Recognizing Interactions: A First-Person Perspective

Alireza Fathi and James M. Rehg

Georgia Institute of Technology

Georgia Tech College of Computing
School of Interactive Computing

Expedition in Computing

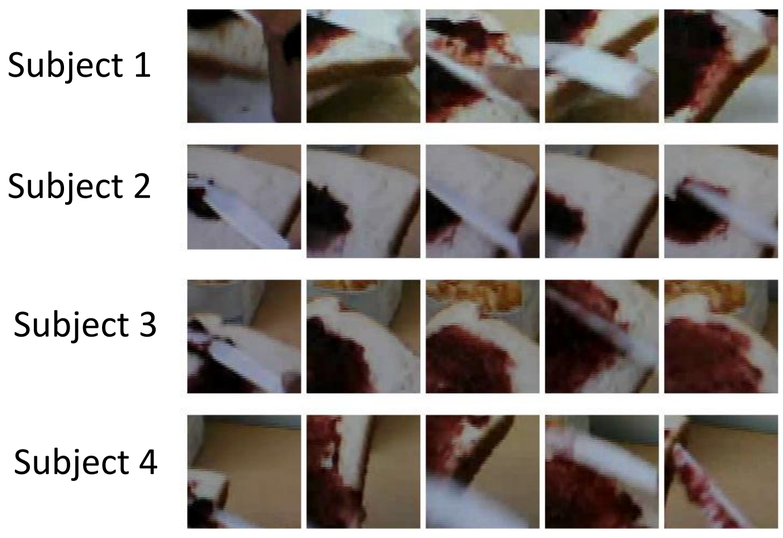
NSF

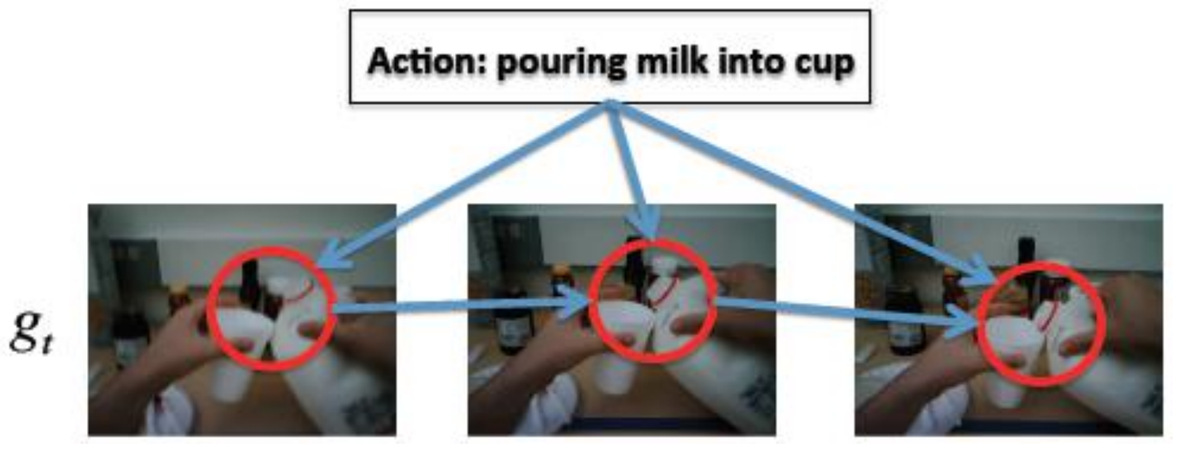
Motivation

- Humans interact with their surrounding environment at every moment in their lives.
- A common type of interaction consists of tasks which involve manipulation and movement of objects.
- Another popular kind of interaction involves social activities.
- Our goal is to detect these behaviors in day-long videos of individuals using a wearable camera system.

