Synergistic Methods for using Language in Robotics

Ching L. Teo
University of Maryland
Dept of Computer Science
College Park, Maryland 20742
+01 3014051762
cteo@cs.umd.edu

Yezhou Yang
University of Maryland
Dept of Computer Science
College Park, Maryland 20742
+01 3014051762
yzyang@cs.umd.edu

Cornelia Fermüller
University of Maryland
Institute for Advanced
Computer Studies
College Park, Maryland 20742
+01 3014051743
fer@umiacs.umd.edu

Yiannis Aloimonos
University of Maryland
Dept of Computer Science
College Park, Maryland 20742
+01 3014051768
yiannis@cs.umd.edu

ABSTRACT

This paper presents an overview of our work on integrating language with vision to endow robots with the ability of complex scene understanding. We propose and motivate the Vision-Action-Language loop as a form of cognitive dialogue that enables us to integrate current tools in linguistics, vision and AI. We present several experimental results of preliminary implementation and discuss future research directions that we view as crucial for developing the cognitive robots of the future.

Categories and Subject Descriptors

I.2.9 [**Artificial Intelligence**]: Robotics; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*perceptual reasoning*

General Terms

Theory, Algorithms

Keywords

Cognitive Robotics, Computer Vision, Computational Linguistics

1. INTRODUCTION

A *cognitive robot* is a robot capable of simulating cognitive processes that mimic human intelligent behavior requiring capabilities such as visual perception, sensorimotor activation and high-level reasoning. In this paper, we argue that *Language* is an important, and till now an overlooked component that is crucial for developing cognitive robots. As we will show in sec. 3, language, when processed appropriately, can be leveraged to bridge the so-called *semantic* gap between low-level sensory signals (visual, auditory, haptics etc.) and high-level concepts (words, ideas etc.). In this work, we focus on visual signals, and show how we can use the Vision-Action-Language loop depicted by Fig. 1 as a form of cognitive dialogue to facilitate several important vision tasks: 1) Object recognition, 2) Action recognition and 3) Scene description. We first motivate why language is useful for cognitive robots followed by an overview of the cognitive dialogue framework.

PerMIS'12 , March 20-22, 2012, College Park, MD, USA.

1.1 Why Language for Robotics?

Let us examine in some more detail what is really going on when a human (a cognitive system with vision and language) is interpreting a visual scene. When we fixate at an object and recognize it, then this means an immediate entry to the linguistic system. Indeed, if we recognize a "street", the word street lights up in the linguistic system, with a number of consequences. The word "street" has many "friends". These are other words that tend to co-occur with "street", such as "human", "car", "house", etc. Modern computational linguistics has created, using a large corpus, resources where this information can be obtained, e.g. probability distributions for the co-occurrence of any two words, lists of the friends of any word, and so on. Thus, recognizing a noun in the scene creates expectations for the existence of other words in the scene that vision can check for. In this case, **language acts as a contextual system** hat aids perception. There is however much more than this. Let's say you are in a kitchen. Because you have prior knowledge about kitchens, their structure and the actions taking place in them and a large part of this knowledge is expressed in language, we can utilize this information during visual inspection. A knife in the kitchen will most probably be used for "cutting" a food item, so the vision can look for it. In this case, **language acts as a high level prior knowledge system** that aids perception. There is still more. Let's say you observe someone pick up an object, put it in the trunk of a car, then get into the car and drive away. Given this, you know that the object is gone, it is inside that car. In this case, **language acts as part of a reasoning process.**

When we visually inspect a scene, it appears that our linguistic system is working in the background together with visual perception to achieve meaning and understanding. This is an aspect of perception that has not been studied systematically. There has been a lot of work on what could be called "parallel vision", i.e. given an image or an image sequence, how do we find edges, contours, motions and other features, how do we segment the scene and group the features into objects, etc. On the other hand, "sequential vi-

sion" has not received as much attention. As you interpret a visual scene, you fixate at some location and you recognize nouns, verbs, adjectives, adverbs and prepositions. Because the linguistic system is highly structured, these recognitions produce a large number of inferences about what could be happening in the scene. This leads you to fixate at a new location, and the same process is repeated. In this case, **language acts as part of an attention mechanism.**

Thus, language is beneficial not so much for communication, but for facilitating the shaping of different cognitive spaces. Finally, it should be clear that instead of language one could use a formal system with properties like the ones of language. The symbols of the system would be labels of the different concepts that the system possesses and they would have to obey a number of constraints. Language gives us this for free. In the next section, we describe how the Vision-Action-Language loop integrates language to realize some of the uses that was described here.

