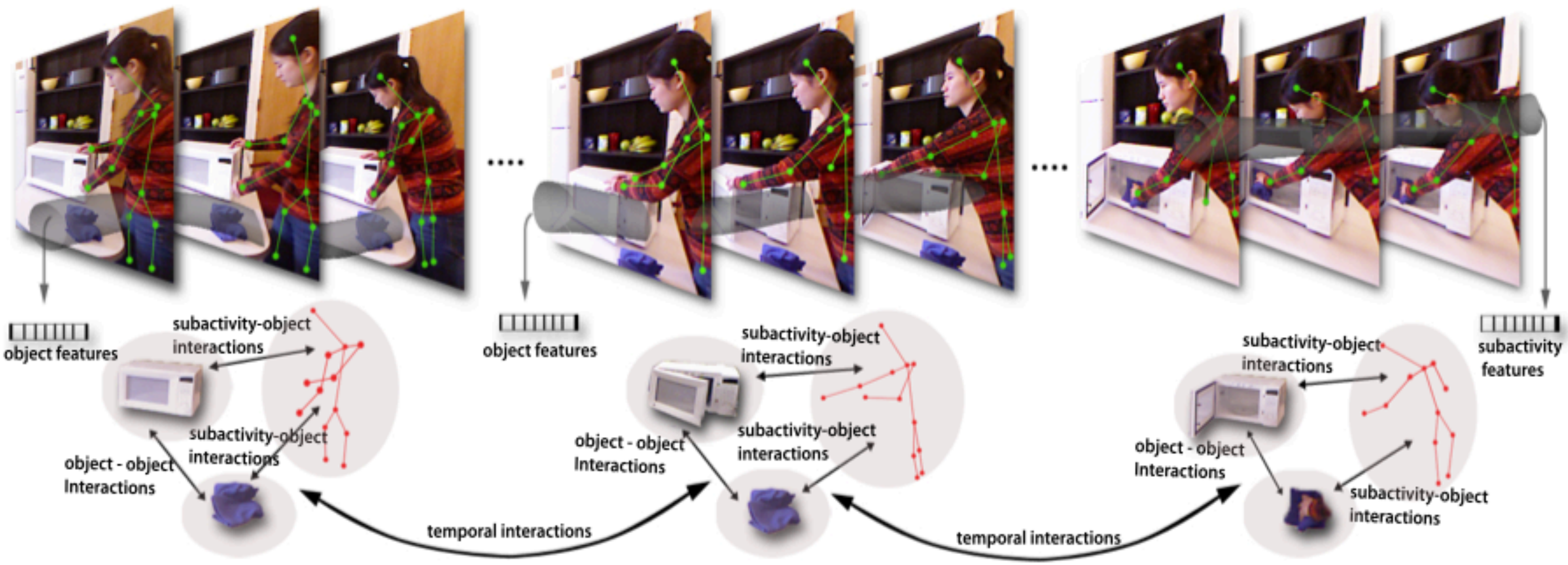
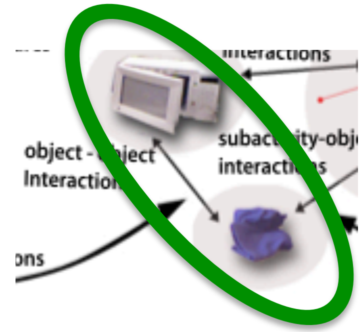


Fig. 3. Pictorial representation of the different types of nodes and relationships modeled in part of the *cleaning objects* activity comprising three sub-activities: *reaching*, *opening* and *scrubbing*. (See Section III.)

Interactions

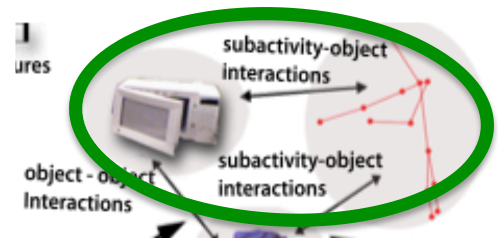
1) Affordance – affordance

(“on top of”, “in front of”)



