Robots with Language: Multi-Label Visual Recognition Using NLP

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

Abstract— **There has been a recent interest in utilizing contextual knowledge to improve multi-label visual recognition for intelligent agents like robots. Natural Language Processing (NLP) can give us labels, the correlation of labels, and the ontological knowledge about them, so we can automate the acquisition of contextual knowledge. In this paper we show how to use tools from NLP in conjunction with Vision to improve visual recognition. There are two major approaches: First, different language databases organize words according to various semantic concepts. Using these, we can build special purpose databases that can predict the labels involved given a certain context. Here we build a knowledge base for the purpose of describing common daily activities. Second, statistical language tools can provide the correlations of different labels. We show a way to learn a language model from large corpus data that exploits these correlations and propose a general optimization scheme to integrate the language model into the system. Experiments conducted on three multi-label everyday recognition tasks support the effectiveness and efficiency of our approach, with significant gains in recognition accuracies when correlation information is used.**

I. Introduction

Recognition tasks for robots are not independent. They are correlated [6]. Researchers from computer vision and robotics disciplines have conducted a great amount of work utilizing the correlation between recognition tasks to boost individual performance, e.g. human identification and action [13], object and action [31]. However, correlation learned from hard-coded boolean charts or human labeling limits the possibility of deploying the approach into an real intelligent agent, aka a robot. In this paper, we show that the field of Natural Language Processing (NLP) has produced many tools that we can use to obtain the correlation. We discuss the different usages of language tools and how to combine natural language and vision tools in a robot.

Computational linguistics have created large text corpora and statistical tools so that we can obtain probability distributions for the co-occurrence of any two words, such as how likely a certain noun co-occurs with a certain verb. Here we present a framework to learn the correlation from corpus and apply it to several correlated multi-label recognition tasks. While we do not claim credits for introducing corpus statistics, we introduce a general way to combine it with vision. Experimental results on three different multi-label recognition tasks support our conjecture that the corpus guided method is able to boost recognition performance.

Classical and computational linguists are interested in modeling lexical semantics and have created resources where information about the conceptual meaning of lexical items and how these items relate to each other [5], such as "cause-effect" or "performs-functions", is organized. For example, the WordNet database [20] relates words through synonymy (words having the same meaning, like `argue` and `contend`) and hypernymy ("is-a" relationships, as between `car` and `vehicle`), among many others [21].

From another perspective, it is widely known that recognition tasks, like many other Artificial Intelligence tasks, require a high level knowledge base. Fanya Montalvo defined any such AI task as an AI-complete problem [18]. Tremendous amount of work has been devoted into building such a knowledge base, from early experts systems, to IBM's recent Watson project. In this work, we build a small database of common daily activities using linguistic tools. We show here how one can exploit the relationships in such a database to predict reasonable language labels for mining a large textual corpus dynamically.

In this paper, we focus on using as labels "nouns" (objects) and "verbs" (actions). Language and vision extract different information about the same entity, and we integrate them at the last computational stage in the recognition. However, we believe that the true power of using language and vision in combination will come through an integration at earlier stages. Both statistical and classical linguistic tools can provide in addition to correlation of labels, other information useful for visual recognition, such as spatial and temporal relations, and information on the visual appearance of objects and actions. We collectively call these information the entities' *attributes*. In some sense, language creates a multi-layer, hierarchical representation. Using this information, we can potentially address visual recognition as a problem of reasoning about the scene, instead of just classifying labels. For example, instead of recognizing a `hammer` using a classifier, we can verify that the segmented image patch "contains a wooden handle", "has a metal part" and "is next to a nail". Although we do not have all the tools yet to fully demonstrate these ideas, we include here some preliminary work on utilizing visual attributes for recognition.

II. Motivation

Human perception has a crucial but straightforward principle: *Principle of Totality* [6], which states that the conscious experience must be considered globally (by taking into account all the physical and mental aspects of the individual simultaneously) because the nature of the mind demands that each component be considered as part of a system of dynamic relationships. If any individual recognition task can be regarded as "some part of the system" here, then where

Computer Vision Lab, UMIACS, University of Maryland College Park, College Park, MD 20740, USA {yzyang,cteo,fer,yiannis}@cs.umd.edu