1.2 The Vision-Action-Language Loop

The Vision-Action-Language loop is depicted in Fig. 1.

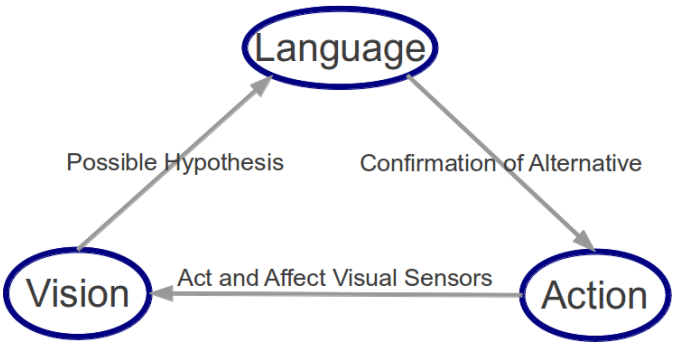

Figure 1: The Vision-Action-Language Loop.

Each of the three nodes can be seen as a distinct process (an *executive*) in the Robot's operating system . The *Visual* executive takes care of low-level visual processing associated with the task at hand: e.g, segmenting an object, or extracting certain visual features. The output of the visual executive are a set of possible hypothesis on the task which is then passed on to the *Language* executive. The Language executive will then act as a reasoner, using high-level knowledge embedded in language to decide which, if any, of the hypothesis makes sense; and provide reasonable alternatives. The output of the Language executive is therefore a set of potentially modified hypothesis which can be acted upon by the *Action* executive. Based on the set of modified hypothesis, the Action executive will then decide the most appropriate next course of action that will affect the visual sensor: e.g. to move to a new location or to change the sensors' pan-tilt-zoom (PTZ) unit. This Vision-Action-Language loop continues until the Action executive decides that a certain end goal or objective had been realized which is then relayed to the rest of the robot's operating system. We call it a cognitive "dialogue" as the three executives are constantly working in a synergistic manner to update each others prior beliefs, so as to achieve a shared goal or objective together.

2. RELATED WORKS

The use of language in robotics has been pursued recently in the fields of Computer Vision, AI and Robotics. We highlight a few prominent related studies in these areas.

In the field of computer vision, the classical view of Marr and others [17] considered language to be part of high-level vision, diametrically opposed to the low-level visual processes that processes the signals directly. As a result, language was only used "at the end" of the visual processing pipeline. With advances on textual processing and detection, several works recently focused on using sources of data readily available "in the wild" to analyze static images. The seminal work of [4] showed how nouns can provide constraints that improve image segmentation. [9] (and references herein) added prepositions to enforce spatial constraints in recognizing objects from segmented images. [1] processed news captions to discover names associated with faces in images, and [11] extended this work to associate poses detected from images with the verbs in the captions. Some studies also considered dynamic scenes. [2] studied the aligning of screen plays and videos, [15] learned and recognized simple human movement actions in movies, and [10] studied how to automatically label videos using a compositional model based on AND-OR-graphs that was trained on the highly structured domain of baseball videos The work of [5] attempts to "generate" sentences by first learning from a set of human annotated examples, and producing the *same* sentence if both images and sentence share common properties in terms of their triplets: (Nouns-Verbs-Scenes). No attempt was made to generate *novel* sentences from images beyond what has been annotated by humans.

In AI, the use of language had been largely confined to classical problems in computational linguistics: 1) speech recognition 2) language modeling (e.g. machine translation) and 3) text generation. In speech recognition, current approaches include automatic speech recognition and understanding, both need language information as prior knowledge. For language modeling, the work of IBM models uses large parallel text corpus to build HMM style language models [12], and then apply it into several applications, such as machine translation. In terms of text generation, classic approaches [25] are based on three steps: selection, planning and realization. A common challenge in generation problems is the question of: what is the input? Recently, approaches for generation have focused on formal specification inputs, such as the output of theorem provers [20] or databases [6]. Most of the effort in those approaches has focused on selection and realization.