2) Affordance – sub-activity

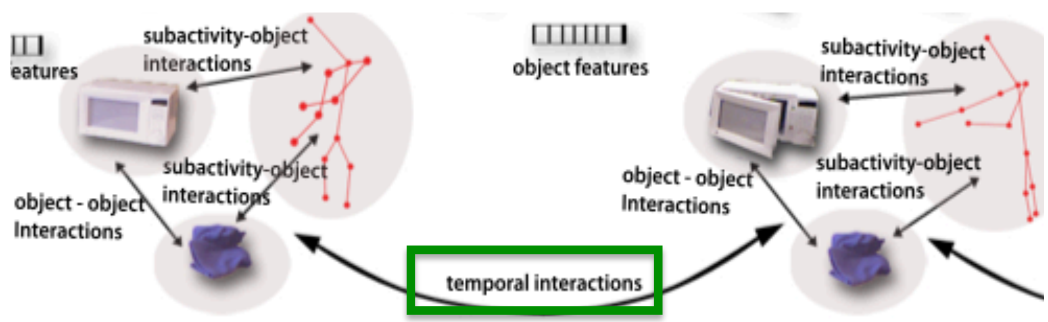
(“pour-to”, “drinkable”)



3) Affordance change over time

(f(appearance, location))

4) Sub-activity over time



Features

Each phi in energy fcn is a set of features

$$E_{oo} = \sum_{(i,j)\in\mathcal{E}_{oo}} \sum_{(l,k)\in K_o\times K_o} y_i^l y_j^k \left[\mathbf{w}_{\mathbf{oo}}^{lk} \cdot \phi_{oo}(i,j)\right],$$

$$E_{oa} = \sum_{(i,j)\in\mathcal{E}_{oa}} \sum_{(l,k)\in K_o\times K_a} y_i^l y_j^k \left[\mathbf{w}_{\mathbf{oa}}^{lk} \cdot \phi_{oa}(i,j)\right],$$

$$E_{oo}^t = \sum_{(i,j)\in\mathcal{E}_{oo}^t} \sum_{(l,k)\in K_o\times K_o} y_i^l y_j^k \left[\mathbf{w}^{t}{}_{\mathbf{oo}}^{lk} \cdot \phi_{oo}^t(i,j)\right]$$

$$E_{aa}^t = \sum_{(i,j)\in\mathcal{E}_{aa}^t} \sum_{(l,k)\in K_a\times K_a} y_i^l y_j^k \left[\mathbf{w}^{t}{}_{\mathbf{aa}}^{lk} \cdot \phi_{aa}^t(i,j)\right]$$

Description	Count
Object Features	**18**
N1. Centroid location	3
N2. 2D bounding box	4
N3. Transformation matrix of SIFT matches between adjacent frames	6
N4. Distance moved by the centroid	1
N5. Displacement of centroid	1
Sub-activity Features	**103**
N6. Location of each joint (8 joints)	24
N7. Distance moved by each joint (8 joints)	8
N8. Displacement of each joint (8 joints)	8
N9. Body pose features	47
N10. Hand position features	16
Object-object Features (computed at start frame, middle frame, end frame, max and min)	**20**
E1. Difference in centroid locations $(\Delta x, \Delta y, \Delta z)$	3
E2. Distance between centroids	1
Object–sub-activity Features (computed at start frame, middle frame, end frame, max and min)	**40**
E3. Distance between each joint location and object centroid	8
Object Temporal Features	**4**
E4. Total and normalized vertical displacement	2
E5. Total and normalized distance between centroids	2
Sub-activity Temporal Features	**16**
E6. Total and normalized distance between each corresponding joint locations (8 joints)	16

Object detection

-SVM on color histogram, HOGs, VFH

-kNN on VFH

Train on set of potential objects (e.g. mugs, cups)

RGB-D object dataset

Procedure:

1) Look only around the users hands
2) Run SVM on color data
3) For all with SVM(·)>Thresh:

   Calculate kNN for VFH

   Shrink box around local peak in kNN score

Tracking

Run particle filter on detections with high likelihood

Only do detection every N frames

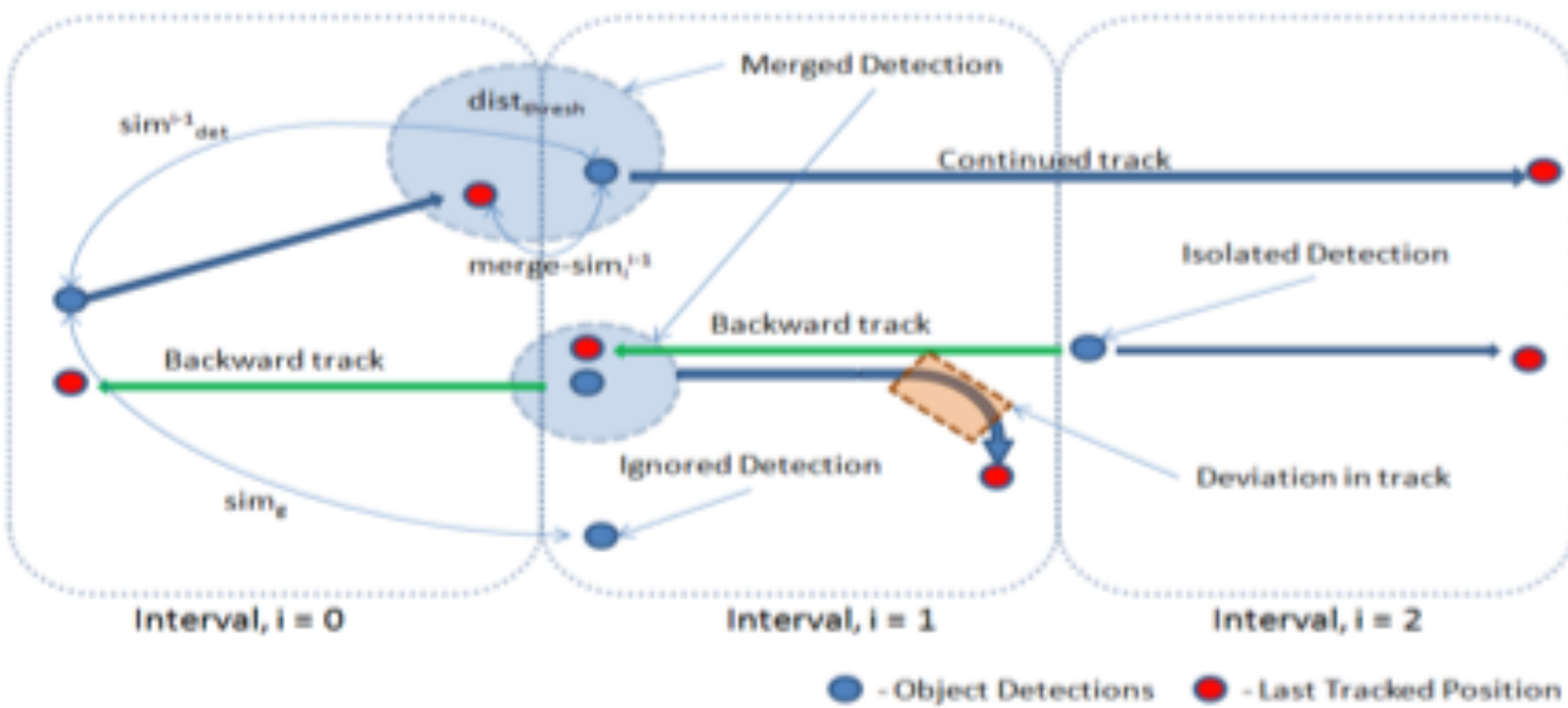


Fig. 4. Pictorial representation of our algorithm for combining object detections with tracking.

Sub-actions

TABLE II

DESCRIPTION OF ACTIVITIES IN TERMS OF SUB-ACTIVITIES. NOTE THAT SOME ACTIVITIES CONSIST OF SAME SUB-ACTIVITIES BUT ARE EXECUTED IN DIFFERENT ORDER.

	reaching	moving	placing	opening	closing	eating	drinking	pouring	scrubbing	null
Making Cereal	✓	✓	✓					✓		✓
Taking Medicine	✓	✓	✓	✓		✓	✓			✓
Stacking Objects	✓	✓	✓							✓
Unstacking Objects	✓	✓	✓							✓
Microwaving Food	✓	✓	✓	✓	✓					✓
Picking Objects	✓	✓								✓
Cleaning Objects	✓	✓		✓	✓				✓	✓
Taking Food	✓		✓	✓	✓					✓
Arranging Objects	✓	✓	✓							✓
Having a Meal	✓	✓				✓	✓			✓

