Towards a Watson That Sees: Language-Guided Action Recognition for Robots

Ching L. Teo, Yezhou Yang, Hal Daumé III, Cornelia Fermüller
and
Yiannis Aloimonos

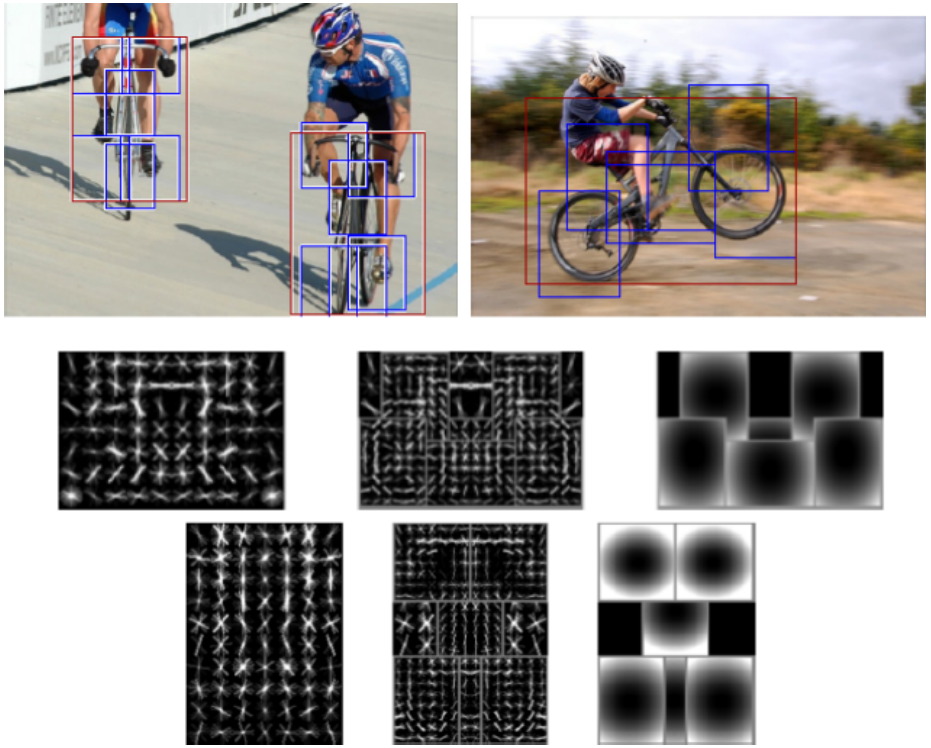
University of Maryland Institute for Advanced Computer Studies