State of the art robotics uses language as a communication system; conversational robots of the new millennium have more or less sophisticated mechanisms to map words to related sensorimotor experiences so that they engage into more natural human robot interaction (e.g. [18], cf. also [22] for an extensive review). Language has been used to trigger action-sensory state associations ([26]) or predesigned control programs ([16]); mappings from natural language to symbolic logic or temporal logic and then to basic control primitives of the robot ([13]) have also been developed for controlling robots with high level task descriptions. The system of [19] describes model that enables the agent to ground evidences from multiple modalities: language, vision, etc. However, none of these approaches takes advantage of language as a contextual system and as part of a reasoning system. With the exception of a few notable approaches on understanding of gestures by robot platforms (cf. for example [14]) or using visual scenes to prime speech understanding ([23]) there has not been much work on scene interpretation by robotic agents. There are many reasons for this, but basically computer vision solutions developed in the image/video databases arena that use language as a contextual system do not transfer to robots.

3. INTEGRATING LANGUAGE

In this section we present preliminary implementations of the Vision-Action-Language cognitive dialogue on three tasks: 1) Object recognition, 2) Action recognition and 3) Scene description.

For each task, we highlight how each implementation is related to the cognitive dialogue and summarize the results from experiments performed on a robot that is endowed with the presented algorithms.

3.1 Attributes-Based Object Recognition

The key goal of any object recognition task is to provide distinct labels to objects within the image. For language to be integrated into this task, we propose to use *attributes* that link visually extracted information to textual descriptions that humans would use to describe these objects. An attribute can be defined as a property that is *innate* to the object, and as a result is *invariant* under most circumstances. In addition, the use of attributes has strong links to human perception [3]. Such properties makes attribute detection an important capability for cognitive robots. Our approach first segments the image into foreground regions and background, and then computes on the foreground object attribute properties. In this study we focused on shape properties. Since our application was the description of kitchen tools (3.4 we have identified the following five computable attributes:

Is elongated: An ellipse was fitted to the mask provided by segmentation, and the ratio of major to minor axis was used to set a threshold (Fig. 2a).

Is round: If the ratio of major to minor axis is about the same, the object was considered round.

Has a handle: If the error from fitting two separate ellipses was lower than from fitting a single ellipse, the object was considered having a handle (Fig. 2b).

Is a container: Depth discontinuities were found in the depth mask. If the object could be segmented into parts, with one part of a mostly concave depth map and the other part of a mostly convex depth map, the object was considered a container (Fig. 2c).

Has a flat part: If an object was classified as consisting of two parts by the 2D shape attribute method, a plane was fitted to the larger part.

(a) (b) (c)

Figure 2: Examples of shape attributes (a) *Elongated* (b) *Has handle* (c) *Is a container*.

The Language Executive is a simple language model that uses the attributes extracted by the Visual executive to perform a classification of the object's identity as shown in Fig. 3 using a decision tree based classifier.

3.2 Action Recognition

For this task, we are interested in recognizing actions associated with certain hand-tools. The basic intuition is to exploit the close semantic relationship between actions and tools in a large text corpus to improve the recognition of actions and tools in the visual space. The basic framework is summarized in Fig. 4

The Visual Executive extracts visual features related to the action (trajectories of hand) and tools. It then performs a classification of these features to produce initial hypothesis of their labels, which is expected to be noisy. The Language Executive first creates a language model that gives the conditional probability of how likely an action has occurred given the tool. This was done by mining a large text corpus [8] for correlated tools and actions. We then combined the probabilities to determine the final labels of tool and associated action. This step can be repeated in a few iterations, where at each iteration, we retain only the top N hypothesis of actions and tools until we do not see any significant updates or only a single pair of tool and action exists.

3.3 Scene Description

The goal of this task is to produce a textual description of an image or video sequence based on a triplet $\mathcal{T}$ of objects, actions and environments (locations) that co-occur in the scene. The full details of the implementation are described in [27], and we link it to the Vision-Action-Language loop described here. The key component of the approach in [27] is the dynamic programming optimization of an HMM that integrates language and visual input as shown in Fig. 5.

