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
Model

Markov Random Field

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

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

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

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

\[
E_{aa}^t = \sum_{(i,j) \in \mathcal{E}_{aa}^t} \sum_{l,k \in K_a \times K_a} y_i^l y_j^k \left[ w_{aa}^{tlk} \cdot \phi_{aa}^t(i,j) \right].
\]
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.
<table>
<thead>
<tr>
<th>Activity</th>
<th>reaching</th>
<th>moving</th>
<th>placing</th>
<th>opening</th>
<th>closing</th>
<th>eating</th>
<th>drinking</th>
<th>pouring</th>
<th>scrubbing</th>
<th>null</th>
</tr>
</thead>
<tbody>
<tr>
<td>Making Cereal</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Taking Medicine</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Stacking Objects</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Unstacking Objects</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Microwaving Food</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Picking Objects</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Cleaning Objects</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Taking Food</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Arranging Objects</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Having a Meal</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
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

\[ w = \text{model parameters} \]

\[ y = \text{label} \]

\[ x = \text{data} \]

\[ z = y_i^l y_j^k \]

\[
\hat{y} = \arg\max_y \max_z \sum_{i \in V_a} \sum_{k \in K_a} y_i^k \left[ w_a^k \cdot \phi_a(i) \right] + \sum_{i \in V_o} \sum_{k \in K_o} y_i^k \left[ w_o^k \cdot \phi_o(i) \right] \\
+ \sum_{t \in T} \sum_{(i,j) \in E_t} \sum_{(l,k) \in T_t} z_{ij}^{lk} \left[ w_t^{lk} \cdot \phi_t(i, j) \right]
\]

\( \forall i, j, l, k: \ z_{ij}^{lk} \leq y_i^l, \ z_{ij}^{lk} \leq y_j^k, \ y_i^l + y_j^k \leq z_{ij}^{lk} + 1, \ z_{ij}^{lk}, y_i^l \in \{0, 1\} \)
Learning

Structural SVM

\[
\begin{aligned}
\min_{w, \xi} & & \frac{1}{2} w^T w + C\xi \\
\text{s.t.} & & \forall \bar{y}_1, \ldots, \bar{y}_M \in \{0, 0.5, 1\}^{N \cdot K} : \\
& & \frac{1}{M} w^T \sum_{M} [\Psi(x_m, y_m) - \Psi(x_m, \bar{y}_m)] \geq \Delta(y_m, \bar{y}_m) - \xi \\
\end{aligned}
\]

