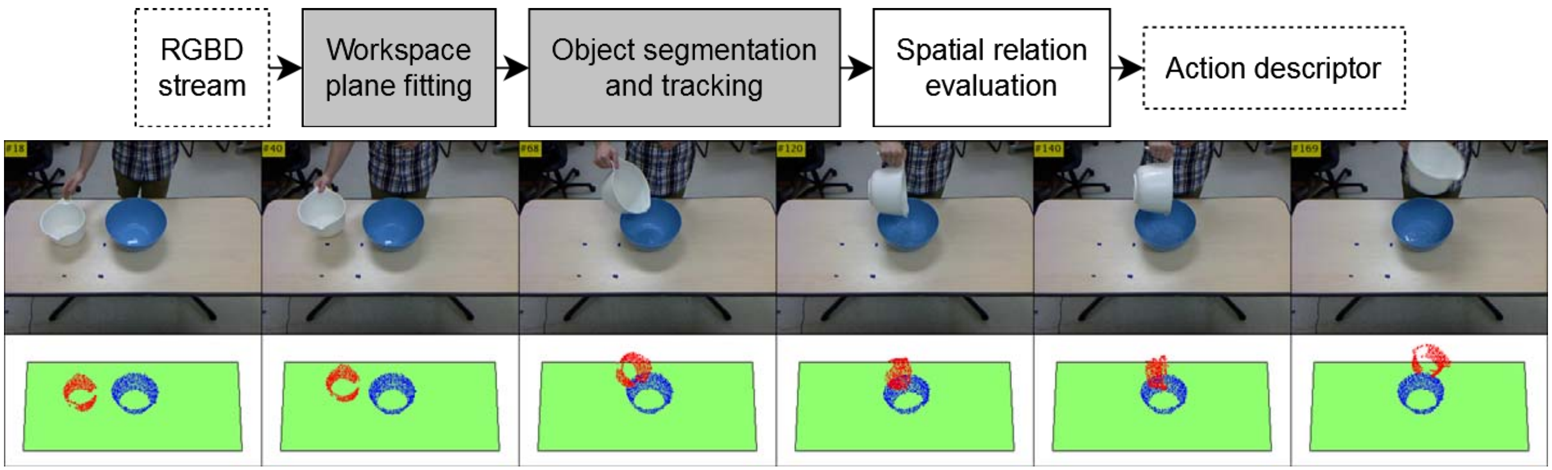
Learning the Spatial Semantics of Manipulation Actions through Preposition Grounding