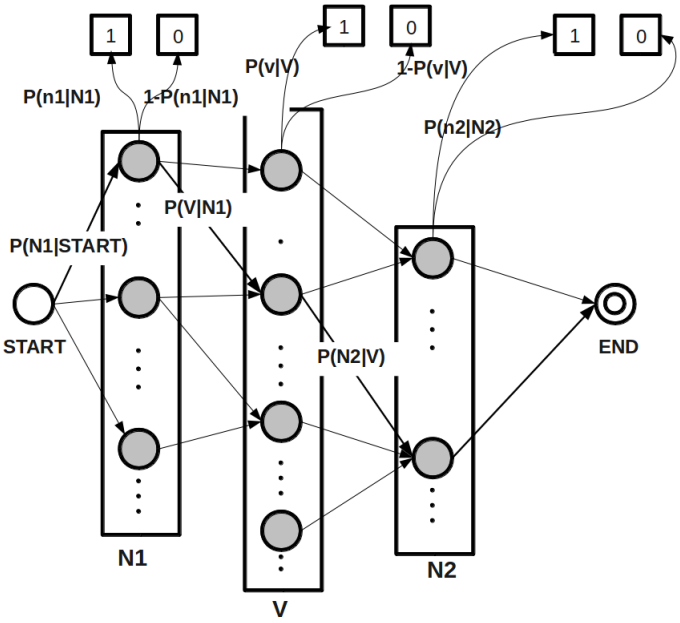

Figure 5: The HMM used to predict the optimal triplet $\mathcal{T}$: N_1, N_2 corresponds to objects and tools, and V corresponds to verbs (actions). The relevant transition and emission probabilities are also shown. See text for more details.

The key idea to this approach is to model the detection scores from visual object and scene detections as *emissions* (observations) in the HMM. This is the Visual Executive in the framework. The transition probabilities, learned from the same large text corpus [8], describe how the different components of $\mathcal{T}$ relate to each other. This forms the Language Executive. Optimizing over the HMM essentially finds the most likely $\mathcal{T}$ that supports both visual observations and linguistic correctness, which simulates the cognitive dialogue between the processes. A template based method of generating sentences is then used to generate a descriptive sentence from $\mathcal{T}$.

3.4 The Telluride Experiments

The algorithms described in the preceding sections are implemented on a mobile robot whose goal is to observe a human perform certain actions with kitchen tools and to ultimately generate a

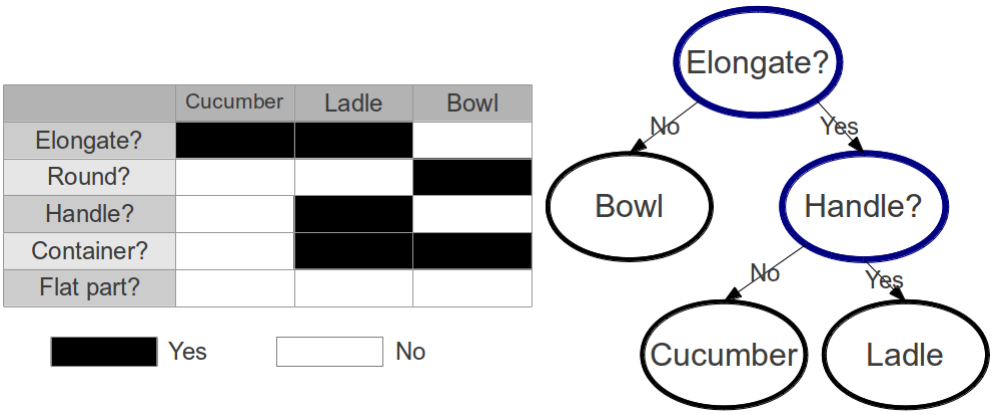

Figure 3: Example of using attributes for object recognition.