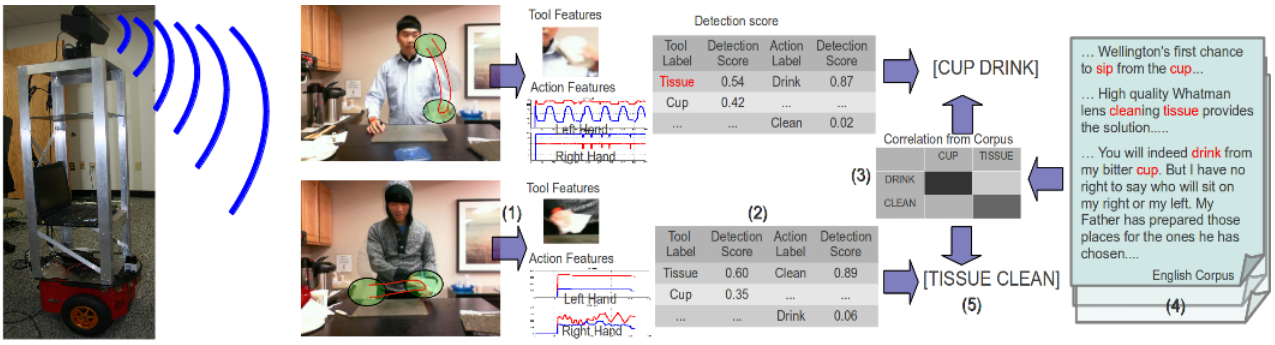


Fig. 1. Overview of framework for two sequences: (above) *drinking*, (below) *cleaning*. Hands, tools and action features (1) are extracted. From the visual detection scores (2), a correlation matrix (3) learned from a textual corpus (4) is used to improve the final label predictions (5).

do the "dynamic relationships" come from? Obviously for human beings, such dynamic relationships are from knowledge we accumulate through either learning or experience. We argue here that for a robot, such dynamic relationships can be obtained from linguistic resources, either from lexical databases and/or estimated by mining a corpus.

Fig. 1 shows a sketched overview of the proposed framework that uses the above mentioned principle. Consider the following scenario: A robot is observing a human being, e.g. drinking a cup of coffee (Fig. 1 left), and it is required to recognize the action and the tool. Current Computer Vision applications require that all entities, that the robot can distinguish, are predefined. Using our Knowledge database, which stores information about common daily activities, the robot can obtain the possible entities (objects and activities in a kitchen setting). Then the agent may be challenged from the uncertainty between a white cup or a white towel. However, by tracking the hand trajectories, it is confident that the action is "drinking" rather than "cleaning". If the agent has knowledge that humans normally use a cup to drink but not a towel, it should be easy to resolve the uncertainty. The knowledge needed here can be achieved by "teaching" the agent with several hard-coded propositional facts, such as [`drink` use `cup`], [`clean` use `towel`]. However, if we want to extend the knowledge, we have to enumerate all the possible facts required. On the other hand, such kind of knowledge is not deterministic: sometimes we do use a cup to clean by pouring water onto a dirty desk. The system should also assign a small possibility to [`clean` use `cup`]. Another way of learning is to learn the correlation from a large corpus automatically. Sentences like "*Wellington's first chance to sip from the cup*" are indicative of a strong correlation between `cup` and `drink`. Using NLP we can examine synonyms from WordNet to find out that `sip` is a form of `drink`. Sec. IV-A reports an experiment inspired from this scenario. We extend the framework to show how one can similarly exploit the correlation between scenes and objects in sec. IV-B. Finally, we know that the appearance of objects and actions can vary greatly, and visual object classifiers don't appear to be scalable. The problem is even harder for action recognition; descriptors robust to viewpoint and variation in movement have not been developed yet. Instead, as was advocated in the last part of sec. I, if we can obtain knowledge about the attributes of labels that do not change under environmental influences, we can use them in a reasoning process within the framework. This is demonstrated here for the case of object attributes in sec. IV-C.

In general, it is reasonable to expect the robot to learn the correlation between labels from the corpus and use it as an approximation of knowledge to guide the recognition task. Our proposed framework (sec. III) details how the corpus is mined, as well as the optimization method for integrating information from language and visual processing so as to help the robot achieve higher recognition accuracies when it is doing several correlated recognition tasks.

III. The Multi-Label Recognition Framework

Consider a multi-label recognition task with two sets of possible labels: $t_1 \in T_1$ and $t_2 \in T_2$, m and n is the size of the two label sets. Our framework first predicts the possible labels given some initial knowledge of the domain. From these labels, we show how we compute corpus statistics, essentially label correlations, to guide recognition in both T_1 and T_2. We then show how this framework can be generalized to a situation with three or more tasks.

A. Predicting Language Labels

An important prerequisite before corpus statistics can be computed is that we need to determine the relevant task labels that are appropriate for the dataset. The simplest and most direct approach, for e.g. in [28], is to *predefine* all the labels prior to any computation of the corpus statistics. While this may work for standalone evaluations of the proposed framework, a system that is fully autonomous should be able to make reasonable predictions of labels, given some knowledge of the domain. In this work, inspired by [22] we build a small ontology of relations inherent in common daily activities, with bidirectional relations all organized in a symmetric, labeled graph. Fig. 2 shows an example result of querying the knowledge base with the prior knowledge "kitchen" corresponding to one of our experimental datasets. The output is a list of relevant kitchen related tools, together with the possible associated actions linked to each tool.