Konstantinos Zampogiannis, Yezhou Yang, Cornelia Fermüller and Yiannis Aloimonos

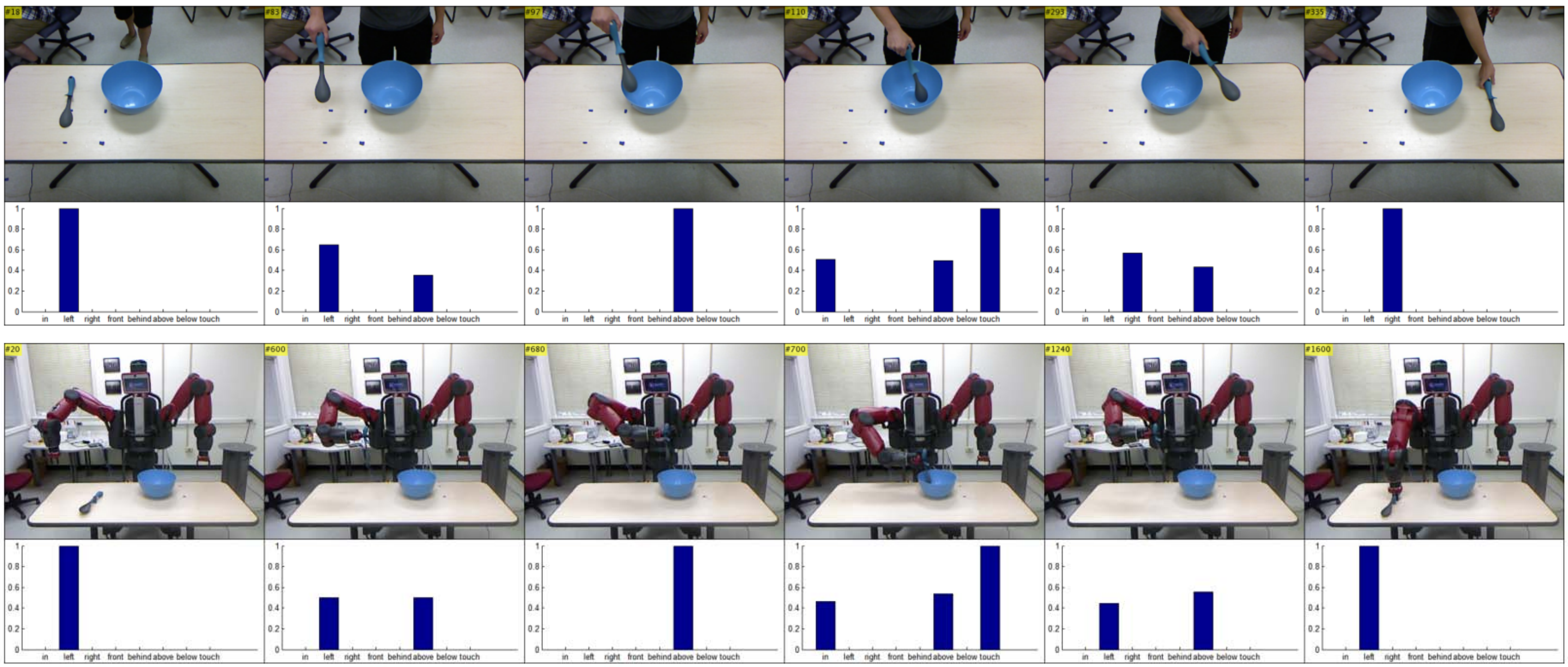
Overview

- We propose an abstract action representation that captures the temporal evolution of spatial pairwise object relations
- Processing steps (pipeline):



- Given the tracked point clouds for all objects involved in the manipulation:
  - A set of spatial relation predicates are evaluated for all object pairs at all video frames
  - Action descriptors are built upon spatial *Predicate Vector Sequences* (PVS)
- An appropriate *time-normalized* distance measure for our representation is introduced

Temporal evolution of spatial relations



Spatial relations evolution: ladle relative to bowl for two instances of `Stir`.

Spatial relation grounding: relative spaces

- Align sensor frame xyz with workspace plane normal:

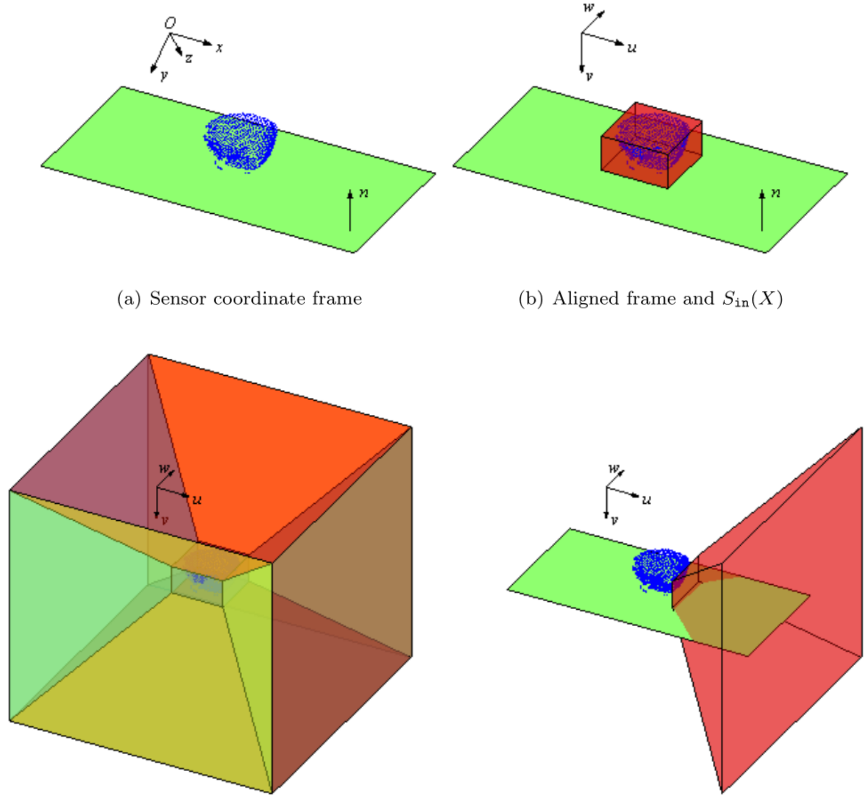
$$\hat{v} = \operatorname{sgn}(\hat{y} \cdot \hat{n})\,\hat{n}$$
$$\hat{w} = (\hat{z} - (\hat{v} \cdot \hat{z})\,\hat{v}) \,/\, \|\hat{z} - (\hat{v} \cdot \hat{z})\,\hat{v}\|$$
$$\hat{u} = \hat{v} \times \hat{w}$$