Apr 9, 2012

Outline

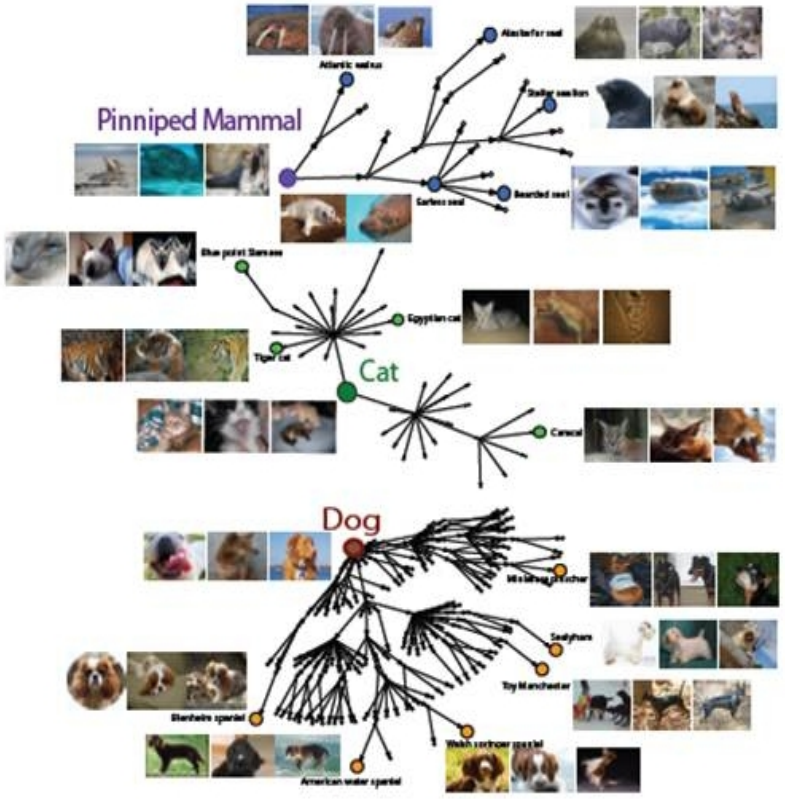
Motivation

- Computer Vision seeks to achieve *perception* through visual signals.
- Current approaches to perception: extract edges, texture, motion, color, surfaces.
- What comes next?



Felzenszwalb, P. et al, *Object detection with discriminatively trained part-based models*, PAMI 2010

- This is not enough: high-level knowledge is required.
- High-level knowledge: semantics, function, attributes, associations.
- Enables *reasoning* from perception.
- Approach: use *language* as a resource of high-level information to solve problems in Computer Vision.



Jia Deng et al, *ImageNet: A Large-Scale Hierarchical Image Database*, CVPR 2009

Why Language?

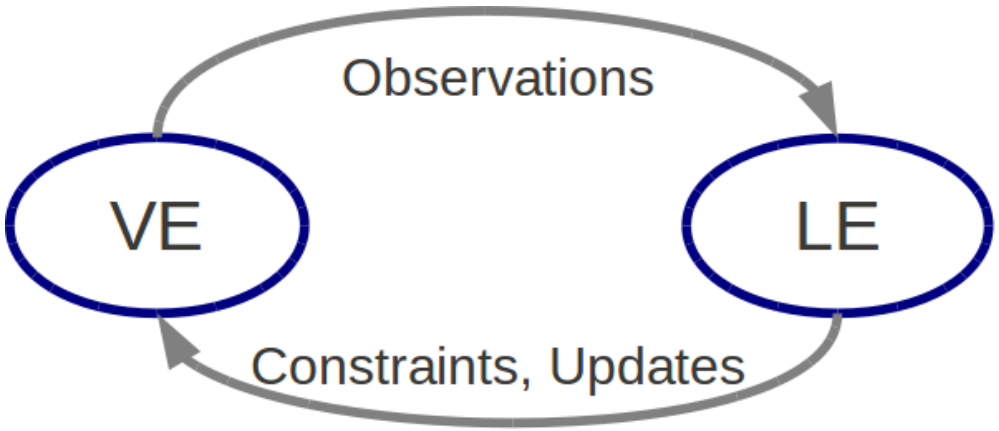
Language provides:

1. A lexicon (dictionary) that encodes contextual relationship between entities: e.g. ladle $\xrightarrow{\text{occurs}}$ kitchen.
2. Prior knowledge: e.g. knife $\xrightarrow{\text{cuts}}$ cucumber.
3. Generalizes to beyond what is "seen" (non-visual): e.g. sun, beaches, people $\xrightarrow{\text{relates}}$ vacation.

Computational Linguistics have provided tools and relational databases that encodes relationships such as: is-a, cause-effect, performs-functions, motivated by etc.