In Fig. 2(a), "kitchen" is one of the subclasses (aka "is-a" relationship) under the superclass "scenes", and every associated tool is a subclass of "tools" in purple lines. In Fig. 2(b), the visualization shows every associated action with each tool in dotted yellow lines. Every concept in this knowledge base starts from the superclass "Thing". Although querying the knowledge base does not provide probabilities that we need directly, the list of possible tool and action labels can then be used as seeds to create the corresponding language model.

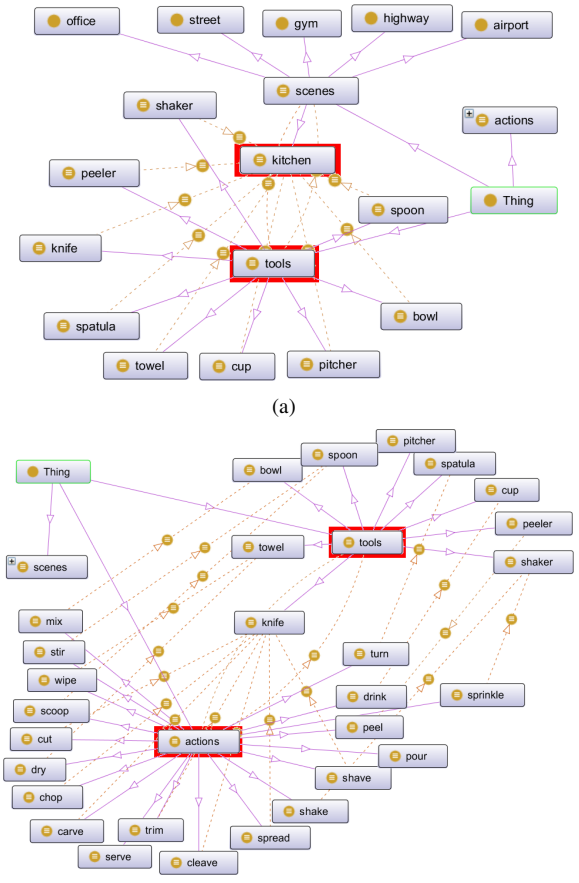


Fig. 2. Label prediction using the Embodied Knowledge Base. Purple lines indicate "subclass" relationships (aka "is-a") while dotted yellow lines indicate "association" relationships (one class is related to another class semantically). (a) The "kitchen" concept with its associated tools. (b) Various tools classes with their associated actions.

B. Correlation Mining

The key component of our approach is the language model that predicts the correlation between T_1 and T_2. We use the Gigaword Corpus [12] as a large text resource that contains the information. We do this by training a language model $P_L(T_1, T_2)$ that returns the maximum likelihood estimates of any label t_1 given the other label t_2. This can be done by counting the number of times t_1 co-occurs with t_2 in sentences in the corpus: $P_L(t_1|t_2) = \frac{\#(t_1,t_2)}{\#(t_2)}$.

As many English words share common meanings, a simple count of the words (labels) defined in T_1 or T_2 is likely to grossly underestimate $P_L(t_1|t_2)$. For example, in the Gigaword Corpus counting how often `drink` co-occurs with `cup` where the actual words are used will not be significantly larger than `pick` and `cup`. The reason is that `cup` can mean a normal drinking cup or a trophy cup. In order to ensure that P_L captures the correct sense of the word: nouns or verbs, we use WordNet to determine the synonyms and hyponymns of the words considered.

We then recompute P_L using these enlarged word classes to capture more meaningful relationships between the co-occurring words. Fig. 3(a) shows the $m \times n$ co-occurrence matrix of likelihood scores over the set of tools and actions considered in the UMD Sushi-Making dataset (sec. IV-A.1), denoted as $P_L(T_1|T_2)$ when we normalize the correlation scores over all labels in T_2. Similarly we obtain $P_L(T_2|T_1)$ when we normalize over all labels in T_1.