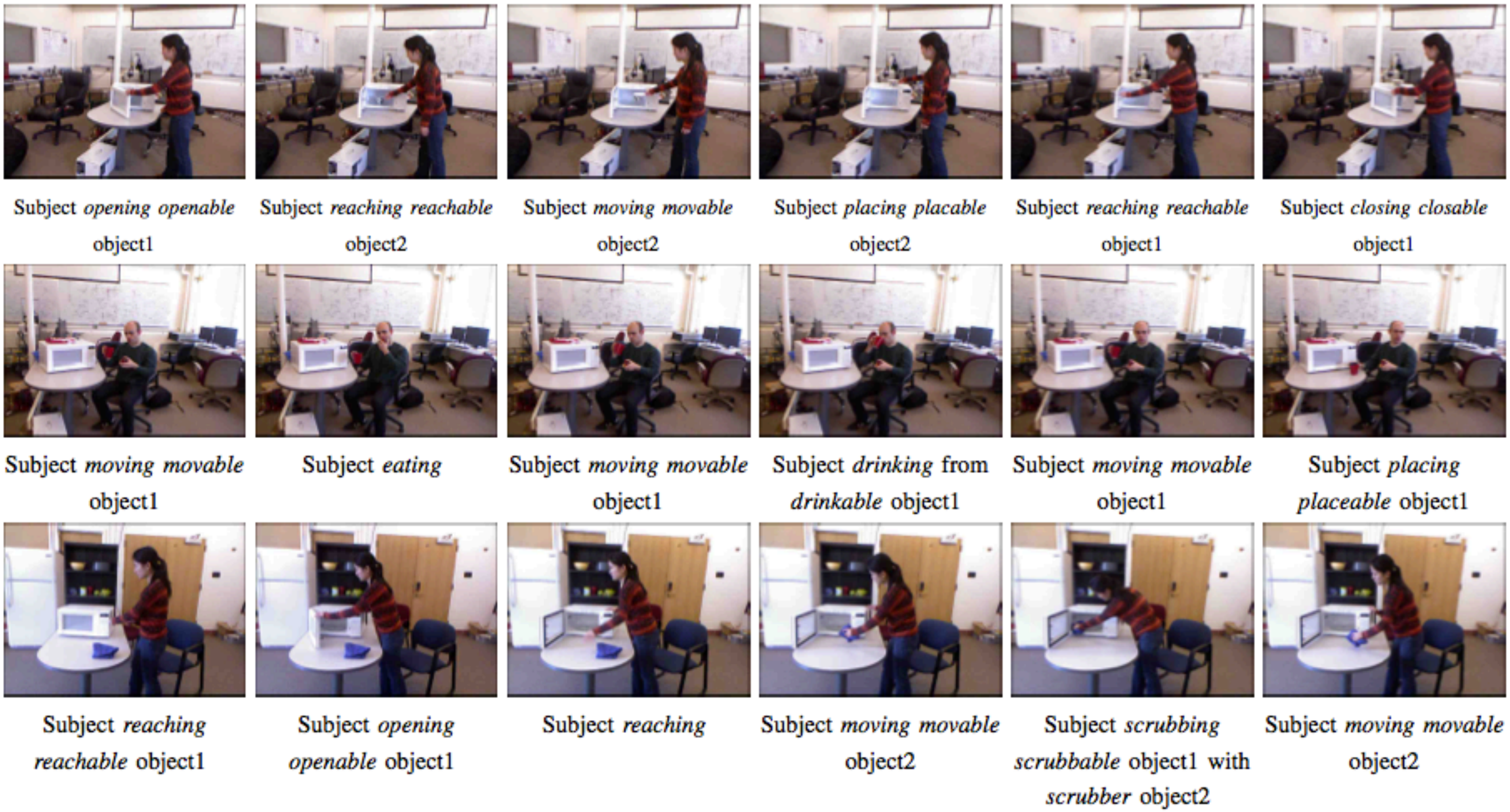


Fig. 8. Descriptive output of our algorithm: Sequence of images from the *taking food* (Top Row), *having meal* (Middle Row) and *cleaning objects* (Bottom Row) activities labeled with sub-activity and object affordance labels. A single frame is sampled from the temporal segment to represent it.