\[
\bar{y}_m = \arg\max_{y \in \{0, 0.5, 1\}^{N \cdot K}} \left[ w^T \Psi(x_m, y) + \Delta(y_m, y) \right].
\]
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Micro Affordance</th>
<th>Macro Affordance</th>
<th>Micro Sub-Activity</th>
<th>Macro Sub-Activity</th>
<th>Micro High-Level Activity</th>
<th>Macro High-Level Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P/R</td>
<td>Prec.</td>
<td>Recall</td>
<td>P/R</td>
<td>Prec.</td>
<td>Recall</td>
</tr>
<tr>
<td>max class</td>
<td>65.7 ± 1.0</td>
<td>65.7 ± 1.0</td>
<td>8.3 ± 0.0</td>
<td>29.2 ± 0.2</td>
<td>29.2 ± 0.2</td>
<td>10.0 ± 0.0</td>
</tr>
<tr>
<td>image only</td>
<td>74.2 ± 0.7</td>
<td>15.9 ± 2.7</td>
<td>16.0 ± 2.5</td>
<td>56.2 ± 0.4</td>
<td>39.6 ± 0.5</td>
<td>41.0 ± 0.6</td>
</tr>
<tr>
<td>SVM multiclass</td>
<td>75.6 ± 1.8</td>
<td>40.6 ± 2.4</td>
<td>37.9 ± 2.0</td>
<td>58.0 ± 1.2</td>
<td>47.0 ± 0.6</td>
<td>41.6 ± 2.6</td>
</tr>
<tr>
<td>MEMM [Sung et al., 2012]</td>
<td>86.9 ± 1.0</td>
<td>72.7 ± 3.8</td>
<td>63.1 ± 4.3</td>
<td>71.9 ± 0.8</td>
<td>60.9 ± 2.2</td>
<td>51.9 ± 0.9</td>
</tr>
<tr>
<td>object only</td>
<td>88.4 ± 0.9</td>
<td>79.5 ± 3.7</td>
<td>66.1 ± 1.5</td>
<td>85.3 ± 1.0</td>
<td>79.6 ± 2.4</td>
<td>74.6 ± 2.8</td>
</tr>
<tr>
<td>sub-activity only</td>
<td>87.0 ± 0.8</td>
<td>79.8 ± 3.6</td>
<td>66.1 ± 1.5</td>
<td>76.0 ± 0.6</td>
<td>74.5 ± 3.5</td>
<td>66.7 ± 1.4</td>
</tr>
<tr>
<td>no temporal interactions</td>
<td>89.8 ± 0.4</td>
<td>83.5 ± 3.4</td>
<td>63.3 ± 3.4</td>
<td>91.8 ± 0.4</td>
<td>89.4 ± 2.5</td>
<td>74.2 ± 3.1</td>
</tr>
<tr>
<td>full model: groundtruth seg</td>
<td>88.2 ± 0.6</td>
<td>74.5 ± 4.3</td>
<td>64.9 ± 3.5</td>
<td>82.5 ± 1.4</td>
<td>72.9 ± 1.2</td>
<td>70.5 ± 3.0</td>
</tr>
<tr>
<td>full model: groundtruth seg + tracking</td>
<td>91.8 ± 0.4</td>
<td>90.4 ± 2.5</td>
<td>74.2 ± 3.1</td>
<td>86.0 ± 0.9</td>
<td>84.2 ± 2.5</td>
<td>76.9 ± 2.6</td>
</tr>
</tbody>
</table>

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
- full, 1 segment. (averaged) 81.3 ± 0.4 67.8 ± 1.1 60.0 ± 0.8
- full, multi-seg learning 83.9 ± 1.5 75.9 ± 4.6 64.2 ± 4.0
- full, multi-seg learning + tracking 79.4 ± 0.8 62.5 ± 5.4 50.2 ± 4.9

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.
**TABLE IV**

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

<table>
<thead>
<tr>
<th></th>
<th>bathroom</th>
<th></th>
<th>bedroom</th>
<th></th>
<th>kitchen</th>
<th></th>
<th>living room</th>
<th></th>
<th>office</th>
<th></th>
<th>Average</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>prec</td>
<td>rec</td>
<td>prec</td>
<td>rec</td>
<td>prec</td>
<td>rec</td>
<td>prec</td>
<td>rec</td>
<td>prec</td>
<td>rec</td>
<td>prec</td>
<td>rec</td>
</tr>
<tr>
<td>Sung et al. (2012)</td>
<td>72.7</td>
<td>65.0</td>
<td>76.1</td>
<td>59.2</td>
<td>64.4</td>
<td>47.9</td>
<td>52.6</td>
<td>45.7</td>
<td>73.8</td>
<td>59.8</td>
<td>67.9</td>
<td>55.5</td>
</tr>
<tr>
<td>Our method</td>
<td>88.9</td>
<td>61.1</td>
<td>73.0</td>
<td>66.7</td>
<td>96.4</td>
<td>85.4</td>
<td>69.2</td>
<td>68.7</td>
<td>76.7</td>
<td>75.0</td>
<td>80.8</td>
<td>71.4</td>
</tr>
</tbody>
</table>
Results

1) Extract features:
   - 3D Local
   - Skeletal Features

2) Combine features

3) Look at different time scales

4) Combine top joint features

5) MKL classification

<table>
<thead>
<tr>
<th>TABLE III</th>
<th>OBJECT TRACKING RESULTS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>≥40%</td>
</tr>
<tr>
<td>tracking w/o detection</td>
<td>49.2</td>
</tr>
<tr>
<td>tracking + detection</td>
<td>53.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE VI</th>
<th>ROBOT OBJECT MANIPULATION RESULTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>task</td>
<td># instance</td>
</tr>
<tr>
<td>object movement</td>
<td>19</td>
</tr>
<tr>
<td>constrained movement</td>
<td>15</td>
</tr>
</tbody>
</table>
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.