Integrating Language into Computer Vision

Ching L. Teo

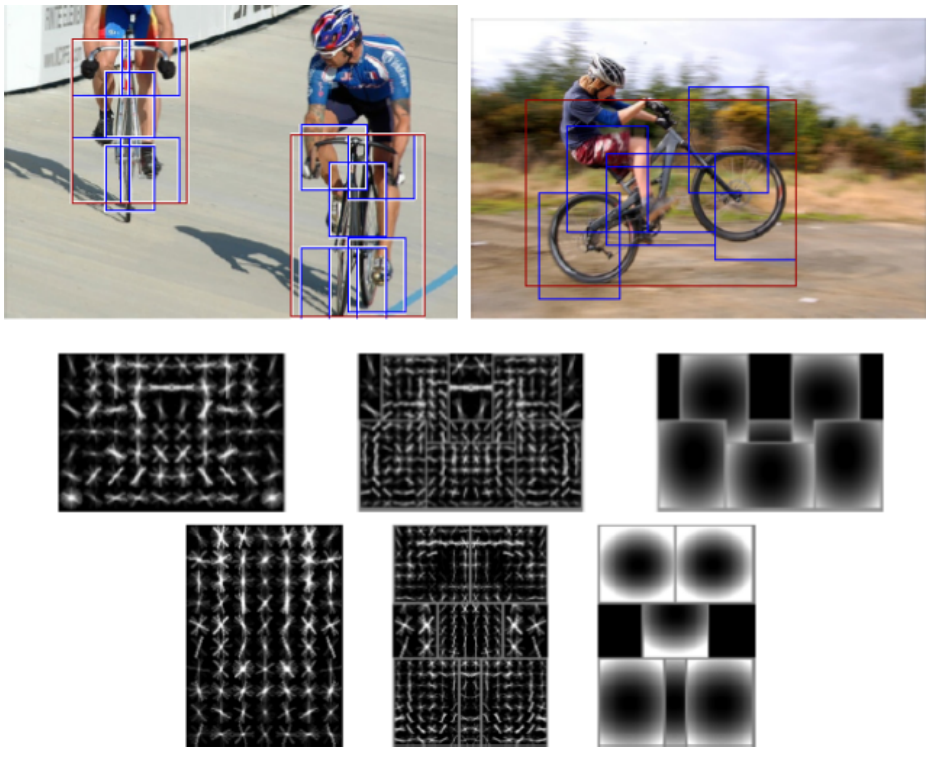
University of Maryland Institute for Advanced Computer Studies

09 Jan 2012

Outline

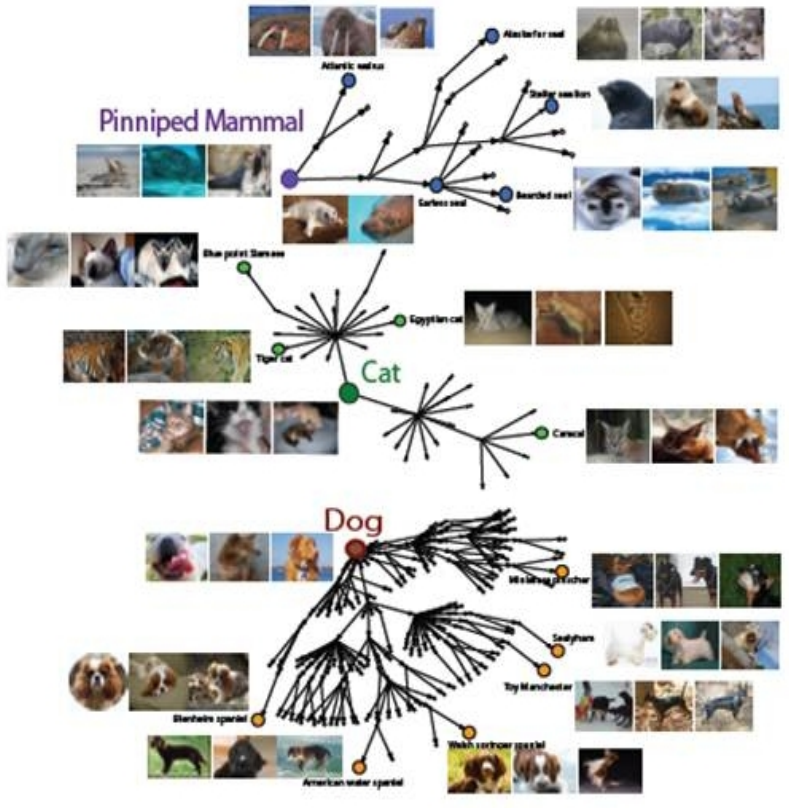
Motivation

- 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?