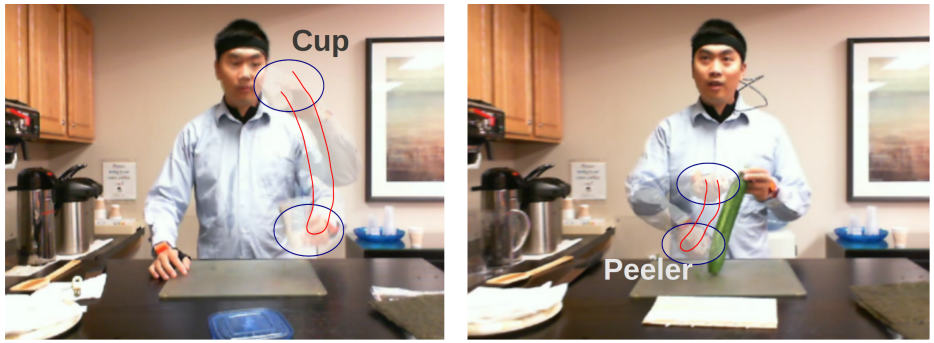
Approach: The Cognitive Dialog

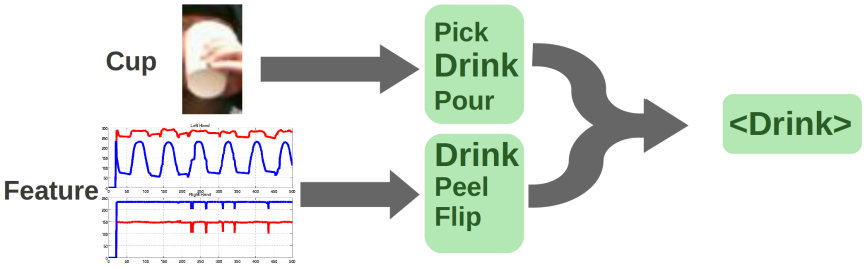
- A model of a reasoning process that involves the Visual Executive (VE) and Language Executive (LE).
- VE provides: observations, low-level feature extraction.
- LE provides: constraints, plausibility of observations from VE.
- Each iteration of the "dialog" seeks to optimize a global function related to the task: e.g. scene/object/action recognition, object segmentation.



Action Recognition is Hard

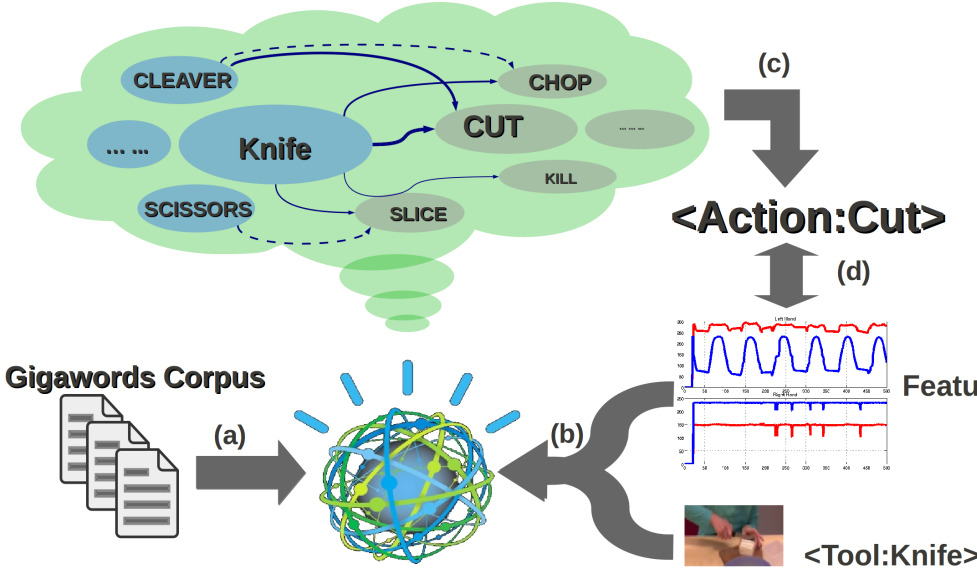
- Goal: Recognizing various actions involving hand-tools.
- Challenge: ambiguity of actions simply from the trajectories of the hand alone.
- Key idea is to use language as source of prior knowledge that relates *tools* with the actions performed.