Temporal Segmentation

Try 3 methods:

1) Uniform lengths

Graph methods (Felzenszwalb and Huttenlocher):

2) edges: sum of Euclidean distances between skeleton joints

3) edges: rate of change of skeleton joints

High Level Activity

Features = Histograms of sub-activity, affordance labels

Use multi-class SVM

This has problems with similar actions

(e.g. stacking objects and unstacking objects)

Inference

Mixed integer programming solver

w = model parameters

y = label

x = data

z = $y_i^l y_j^k$

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}} \max_{\mathbf{z}} \sum_{i \in \mathcal{V}_a} \sum_{k \in K_a} y_i^k \left[\mathbf{w}_\mathbf{a}^k \cdot \phi_a(i) \right] + \sum_{i \in \mathcal{V}_o} \sum_{k \in K_o} y_i^k \left[\mathbf{w}_\mathbf{o}^k \cdot \phi_o(i) \right]$$
$$+ \sum_{t \in \mathcal{T}} \sum_{(i,j) \in \mathcal{E}_t} \sum_{(l,k) \in T_t} z_{ij}^{lk} \left[\mathbf{w}_\mathbf{t}^{lk} \cdot \phi_t(i,j) \right] \qquad (12)$$
$$\forall i,j,l,k: \; z_{ij}^{lk} \le y_i^l, \; z_{ij}^{lk} \le y_j^k, \; y_i^l + y_j^k \le z_{ij}^{lk} + 1, \; z_{ij}^{lk}, y_i^l \in \{0,1\}$$

Learning

Structural SVM

$$\begin{aligned} \min_{w,\xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\xi \qquad (14) \\ s.t. \quad & \forall \bar{\mathbf{y}}_1, ..., \bar{\mathbf{y}}_M \in \{0, 0.5, 1\}^{N\cdot K}: \\ & \frac{1}{M}\mathbf{w}^T \sum^{M} [\Psi(\mathbf{x}_m, \mathbf{y}_m) - \Psi(\mathbf{x}_m, \bar{\mathbf{y}}_m)] \geq \Delta(\mathbf{y}_m, \bar{\mathbf{y}}_m) - \xi \end{aligned}$$

$$\bar{\mathbf{y}}_m = \operatorname*{argmax}_{\mathbf{y}\in\{0,0.5,1\}^{N\cdot K}} \left[\mathbf{w}^T\Psi(\mathbf{x}_m, \mathbf{y}) + \Delta(\mathbf{y}_m, \mathbf{y})\right].$$

Results

Results on our CAD-120 dataset, SHOWING AVERAGE MICRO PRECISION/RECALL, AND AVERAGE MACRO PRECISION AND RECALL FOR AFFORDANCE, SUB-ACTIVITIES AND HIGH-LEVEL ACTIVITIES. STANDARD ERROR IS ALSO REPORTED.