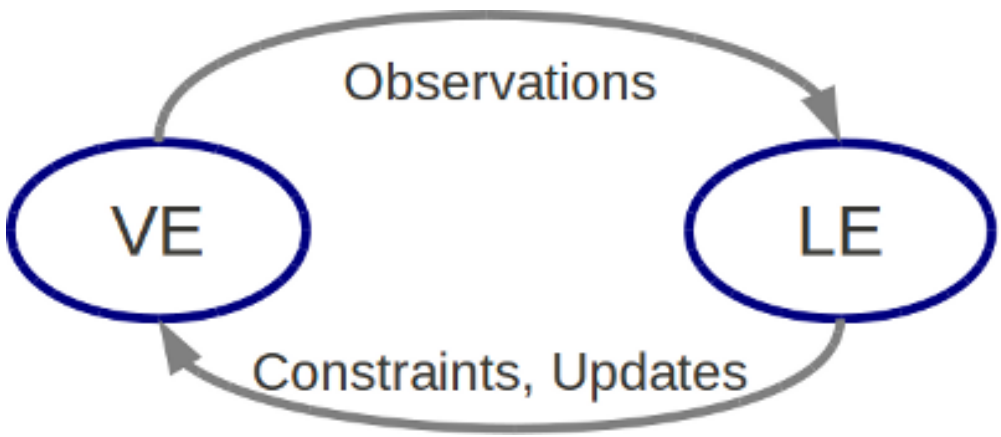
Language provides:

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, plausiblity 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.



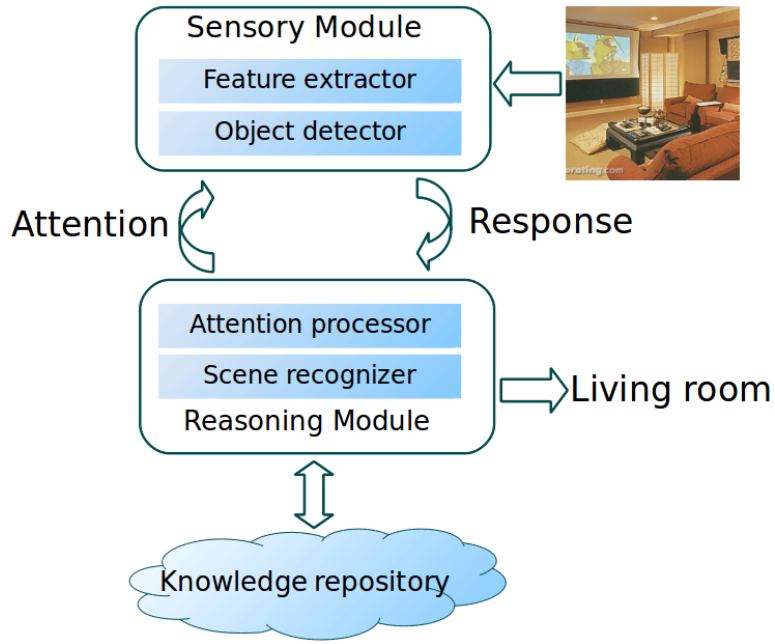
Implementing the Dialog

Three different works explored so far:

1. Active scene recognition.
2. Action-Tool Recognition.
3. Generating descriptions of images/videos.

Active Scene Recognition

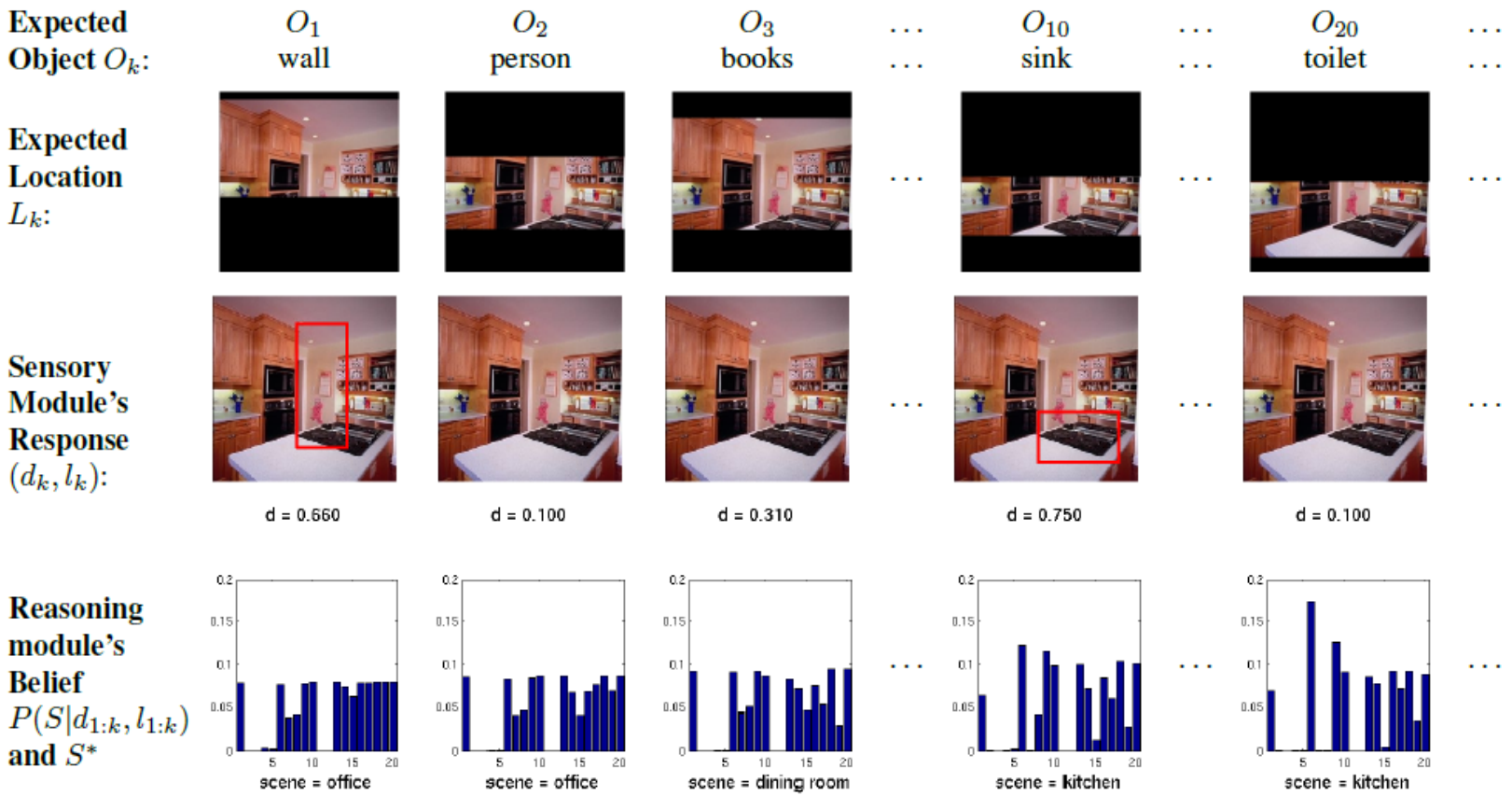
- Goal: Scene recognition guided by high-level knowledge.
- VE: Object Detection + Recognition.
- LE: Guides VE to select the type and position of the object detector to use in the *next* iteration.
- Determines the most likely scene category based on history of detection scores and object's location.



X. Yu et al, *Active Scene Recognition with Vision and Language*, ICCV 2011

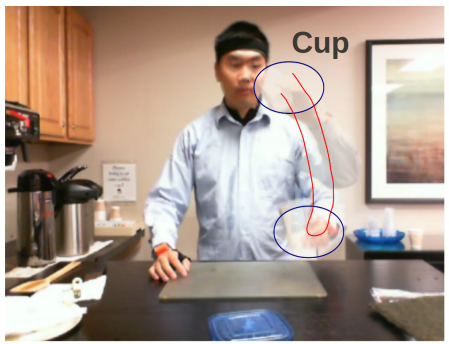
Information Gain Criterion

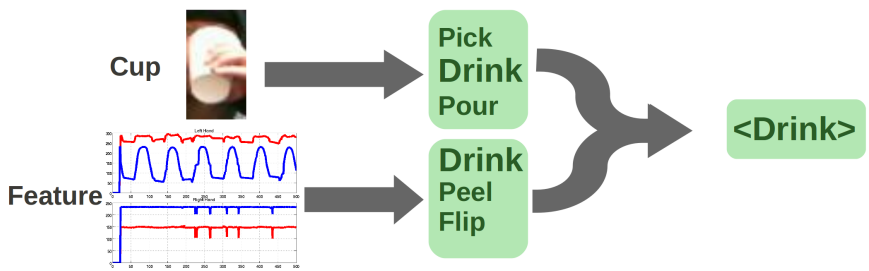
- Selection of the next object category to detect is based on computing the expected information gain across all objects and locations:
  $\{O_k^*, L_k^*\} = \arg\max_{O_k \in N_{k-1}, L_k \in \mathcal{L}_k} \mathbf{I}(S; d_k, l_k | d_{k-1}, l_{k-1})$



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.