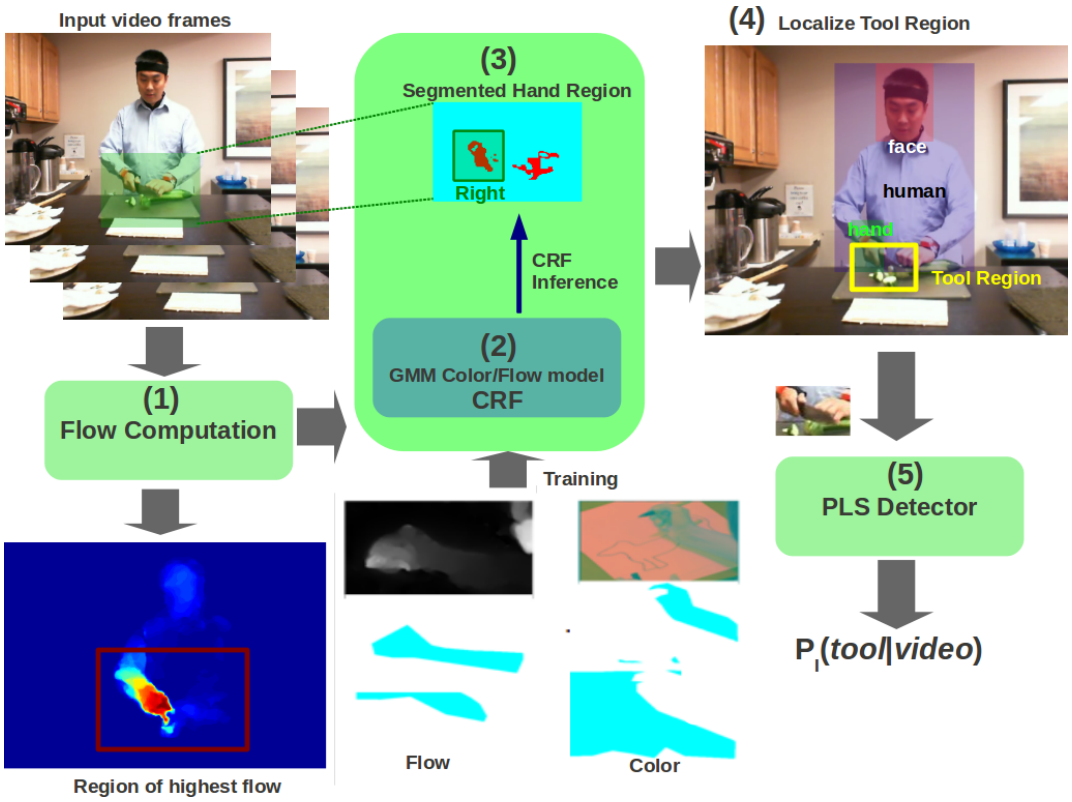
Corpus-Guided Action Recognition

- Input: set of unlabeled videos. Goal is to estimate a model that assigns clusters of videos to the correct actions.
- VE: detects tools and action features (noisy).
- LE: provides a *language model* for computing likelihoods of tool-action co-occurrence.
- Strategy: EM to update the action assignment model at each iteration.



(a) Training the language model from a large text corpus. (b) Detected tools are queried into the language model. (c) Language model returns prediction of action. (d) Action features are compared and beliefs updated.