- Aligned frame uvw captures six basic directions:

Spatial relation defining directions

Direction	left	right	front	behind	above	below
Reference vector	$-\hat{u}$	$+\hat{u}$	$-\hat{w}$	$+\hat{w}$	$-\hat{v}$	$+\hat{v}$

- uvw-aligned bounding box for object X (blue point cloud) models object interior space $S_{\mathtt{in}}(X)$
- Seven spaces $S_r(X)$ are defined *relative* to object X, for $r \in$ `{in, left, right, front, behind, below, above}`



(a) Sensor coordinate frame

(b) Aligned frame and $S_{\mathtt{in}}(X)$

(c) All 7 relative spaces

(d) $S_{\mathtt{right}}(X)$

Spatial relation grounding: predicates

- We define real-valued predicates $R_r(X,Y)$, for all relations $r \in \mathcal{R}^S = \{$`in`, `left`, `right`, `front`, `behind`, `below`, `above`$\}$, simply as the *fraction* of object X that lies in $S_r(Y)$:

$$R_r(X,Y) = \frac{|X \cap S_r(Y)|}{|X|}$$

  The set of $R_r(X,Y)$, for all r, gives the spatial distribution of X relative to Y.

- Binary-valued contactual predicate (touch relation):

$$R_{\texttt{touch}}(X,Y) = \begin{cases} 1 & \text{if } R_{\texttt{in}}(X,Y) > 0 \\ & \text{or } R_{\texttt{in}}(Y,X) > 0 \\ & \text{or } d_m < d_T \\ 0 & \text{otherwise} \end{cases}$$

  where d_m is the linear SVM margin between point sets X and Y

- $\mathcal{R}^f = \{$`in`, `left`, `right`, `front`, `behind`, `below`, `above`, `touch`$\}$ is the full set of our modeled spatial relations

- **Spatial abstraction:** `left`, `right`, `front` and `behind` relations mostly capture viewpoint-specific information and may depend on execution-specific object arrangements, while having little to do with the manipulation semantics. We combine them into a new relation (`around`):

$$\begin{aligned} R_{\texttt{around}}(X,Y) = & R_{\texttt{left}}(X,Y) + R_{\texttt{right}}(X,Y) + \\ & R_{\texttt{front}}(X,Y) + R_{\texttt{behind}}(X,Y) \end{aligned}$$

- $\mathcal{R}^a = \{$`in`, `around`, `below`, `above`, `touch`$\}$ is the set of relations we will use to build our **action descriptors**!