	Object Affordance			Sub-activity			High-level Activity		
	micro	macro		micro	macro		micro	macro	
method	P/R	Prec.	Recall	P/R	Prec.	Recall	P/R	Prec.	Recall
max class	65.7 ± 1.0	65.7 ± 1.0	8.3 ± 0.0	29.2 ± 0.2	29.2 ± 0.2	10.0 ± 0.0	10.0 ± 0.0	10.0 ± 0.0	10.0 ± 0.0
image only	74.2 ± 0.7	15.9 ± 2.7	16.0 ± 2.5	56.2 ± 0.4	39.6 ± 0.5	41.0 ± 0.6	34.7 ± 2.9	24.2 ± 1.5	35.8 ± 2.2
SVM multiclass	75.6 ± 1.8	40.6 ± 2.4	37.9 ± 2.0	58.0 ± 1.2	47.0 ± 0.6	41.6 ± 2.6	30.6 ± 3.5	27.4 ± 3.6	31.2 ± 3.7
MEMM (Sung et al., 2012)	-	-	-	-	-	-	26.4 ± 2.0	23.7 ± 1.0	23.7 ± 1.0
object only	86.9 ± 1.0	72.7 ± 3.8	63.1 ± 4.3	-	-	-	59.7 ± 1.8	56.3 ± 2.2	58.3 ± 1.9
sub-activity only	-	-	-	71.9 ± 0.8	60.9 ± 2.2	51.9 ± 0.9	27.4 ± 5.2	31.8 ± 6.3	27.7 ± 5.3
no temporal interactions	87.0 ± 0.8	79.8 ± 3.6	66.1 ± 1.5	76.0 ± 0.6	74.5 ± 3.5	66.7 ± 1.4	81.4 ± 1.3	83.2 ± 1.2	80.8 ± 1.4
no object interactions	88.4 ± 0.9	75.5 ± 3.7	63.3 ± 3.4	85.3 ± 1.0	79.6 ± 2.4	74.6 ± 2.8	80.6 ± 2.6	81.9 ± 2.2	80.0 ± 2.6
full model: groundtruth seg	91.8 ± 0.4	90.4 ± 2.5	74.2 ± 3.1	86.0 ± 0.9	84.2 ± 1.3	76.9 ± 2.6	84.7 ± 2.4	85.3 ± 2.0	84.2 ± 2.5
full model: groundtruth seg + tracking	88.2 ± 0.6	74.5 ± 4.3	64.9 ± 3.5	82.5 ± 1.4	72.9 ± 1.2	70.5 ± 3.0	79.0 ± 4.7	78.6 ± 4.1	78.3 ± 4.9
Full model. End-to-end results, *without* assuming any ground-truth temporal segmentation is given.									
full, 1 segment. (best)	83.1 ± 1.1	70.1 ± 2.3	63.9 ± 4.4	66.6 ± 0.7	62.0 ± 2.2	60.8 ± 4.5	77.5 ± 4.1	80.1 ± 3.9	76.7 ± 4.2
full, 1 segment. (averaged)	81.3 ± 0.4	67.8 ± 1.1	60.0 ± 0.8	64.3 ± 0.7	63.8 ± 1.1	59.1 ± 0.5	79.0 ± 0.9	81.1 ± 0.8	78.3 ± 0.9
full, multi-seg learning	83.9 ± 1.5	75.9 ± 4.6	64.2 ± 4.0	68.2 ± 0.3	71.1 ± 1.9	62.2 ± 4.1	80.6 ± 1.1	81.8 ± 2.2	80.0 ± 1.2
full, multi-seg learning + tracking	79.4 ± 0.8	62.5 ± 5.4	50.2 ± 4.9	63.4 ± 1.6	65.3 ± 2.3	54.0 ± 4.6	75.0 ± 4.5	75.8 ± 4.4	74.2 ± 4.6

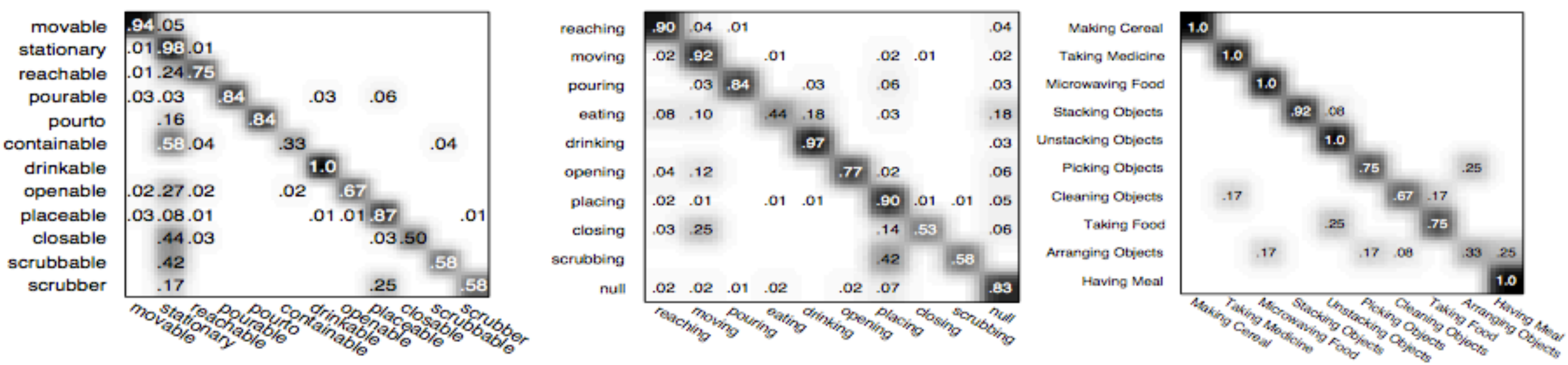


Fig. 7. Confusion matrix for affordance labeling (left), sub-activity labeling (middle) and high-level activity labeling (right) of the test RGB-D videos.