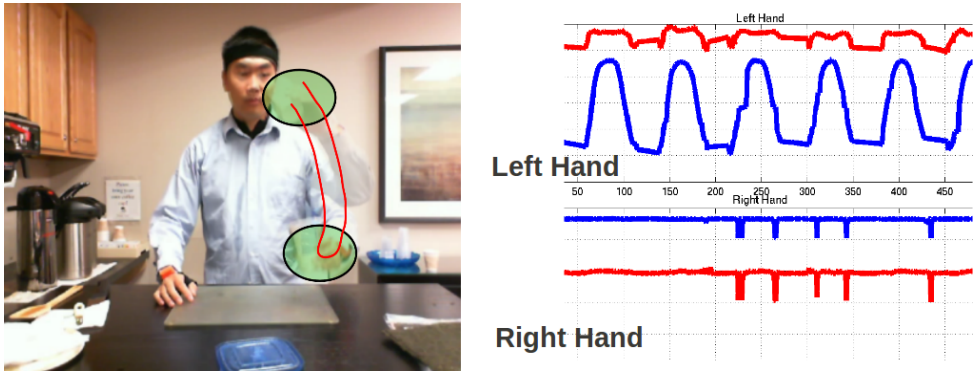
Active Tool Detection



Overview of the tool detection strategy: (1) Optical flow is first computed from the input video frames. (2) We train a CRF segmentation model based on optical flow + skin color. (3) Guided by the flow computations, we segment out hand-like regions (and removed faces if necessary) to obtain the hand regions that are moving (the active hand that is holding the tool). (4) The active hand region is where the tool is localized. Using the PLS detector (5), we compute a detection score for the presence of a tool.

Extracting Action Features

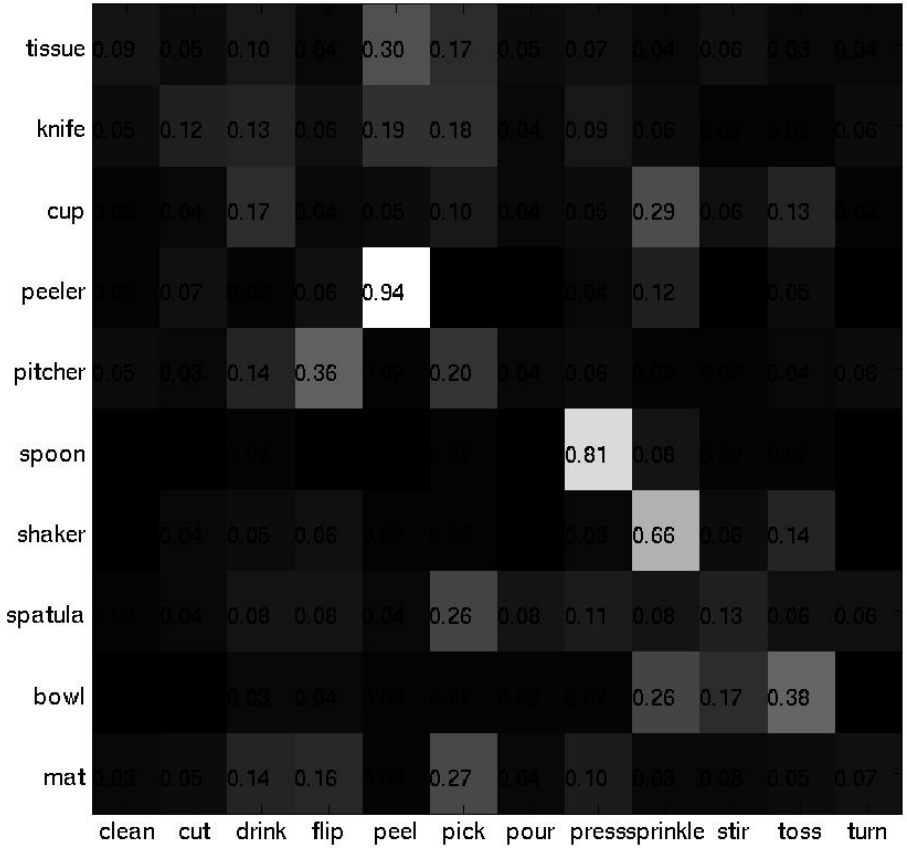
- Detected hands are tracked using a Kalman Filter across the entire video sequence.
- Creates a 20-dim feature vector consisting of Fourier coefficients and normalized averaged velocities.