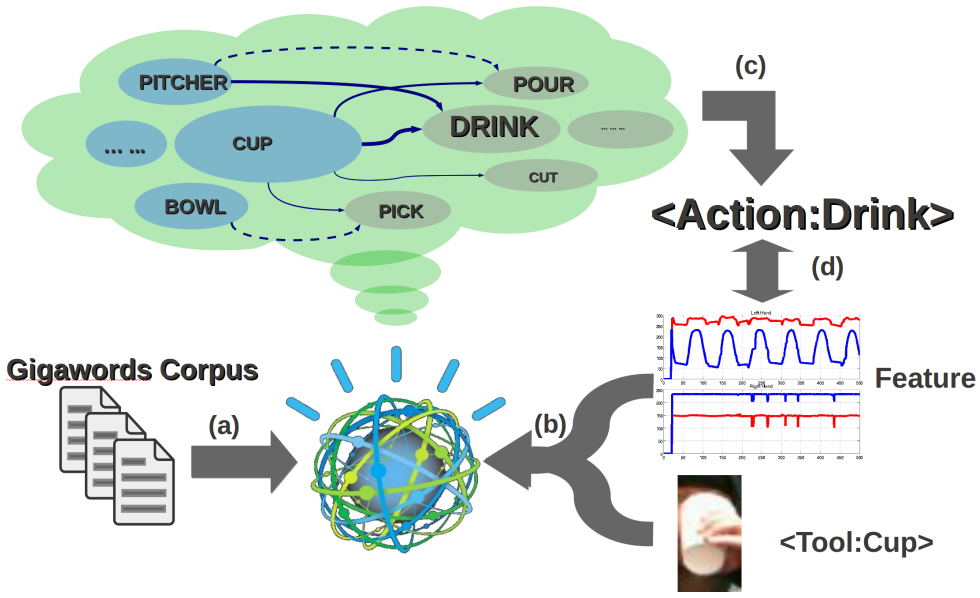

Figure 4: Key components of the approach.: (a) Training the language model from a large text corpus. (b) Detected tools are queried into the language model. (c) Language model returns prediction of action. (d) Action features are compared and beliefs updated.

sentence that describes the actions. All the experiments were conducted during the 2011 Telluride Neuromorphic Workshop[1], and we first describe the experimental setup and procedure and report accuracy results.

3.4.1 Experimental Setup

The robot (Fig. 6), is looking at the table where humans perform tasks using a number of tools and objects. The session begins with a number of objects, $o \in O$ and tools $t \in T$ on the table which the robot observes. Then a person approaches and begins an action $a \in A$, out of set of $|A|$ actions.

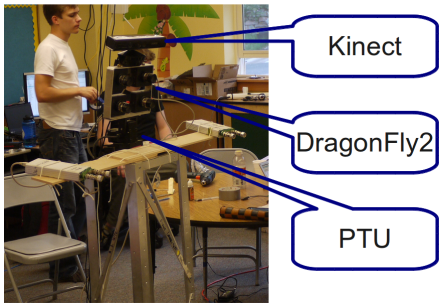

Figure 6: The Telluride Robot used in the experiments with its sensory hardware.

The robot first extracts visual features of objects and tools from the table and attempts to label them using attributes as described in sec. 3.1. This yields a set of scores over all objects o and tools t. When the action starts, it tracks the hand and elbow locations of the human (using the on-board kinect sensor) to extract action features (velocity and Fourier coefficients). Together with the labels of the tools, we use the approach described in sec. 3.2 to compute a detection score for each action a. With these initial detection scores, we use the algorithm described in sec. 3.3 to generate the final triplet $\mathcal{T}$ of object, tools and actions in order to generate a reasonable sentence that describes the scene. The overview of the processing pipeline is shown in Fig. 7.

The experimental test dataset consists of 9 actions: A={slice, mash, peel, chop, pour, stir, toss, sprinkle, pour} performed by 2 different human actors using 9 common tools: T={knife, masher, peeler, pitcher, ladle, fork/spoon, shaker, mug, bowl} and 7 other objects: O={bowl, mug, tomato, cucumber, coffee, soup, salt}. In total, there are 18 video clips, each with 9 actions performed by the 2 actors.