Action descriptors

- Let $\Phi^t(i,j)$ be the *predicate vector* for all relations in $\mathcal{R}^a$ between objects with indices i and j at time t, where $i,j = 1,\dots,N_o$:

$$\Phi^t(i,j) \equiv \big(R_{\mathtt{in}}(X_i^t, X_j^t), R_{\mathtt{around}}(X_i^t, X_j^t), R_{\mathtt{below}}(X_i^t, X_j^t), R_{\mathtt{above}}(X_i^t, X_j^t), R_{\mathtt{touch}}(X_i^t, X_j^t)\big)$$

- We will call the temporal sequence of $\Phi^t(i,j)$, for $t = 1,\dots,T$, the *Predicate Vector Sequence* (PVS) for object pair (i,j):

$$\Phi(i,j) \equiv \big(\Phi^1(i,j), \dots, \Phi^T(i,j)\big)$$

- Our action descriptor will be an ordered set of the PVSes for all $N_r = N_o(N_o - 1)$ *ordered* object pairs, arranged in a *known* order imposed by function I_{N_o}:

$$A \equiv (\Phi_1, \dots, \Phi_{N_r})$$

  where, for $k = 1,\dots,N_r$, $\Phi_k \equiv \Phi(i,j)$ and $(i,j) = I_{N_o}(k)$. Function I_{N_o} can be any bijection from $\{1,\dots,N_r\}$ to the set of all ordered object pairs.

Pairwise distance function

- Calculating the distance between action descriptors A^1 (N_o^1 objects, N_r^1 PVSes) and A^2 (N_o^2 objects, N_r^2 PVSes) is based on finding an optimal **object correspondence** between them
- An object correspondence is encoded in a $N_o^1 \times N_o^2$ binary assignment matrix $X = (x_{ij})$: $x_{ij} = 1$ if and only if object i in A^1 is matched to object j in A^2
- Object correspondence X induces a **PVS correspondence** $Y_X = (y_{r^1 r^2})$ ($N_r^1 \times N_r^2$ binary matrix)
- The **cost** of assignment is then:

$$J(Y_X) = \sum_{r^1=1}^{N_r^1} \sum_{r^2=1}^{N_r^2} c_{r^1 r^2} y_{r^1 r^2}$$

where $c_{r^1 r^2}$ is the **Dynamic Time Warping** (DTW) distance between PVS r^1 in A^1 and PVS r^2 in A^2:

$$c_{r^1 r^2} = \mathrm{DTW}(\Phi^1_{r^1}, \Phi^2_{r^2})$$

- PVS r^1 in A^1 refers to object pair:

$$(o_1^1, o_2^1) = I_{N_o^1}(r^1)$$

PVS r^2 in A^2 refers to object pair:

$$(o_1^2, o_2^2) = I_{N_o^2}(r^2)$$

Clearly, $y_{r^1 r^2} = 1$ if and only if o_1^1 is mapped to o_1^2 and o_2^1 to o_2^2, or:

$$y_{r^1 r^2} = x_{o_1^1 o_1^2} x_{o_2^1 o_2^2}$$

- Binary Quadratic Program

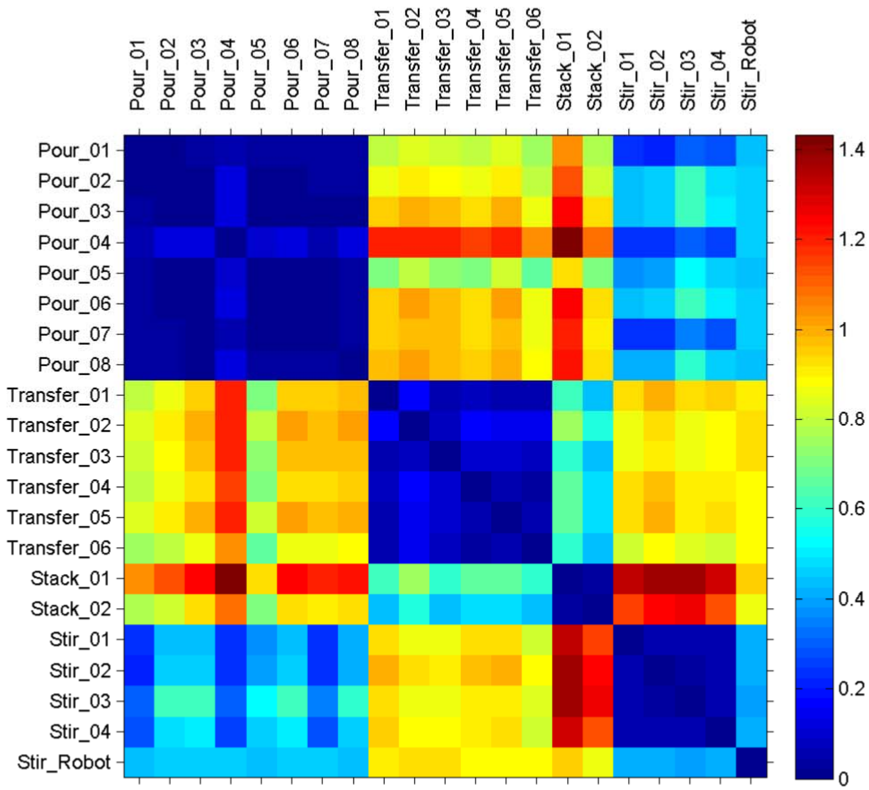
$$\begin{array}{ll}
\underset{X}{\text{Minimize}} & J(X) = \sum_{r^1=1}^{N_r^1} \sum_{r^2=1}^{N_r^2} c_{r^1 r^2} x_{o_1^1 o_1^2} x_{o_2^1 o_2^2} \\
\text{where} & (o_1^1, o_2^1) = I_{N_o^1}(r^1),\ (o_1^2, o_2^2) = I_{N_o^2}(r^2) \\
\text{subject to} & \sum_{j=1}^{N_o^2} x_{ij} \le 1, \quad i = 1, \ldots, N_o^1 \\
& \sum_{i=1}^{N_o^1} x_{ij} \le 1, \quad j = 1, \ldots, N_o^2 \\
& \sum_{i=1}^{N_o^1} \sum_{j=1}^{N_o^2} x_{ij} = \min(N_o^1, N_o^2) \\
& x_{ij} \in \{0, 1\}, \quad i \in \{1, \ldots, N_o^1\} \\
& \qquad\qquad\quad\ \ j \in \{1, \ldots, N_o^2\}
\end{array}$$

- Distance value:

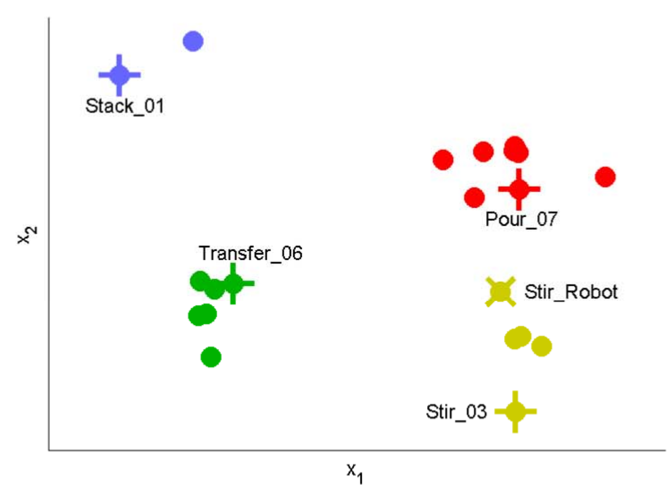
$$d(A^1, A^2) = \min_X (J(X))$$

Case study: unsupervised clustering

- 20+1 action executions in 4 semantic classes (`Pour`, `Transfer`, `Stack`, `Stir`)
  - Different performers, initial/final object arrangements, significant timing variations
- We used Affinity Propagation, with similarities $s_{ij} = -d_{ij}$ and uniform preferences for all points (equal to the median of all similarities).
  - Correct number of clusters was returned and there were no classification errors



(a) Pairwise distances matrix $D = (d_{ij})$, where $d_{ij} = d(A^i, A^j)$.



(b) Embedding of our action descriptors in 2 dimensions, based on our distance measure. Different colors correspond to different clusters, as returned by Affinity Propagation, and cluster representatives are marked by crosses.

Conclusions and future work

- The evolution of pairwise spatial relations between objects is very descriptive of the high-level manipulation semantics
- Online action matching:
  - Robot control policy
  - Prediction
- Our descriptors can be part of a more *complete*, multi-layered action representation