Detected hand trajectories. x and y coordinates are denoted as red and blue curves respectively.

Action-Tool Language Model

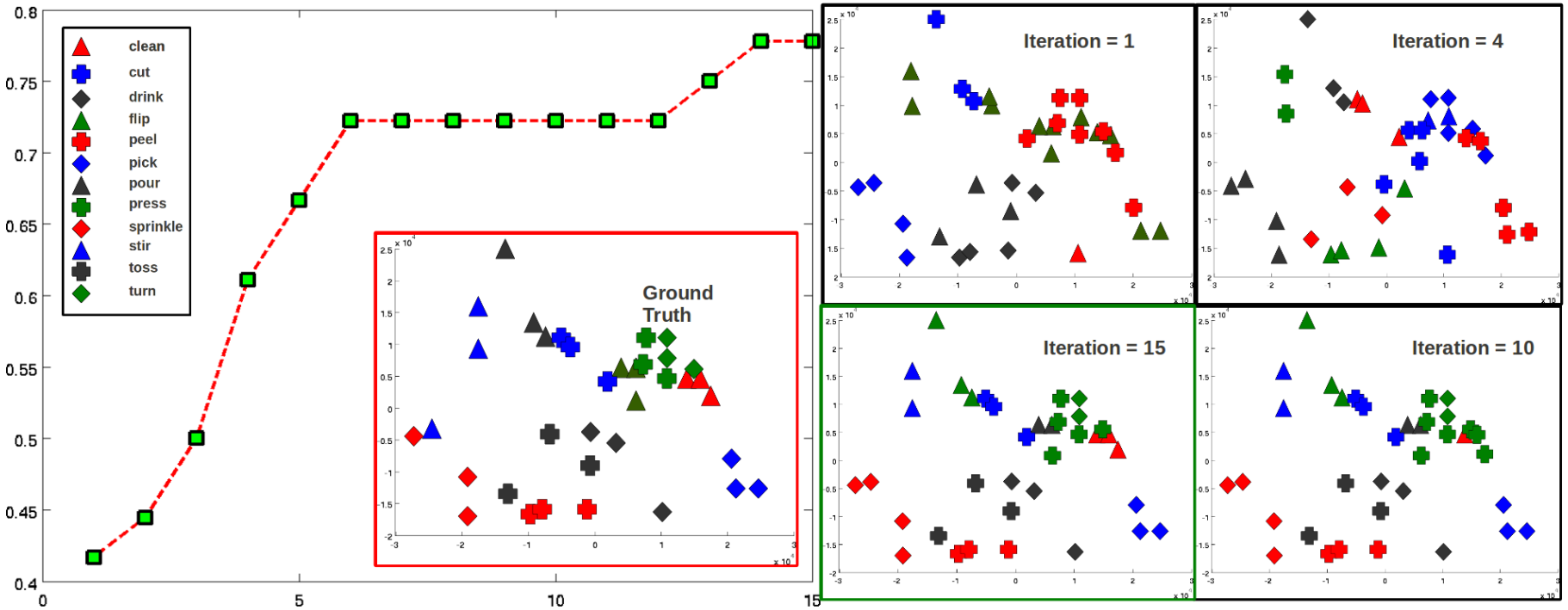
- Computes the conditional probability that an action occured *given* that a tool has been detected: $\mathcal{P}_L(v_j|n_i) = \frac{\#(v_j,n_i)}{\#(n_i)}$.
- Uses the Gigaword Corpus containing 10 years worth of NYT newswire data.



Gigaword co-occurrence matrix of tools and predicted actions.