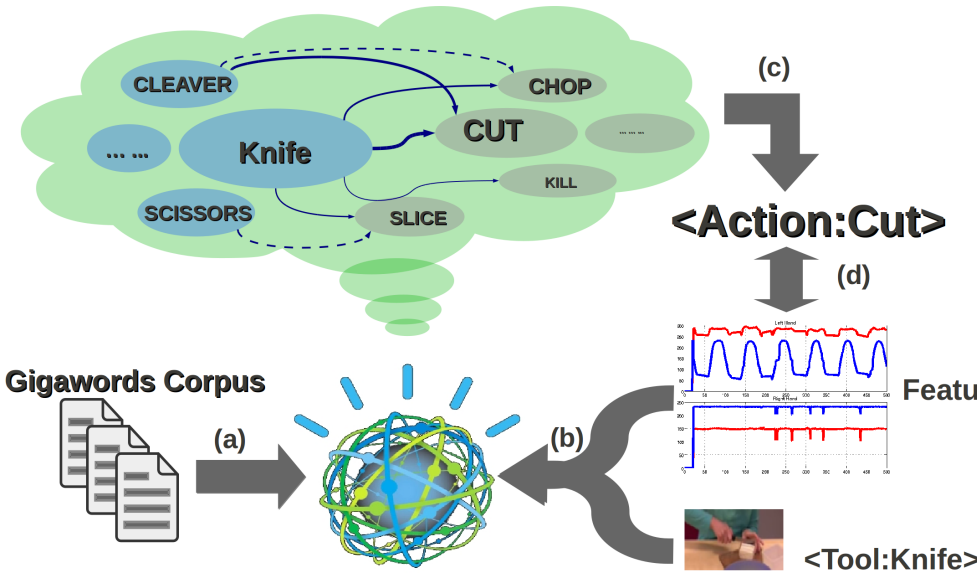




C. L. Teo et al, *The Watson That Sees: Language-Guided Action Recognition for Robots*, ICRA 2012

Corpus-Guided Action Recognition

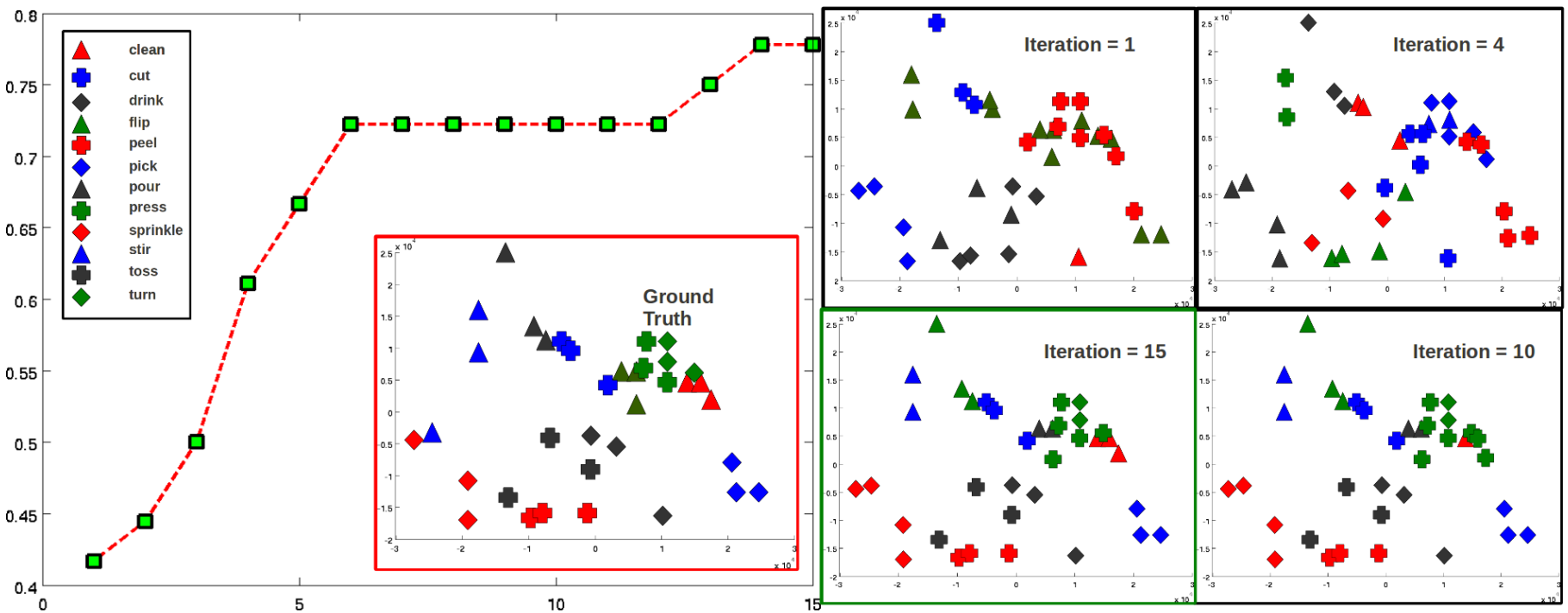
- Input: set of unlabeled videos. Goal is to estimate a model that assigns clusters of videos to the correct actions.
- VE: detects tools and action features (noisy).
- LE: provides a *language model* for computing likelihoods of tool-action co-occurance.
- Strategy: EM to update the action assignment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

