Using a Minimal Action Grammar for Activity Understanding in the Real World

Douglas Summers-Stay, Ching L. Teo, Yezhou Yang, Cornelia Fermuller and Yiannis Aloimonos

University of Maryland, College Park
Why a grammar for activities?

• How do humans come to understand, recognize, and replicate complex activities?

• Each observed instance of an activity is unique in terms of
  – Order
  – Limb motion
  – Appearance

• Somehow, the sensory data must be stored in a greatly compressed representation that captures relevant information

• Must be capable of handling actions of any complexity, where activities are composed of previously known actions and subactions

• Suggests that the brain uses a similar method for understanding both language and actions
Basic principle

• Many human activities consist of multiple actions, which themselves may be broken down into sub-actions
• We want to
  – Decide which sub-actions are part of a larger activity
  – Recognize activities when the sub-actions may be performed in various orders
Forming an activity tree

- Form a tree where two objects merge whenever they begin or stop co-moving
- Search tree for subtrees to recognize activities
- for example, making a closed sandwich involves
  - some kind of bread
  - adding any of several possible ingredients in any order
  - adding another piece of bread (which may itself have condiments)
Events Detected

- Hand grasps bagel
- Bagel and plate touch
- Hand touches knife
- Knife touches cheese
- Knife touches bagel

Activity Tree Formation

- bagel
- plate
- knife
- cheese

[Diagram showing the sequence of events and corresponding activity tree]
How we recognize merge events

- Collect 3D pointcloud
- Recognize hand location from Kinect skeleton
- Extract objects
- Recognize when objects come into contact with hands or each other
The trees allow us to distinguish what parts of a video belong to which activities, even when the activities interrupt each other.
Results

The Tree Edit Distance between activity trees provides a measure of similarity that can be used to recognize activities.