3.4.2 Results

The output of the initial visual processes is the triplet of $\mathcal{T} = \{a, o, t\}$ of action, objects, tools associated with the video observed. The initial output triplet $\mathcal{T}_1, \mathcal{T}_2$ (one for each actor) is then passed on to scene description algorithm (sec. 3.3) which then modifies the triplet if necessary to form the final output triplet $\mathcal{T}_1^*, \mathcal{T}_2^*$. We evaluate the effectiveness of our approach by comparing the overall recognition accuracy Acc, computed as the weighted average from the recognition of the three components in $\mathcal{T}_{1,2}$ and $\mathcal{T}_{1,2}^*$ with the ground truth. The results over the 18 videos are summarized in Table 1.

These results show that on average, we are able to improve upon the recognition accuracies of objects, tools and actions from pure visual processes with the help of the Language Executive. Mistakes still occur and this is because we have not exploited the "Action" Executive of the cognitive dialogue. We address this issue (along with others) in the next section.

[1]http://ine-web.org/telluride-conference-2011/telluride-2011/index.html

Test Video (Truth)	$\mathcal{T}_1, \mathcal{T}_2(Acc)$	$\mathcal{T}_1^*, \mathcal{T}_2^*(Acc)$
{slice,tomato,knife}	{slice,tomato,knife} {slice,tomato,knife}(1.0)	{slice,tomato,knife} {slice,tomato,knife}(1.0)
{mash,bowl,masher}	{mash,bowl,mug} {sprinkle,bowl,mug}(0.5)	{mash,bowl,masher} {sprinkle,bowl,shaker}(0.67)
{peel,cucumber,peeler}	{toss,cucumber,peeler} {peel,cucumber,peeler}(0.83)	{peel,cucumber,peeler} {peel,cucumber,peeler}(1.0)
{chop,cucumber,knife}	{chop,cucumber,knife} {mash,cucumber,knife}(0.83)	{chop,cucumber,knife} {chop,cucumber,knife}(1.0)
{toss,bowl,fork/spoon}	{toss,bowl,fork} {toss,bowl,spoon}(1.0)	{toss,bowl,fork} {toss,bowl,spoon}(1.0)
{sprinkle,bowl,shaker}	{sprinkle,cucumber,bowl} {sprinkle,bowl,shaker}(0.83)	{sprinkle,bowl,shaker} {sprinkle,bowl,shaker}(1.0)
{stir,bowl,fork/spoon}	{pour,bowl,spoon} {pour,bowl,spoon}(0.67)	{pour,bowl,spoon} {pour,bowl,spoon}(0.67)
{pour,mug,pitcher}	{stir,mug,pitcher} {stir,mug,pitcher}(0.67)	{pour,mug,pitcher} {pour,mug,pitcher}(1.0)
{pour,bowl,ladle}	{pour,bowl,ladle} {pour,bowl,ladle}(1.0)	{pour,bowl,ladle} {pour,bowl,ladle}(1.0)
Overall	0.81	**0.93**

Table 1: Triplet accuracy: Initial predictions and final predictions

4. FUTURE WORK

In this section, we discuss possible future research directions that we believe are important for integrating language into vision and AI for solving problems of scene recognition.

4.1 Adding Action

As we have noted in sec. 3.4.2, the mistakes observed in the Telluride Experiments are due to the fact that the robot is stationary and is passively observing the scene. If the robot becomes an active mobile agent, endowed with an Action Executive, problems that had limited the visual processing performance could be mitigated via several strategies:

- Fixation based tracking: As the scene is dynamically changing, with the human actor moving from one part of the scene to another, tracking where the humans are moving the PTZ unit to focus on them will improve the recognition accuracy of the visual processing by reducing false alarms (limited search space)
- Moving to a new location: Objects and tools that are manipulated will change position throughout the process, and may become occluded from time to time. By moving closer or changing its location, the robot could actively aid the recognition by re-tracking the occluded objects or bringing them closer, aiding visual processing.
- Reacting in a reasonable manner: Adding a robotic arm would allow the robot to directly manipulate objects, which would bring the Vision-Action-Language loop to a deeper level. For example, if an object is determined to be occluded by another in front of it, the language reasoner will hint at the robot to attempt to move the occluder so that recognition can be enhanced, by an action called "move", which is then mapped to the robot's motor system to perform the required action.