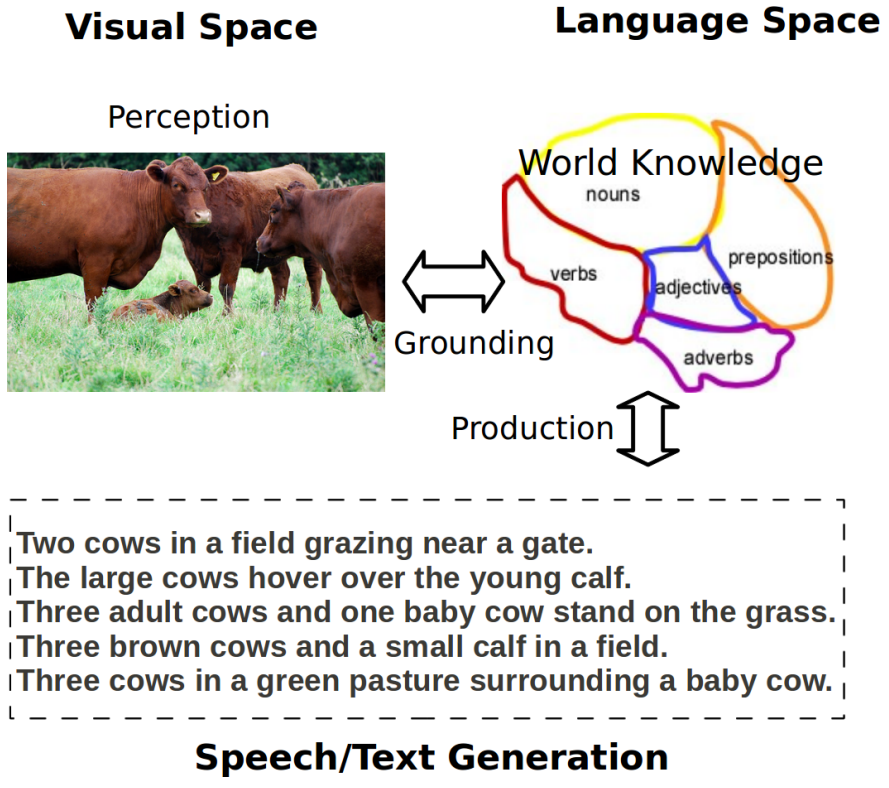
EM steps

- 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: $\arg\max_{\mathcal{C}} \left(-\sum_{i,j,d} \mathcal{W}_{ijd} ||F_d - \mathcal{C}_j||^2 \right)$