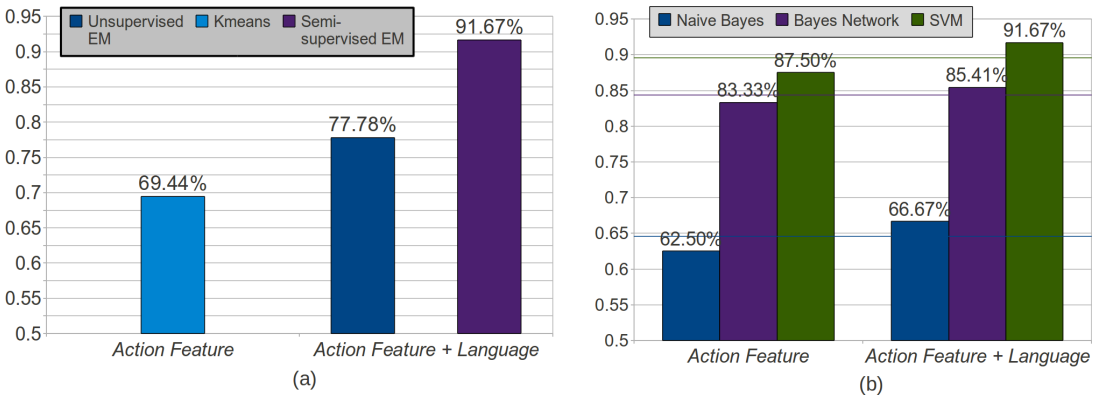
Guiding Action Recognition Via EM

- E step: compute the expectation of the assignment variable for action j, tool i, video d: $\mathcal{W}_{ijd} \propto \mathcal{P}_I(i)\mathcal{P}_L(j|i)Pen(d|j)$
- M step: update the model parameter $\hat{\mathcal{C}}$ via: $\arg\max_{\mathcal{C}} \left(-\sum_{i,j,d} \mathcal{W}_{ijd}||F_d - \mathcal{C}_j||^2\right)$



Experiments and Results

- Evaluated over a dataset of 4 humans making sushi – UMD-Sushi Making Dataset[a].
- 2 experiments: (1) Unsupervised Clustering (K-Means vs EM) and (2) Supervised Classification.

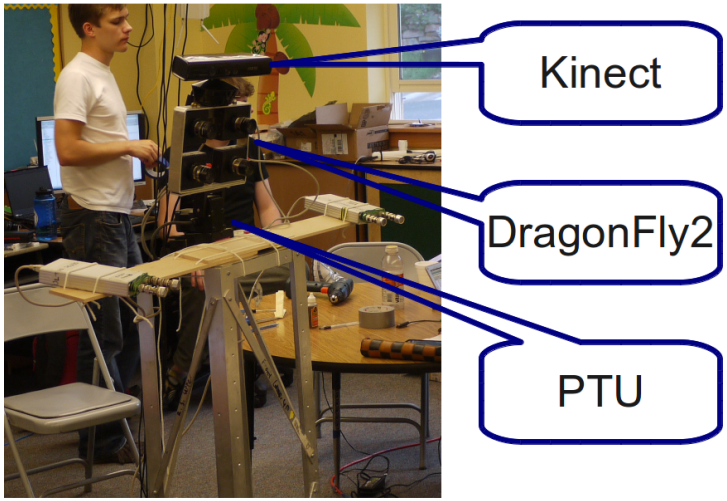


(a) Unsupervised recognition accuracy: no language (K-Means) versus language (EM). (b) Classification accuracy: no language versus language. All reported results have variances within $\pm 0.5\%$.

[a] http://www.umiacs.umd.edu/research/POETICON/umd_sushi/

Future Work

- Integrating the framework into a robotic agent – completes the Cognitive Dialog.
- Better language modelling – using NER or shallow parsing methods to obtain better $\mathcal{P}_L$ estimates.
- Using Kinect as input to reduce viewpoint dependency.
- Using *attributes* for more robust tool and action detection.



Conclusions

- When used properly, language can be exploited to solve important problems in vision.
- In this talk, used a Cognitive Dialog framework for action recognition.
- Promising results point to feasibility of using Language in even more Computer Vision problems.

Thank you

- Contact information: cteo@cs.umd.edu
- Questions?