Results

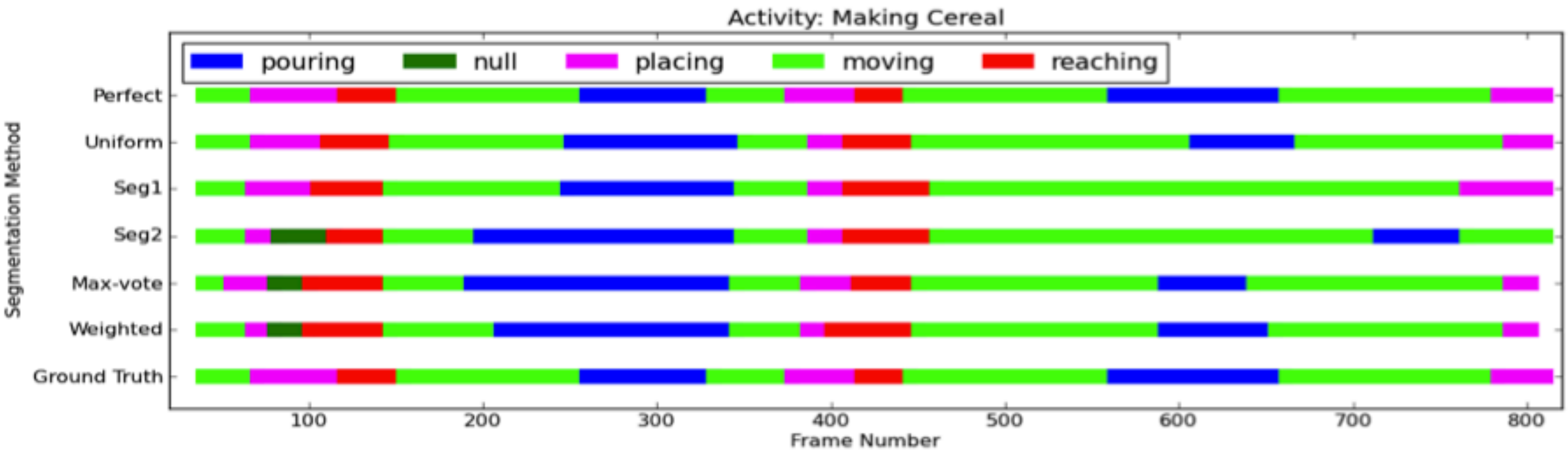


Fig. 9. Comparison of the sub-activity labeling of various segmentations. This activity involves the sub-activities: *reaching*, *moving*, *pouring* and *placing* as colored in red, green, blue and magenta respectively. The x-axis denotes the time axis numbered with frame numbers. It can be seen that the various individual segmentation labelings are not perfect and make different mistakes, but our method for merging these segmentations selects the correct label for many frames.

New person

TABLE IV

RESULTS ON CORNELL ACTIVITY DATASET (SUNG ET AL., 2012), TESTED ON *"New Person"* DATA FOR 12 ACTIVITY CLASSES.

	bathroom		bedroom		kitchen		living room		office		Average	
	prec	rec	prec	rec	prec	rec	prec	rec	prec	rec	prec	rec
Sung et al. (2012)	72.7	**65.0**	**76.1**	59.2	64.4	47.9	52.6	45.7	73.8	59.8	67.9	55.5
Our method	**88.9**	61.1	73.0	**66.7**	**96.4**	**85.4**	**69.2**	**68.7**	**76.7**	**75.0**	**80.8**	**71.4**

Results

1) Extra

3D Local

Skeletal F

TABLE III

Object Tracking Results

	≥40%	≥20%	≥10%
tracking w/o detection	49.2	65.7	75
tracking + detection	53.5	69.4	77.8

2) Combine features

3) Look at different time scales

4) Comb

5) MKL c

TABLE VI

Robot Object Manipulation Results

task	# instance	accuracy	accuracy (multi. obvs.)
object movement	19	100	100
constrained movement	15	80	100

Output sequences

Fig. 2. **Significant Variations, Clutter and Occlusions**: Example shots of *reaching* sub-activity from our dataset. First and third rows show the RGB images, and the second and bottom rows show the corresponding depth images from the RGB-D camera. Note that there are significant variations in the way the subjects perform the sub-activity. In addition, there is significant background clutter and subjects are partially occluded (e.g., column 1) or not facing the camera (e.g., row 1 column 4) in many instances.