Generating Descriptions of Images

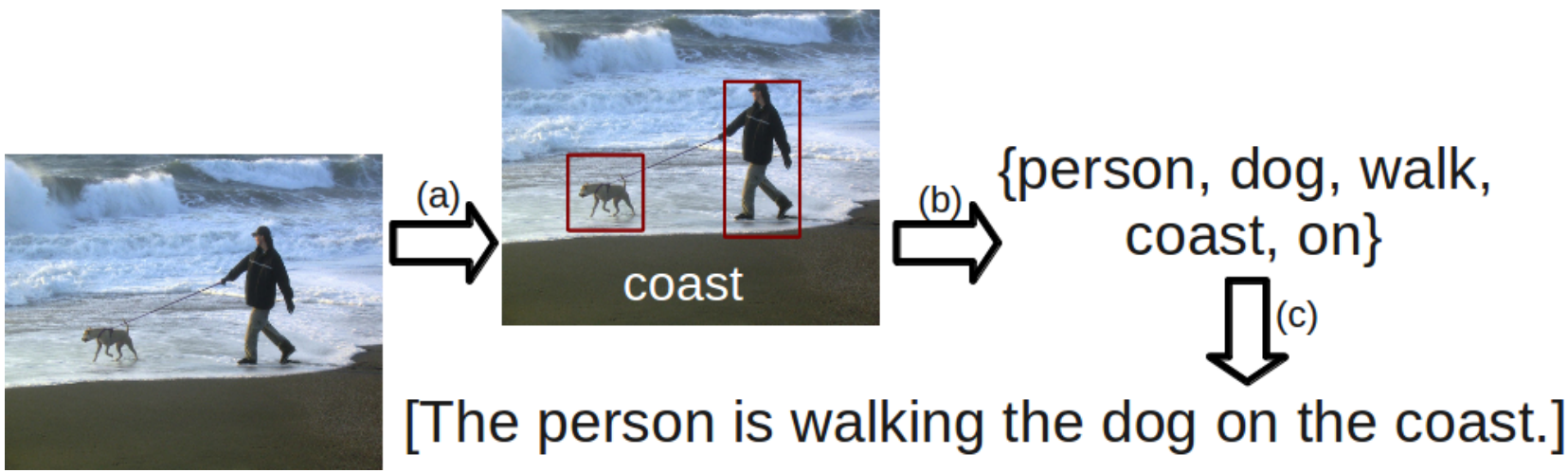
- Goal: Automatic generation of a sentence that *describes* an image.
- Three processes: visual perception, grounding via language and sentence production.
- Key hypothesis: natural images describe common everyday scenarios which are captured in language.
- Assumption: images are described by $\mathcal{T} = n, v, s, p$ (nouns, verb, scene, preposition).