4.2 Multi-level recognition of actions

Actions are compositional in nature. Starting from simple actions occurring on a part of the body, we can compose actions from

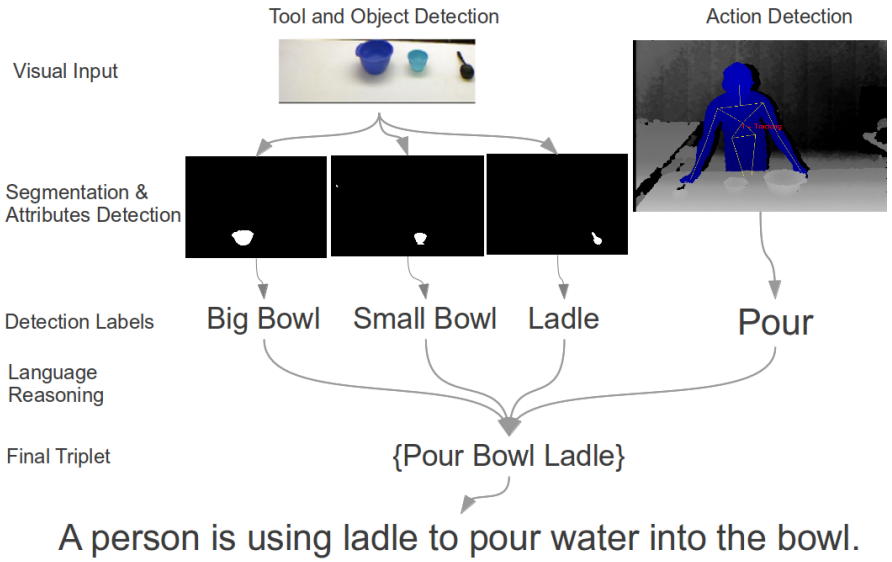

Figure 7: Processing pipeline: from visual features to sentence generation.

several limbs to create more complex actions, and we can further combine a sequence of simple actions with tools together to form an activity. Language can be used to enhance the action recognition at the higher levels and its composition from lower-levels onwards. The key idea is that Language provides a structure that enforces certain constraints on how actions can be composed. For example, focusing on hand-tools alone, there are sets of reasonable actions associated with tools (sec. 3.2). Yet, these actions together are often used to accomplish a global purpose, such as baking a cake. We are in the process of creating several datasets based on cooking recipes so that Language can be used to enforce temporal and logical constraints on how actions can be chained together. The Language executive will work across all levels, from bi-grams of actions to inferring the most likely activity from the sequence of such bi-grams, with a corpus learned from digital cooking recipes.

4.3 Mining from Text and Corpus

We have till now considered the Language Executive to be derived from static sources of corpora. However, for an active agent to be able to accommodate to changes in its surroundings, it is more practical to construct such models "on the fly". Methods such as [7] that perform approximate search through large databases are most promising. In addition, more sophisticated methods that utilize algorithms for relational database mining can be used to extract indirect correlations between objects and their attributes. One interesting way is to exploit relevant questions that humans pose for such objects, and use them to infer possible attributes: e.g. "Is X round? Is Y sharp?". Additionally, one can use various bootstrapping algorithm e.g. [24] using seeds derived from various semantic databases: ImageNet, WordNet etc [21] to extract adjectives where such objects occur.

5. CONCLUSIONS

In this paper, we have argued for the importance of exploiting language in the context of endowing artificial agents with cognitive capabilities. We have demonstrated how the Vision-Action-Language loop can be viewed as a cognitive dialogue between various processes, and we have implemented this dialogue on three tasks, namely object action, and scene recognition. Experiments on our data collected at Telluride confirm that language is a powerful tool which improved object, tool and action recognition. We also discussed potential directions for future work needed to complete the Vision-Action-Language framework in more general settings and for active mobile agents.

6. ACKNOWLEDGMENTS

The support of the European Union under the Cognitive Systems program (project POETICON) and the National Science Foundation under the Cyberphysical Systems Program and the Institute for Neuromorphic Engineering, is gratefully acknowledged. Ching Teo and Yezhou Yang are supported in part by the Qualcomm Innovation Fellowship.

7. REFERENCES

[1] T. L. Berg, A. C. Berg, J. Edwards, and D. A. Forsyth. Who's in the picture? In *NIPS*, 2004.