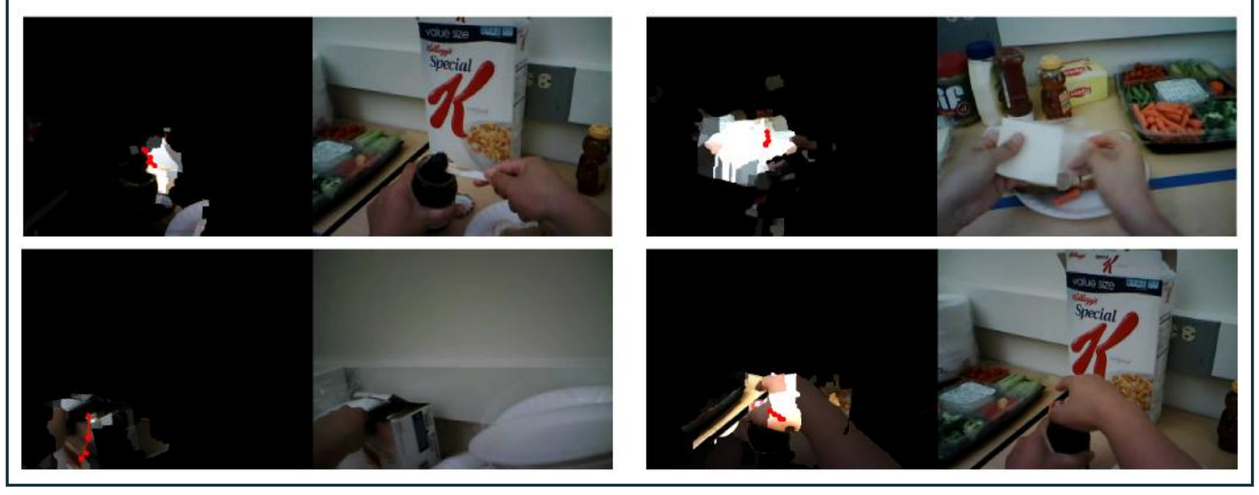
Object-Manipulation Tasks

- A small window of pixels extracted from around gaze is:
  (1) very informative on what the action is
  (2) appears consistently similar among different instances of the same action performed by different subjects

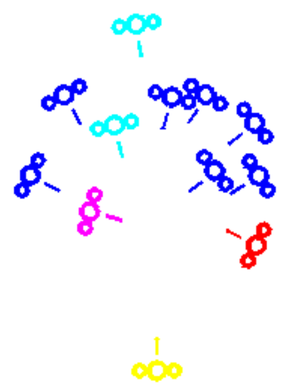




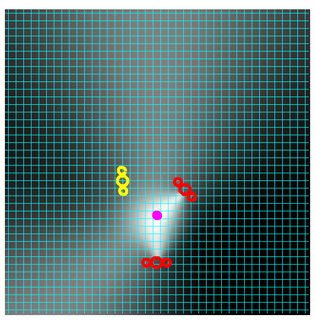
- We use three sets of features for each pixel location in image: Object-based features, Appearance Features, and future manipulation features.

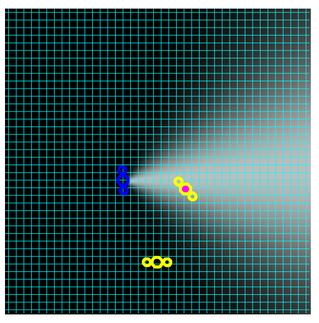


Social Interactions

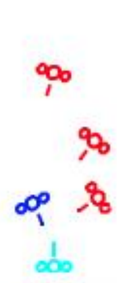


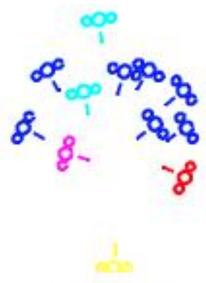
- Face location in image → Angle from first-person view
- Face size in image → Distance from first-person

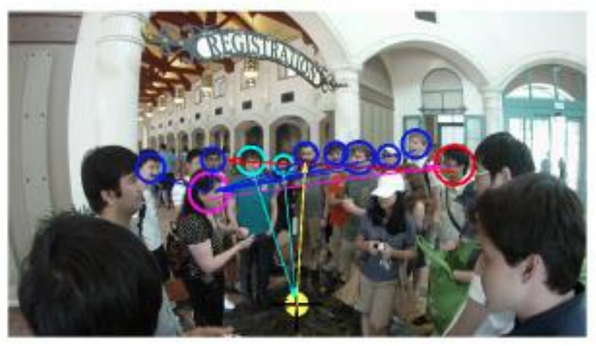




- We discretize the space into a grid.
- Our goal is to estimate at which grid point each face is looking.
- People more likely look at where their head is oriented to.
- People more likely look at faces rather than random locations.
- People more likely look at what others are looking at.







- Roles are assigned to an individual (x) based on the following four features:
  - Number of faces looking at x
  - Whether first-person looks at x
  - If there is mutual attention between first-person and x
  - Number of faces looking at where x is attending
- We learn a HCRF on top of attention, location and head-motion features.

$$\Psi(y, \mathbf{h}, \mathbf{x}; w) = \sum_{i=1}^{n} w_{h_i} . \varphi_{x_i} + \sum_{i=1}^{n} w_{y,h_i} + \sum_{(k,l) \in E} w_{y,h_k,h_l}$$

- Confusion matrix for recognizing different types of social interaction.

- We can build a social network as an immediate result of first-person vision
- Our algorithm simply counts the number of times each face cluster appears in a subject's video