Y. Yang et al, *Corpus-Guided Sentence Generation of Natural Images*, EMNLP 2011

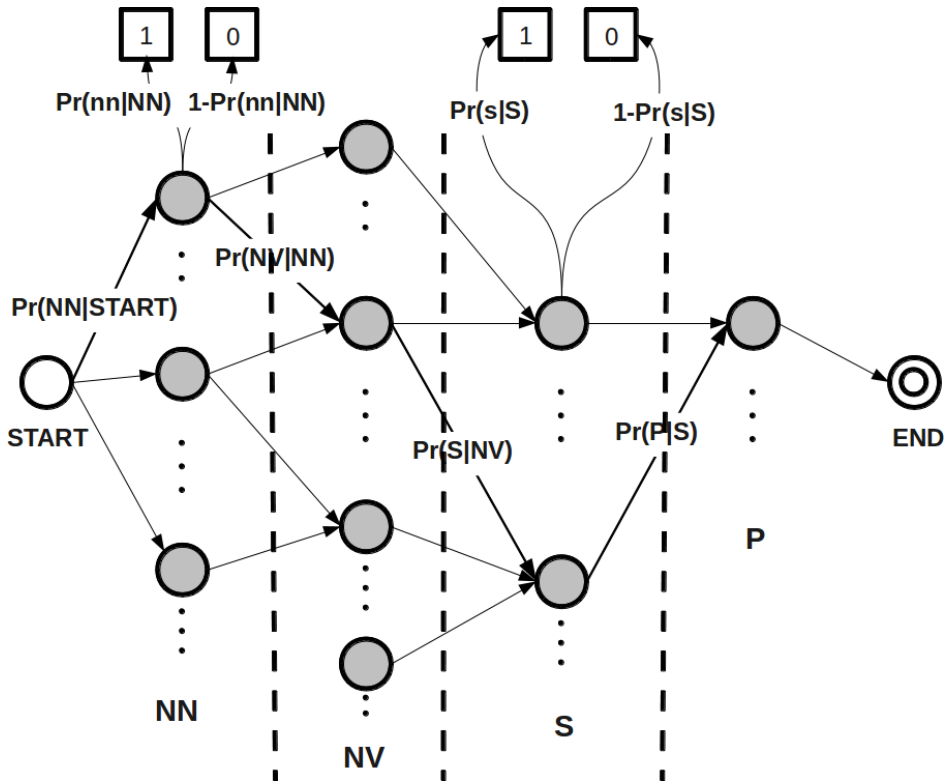
Approach

- Input: static image of a natural scene.
- VE: object and scene detection.
- LE: predicts the verb that co-occurs with detected objects and scene to determine $\mathcal{T}^*$.
- A plausible sentence is then generated from $\mathcal{T}^*$.



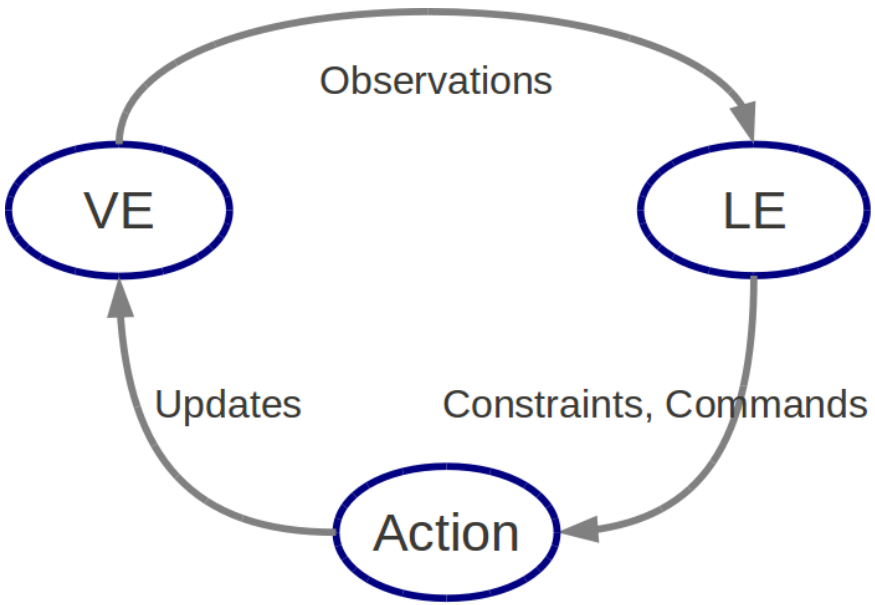
Predicting $\mathcal{T}^*$

- A dynamic programming approach using HMM.
- Observations: Detections from VE, $P_r(n|I)$ and $P_r(s|I)$ (nouns and scenes).
- Transition probabilities: statistics from large text corpus (NYT), $P_r(v|n_1, n_2), P_r(s|n, v), P_r(p|s)$.
- Optimize: $\mathcal{T}^* = \arg\max_{n,v,s,p} P_r(\mathcal{T}|n, v, s, p)$



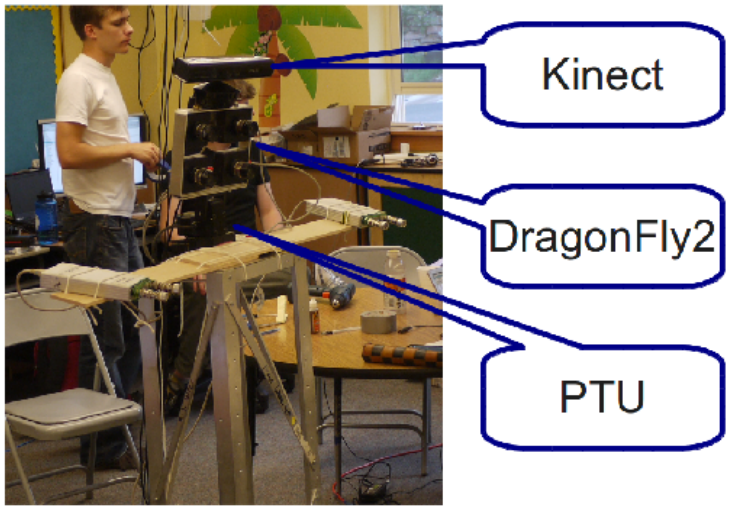
Completing the Cognitive Dialog

- A missing component of the dialog is *action*, implemented on a mobile robot.
- The action component acts upon results from the LE, which in turn affects the VE.
- Applications: navigation (where to go next), attention (where to look next), scene understanding.