[2] T. Cour, C. Jordan, E. Miltsakaki, and B. Taskar. Movie/script: Alignment and parsing of video and text transcription. In *ECCV*. 2008.

[3] A. Desolneux, L. Moisan, and J. M. Morel. *From Gestalt Theory to Image Analysis*, volume 34. 2008.

[4] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, editors, *ECCV (4)*, volume 2353 of *Lecture Notes in Computer Science*, pages 97–112. Springer, 2002.

[5] A. Farhadi, S. M. M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. A. Forsyth. Every

picture tells a story: Generating sentences from images. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *ECCV (4)*, volume 6314 of *Lecture Notes in Computer Science*, pages 15–29. Springer, 2010.

[6] D. Golland, P. Liang, and D. Klein. A game-theoretic approach to generating spatial descriptions. In *Proceedings of EMNLP*, 2010.

[7] A. Goyal and H. Daumé III. Approximate scalable bounded space sketch for large data NLP. In *Empirical Methods in Natural Language Processing (EMNLP)*, Edinburgh, Scotland, 2011.

[8] D. Graff. English gigaword. In *Linguistic Data Consortium, Philadelphia, PA*, 2003.

[9] A. Gupta and L. S. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In D. A. Forsyth, P. H. S. Torr, and A. Zisserman, editors, *ECCV (1)*, volume 5302 of *Lecture Notes in Computer Science*, pages 16–29. Springer, 2008.

[10] A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Trans on PAMI*, 31(10):1775–1789, 2009.

[11] L. Jie, B. Caputo, and V. Ferrari. Who's doing what: Joint modeling of names and verbs for simultaneous face and pose annotation. In NIPS, editor, *Advances in Neural Information Processing Systems*, NIPS. NIPS, December 2009.

[12] D. Jurafsky and J. H. Martin. *Speech and Language Processing (2nd Edition) (Prentice Hall Series in Artificial Intelligence)*. Prentice Hall, 2 edition, 2008.

[13] H. Kress-Gazit, G. Fainekos, and G. Pappas. From structured english to robot motion. In *IEEE/RSJ Conference on Intelligent Robots and Systems*, San Diego, CA, 2007.

[14] V. Krüger, D. Herzog, Sanmohan, A. Ude, and D. Kragic. Learning actions from observations' robotics and automation magazine. *Robotics and Automation Magazine*, 17(2):30–43, 2010.

[15] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[16] C. Madden, M. Hoen, and P. Dominey. A cognitive neuroscience perspective on embodied language for human-robot cooperation. *Brain and Language*, 112:180–188, 2010.

[17] D. Marr. *Vision*. W.H. Freeman, San Francisco, CA, 1982.

[18] C. Matuszek, D. Fox, and K. Koscher. Following directions using statistical machine translation. In *Proceeding of the 5th ACM/IEEE International Conference on Human-Robot Interaction*, 2010.

[19] N. Mavridis and D. Roy. Grounded situation models for robots: Bridging language, perception and action. In *In Proceedings of the AAAI-O5 workshop*, pages 32–39, 2005.

[20] K. McKeown. Query-focused summarization using text-to-text generation: When information comes from multilingual sources. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation (UCNLG+Sum 2009)*, page 3, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[21] G. A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41, 1995.

[22] K. Pastra. *Vision-Language Integration: a Double-Grounding Case*. PhD thesis, Department of Computer Science, University of Sheffield, 2005.

[23] D. Roy and N. Mukherjee. Towards situated speech understanding: visual context priming of language models. *Computer Speech & Language*, 19(2):227–248, Apr. 2005.

[24] M. Thelen and E. Riloff. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 214–221, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[25] D. Traum, M. Fleischman, and E. Hovy. Nl generation for virtual humans in a complex social environment. In *In Proceedings of he AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue*, pages 151–158, 2003.

[26] J. Weng. Developmental robotics: Theory and experiments. *International Journal of Humanoid Robotics*, 1:199–236, 2004.

[27] Y. Yang, C. Teo, H. Daume, and Y. Aloimonos. Corpus-guided sentence generation for natural images. *EMNLP*, 2011.