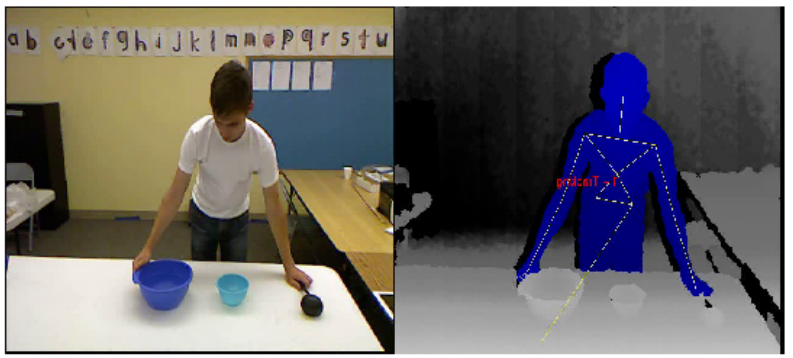
Robotic platform

- Telluride Workshop 2011: integrated components onto a mobile robot.
- Task: recognize kitchen activites performed by a human actor.



VE modules

- Object/tools segmentation.
- Skeletal tracking of human actor, extract trajectories.
- Extracting basic 3D shape description of objects:
  a) elongated
  b) has_handle
  c) container.
- Output: a triplet of {tool, object, action}.

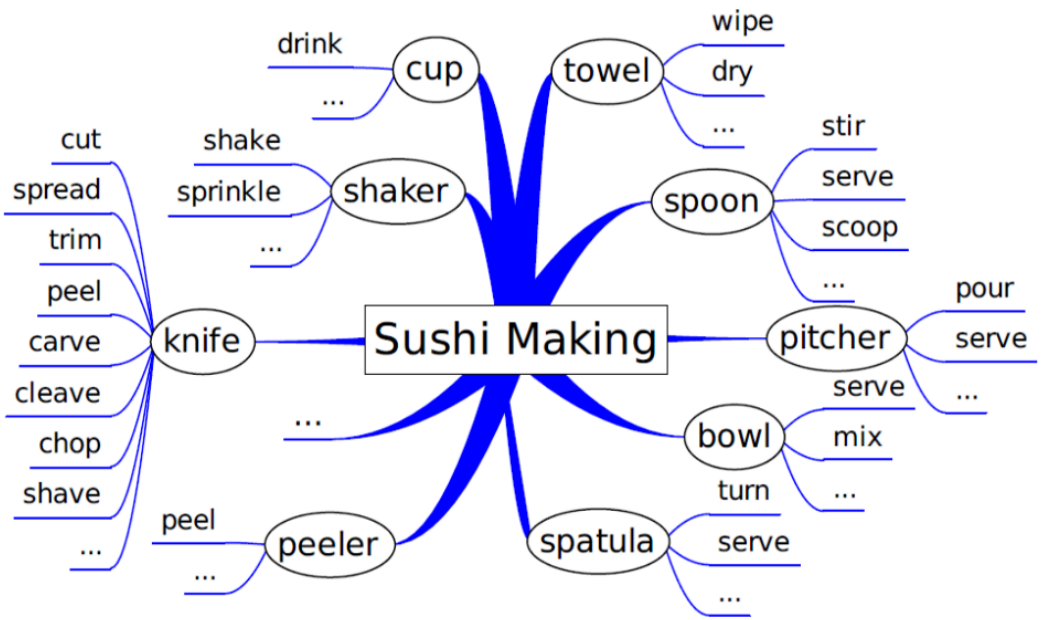


(a) (b) (c)

LE modules

- Uses a language reasoner to determine if the triplet from VE makes sense:
- E.g. {knife, tomato, slice} → ACCEPT
- E.g. {bowl, mug, mash} → ALTERNATIVE: {bowl, masher, mug}

Language reasoner is based on a respository of semantic concepts built from a variety of resources: WordNet, FrameNet, etc...



Conclusions

- When used properly, language can be exploited to solve important problems in vision.
- For this talk, used a Cognitive Dialog framework:
  1) scene recognition,
  2) action recognition,
  3) image description.
- Full potential of this integration will be realized on an active agent.

Acknowledgements

- Advisors: Prof. Yiannis Aloimonos, Dr. Cornelia Fermüller, A/Prof. Hal Daumé III.
- Collaborators: Xiaodong Yu, Yezhou Yang.
- Contact information: cteo@cs.umd.edu