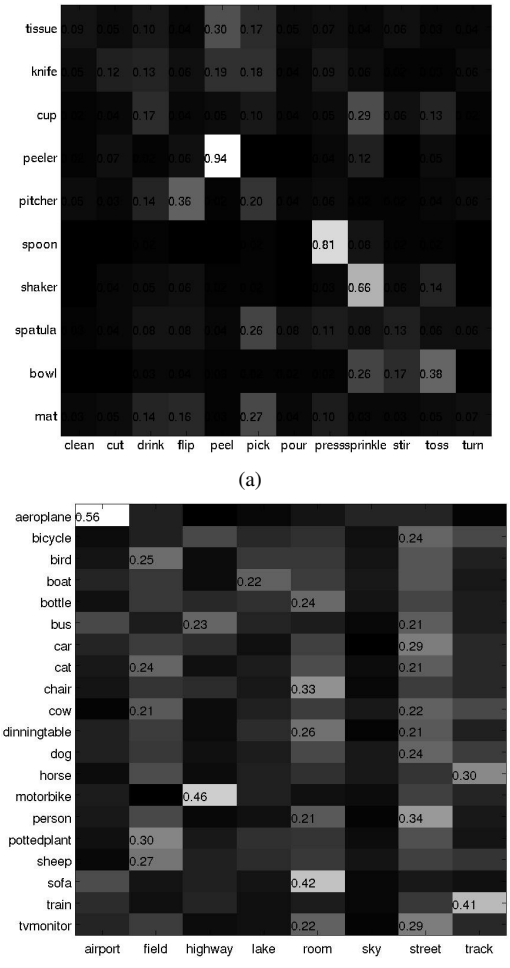


Fig. 3. (a): Gigaword co-occurrence matrix over tools and actions. (b): Gigaword co-occurrence matrix over objects and scenes

A quick analysis shows that for most of the tool classes, the predicted actions are correct (large values along the diagonals): e.g `peeler` predicts `peeling` with high probability (0.94). However, there are many co-occurrences which we could not anticipate: e.g. `sprinkling` has some synonyms such as `drizzle moisten splash splosh`

which have uses that are also close to `cup`, resulting in a higher score (0.29) versus `drinking` (0.17). Other mis-selected tools-action are also due to the confusion at the synonyms/hyponymns levels. We also notice that more *general* actions such as `picking` have a more uniform distribution across the tools, which is expected. Despite this simplistic model, most of the entries in P_L make sense – and it properly reflects the innate complexity of language. As will be shown in sec. IV-A.4, although the prior from language is weak, it is still helpful for the task of action and tool recognition. We used the same approach to extract the relationship between objects and scenes. Fig. 3(b) shows the $m \times n$ co-occurrence matrix of likelihood scores over the set of objects T_1 and scenes T_2 considered in Sec. IV-B.2.

C. Correlation Guidance

We use state-of-the-art features and machine learning techniques (specifically SVM) to train two classifiers on both recognition tasks that take in the labels and return a recognition confidence score on every label. The confidence scores are converted to probabilities using Platt's method [17]. We then normalize the scores over all possible labels to get $P_V(T_1)$ and $P_V(T_2)$, where the subscript P_V is used to denote probabilities from visual processing.

Together with $P_L(T_1|T_2)$ and $P_L(T_2|T_1)$, the joint probability of the multi-recognition task can be modeled as:

$$P(t_1, t_2) = P_V(t_1)P_L(t_2|t_1) = P_V(t_2)P_L(t_1|t_2). \quad (1)$$

We first focus on predicting t_1. We can get an estimated probability from language correlation, $P_L(t_1)$, by marginalizing over T_2:

$$P_L(t_1) = \sum_{t_2 \in T_2} P(t_1, t_2) = \sum_{t_2 \in T_2} P_V(t_2)P_L(t_1|t_2). \quad (2)$$

By introducing $\lambda \in [0, 1]$ as a regularization factor that controls the balance between the influence of visual detections P_V and corpus statistics P_L, as well as taking logs on both $P_V(t_1)$ and $P_L(t_1)$, we obtain the log-likelihood $\mathcal{L}$ of the labeling task by:

$$\mathcal{L}(t_1) = \log(P_V(t_1)) + \lambda \log(P_L(t_1)). \quad (3)$$

We can derive $\mathcal{L}(t_2)$ in a similar manner. The final label prediction pair (t_1^f, t_2^f) is then obtained by:

$$t_1^f = \arg\max_{t_1 \in T_1} \mathcal{L}(t_1) \text{ and } t_2^f = \arg\max_{t_2 \in T_2} \mathcal{L}(t_2) \quad (4)$$

As $|T_1|$ and $|T_2|$ are usually small, an exhaustive search over every label t_1 or t_2 is practical which guarantees a global optimal solution.

D. Generalizing to ≥ 3 Multi-Label Tasks

When generalizing the framework over three or more labels tasks, $(T_1, T_2 \ldots T_n)$, the cost of the marginalization step in eq. (2) increases exponentially and a naive brute force summation over all possible labels in T_n becomes impractical. In this case, we cast the problem of determining the optimal solution into a general graphical model optimization (Fig. 4(a)) where approximate inference methods, such as message passing [7] can be applied. This occurs, for e.g. when a third task, such as attributes of tools and actions, is added into the framework in which we solve a joint likelihood over the entire graphical model.

Further simplification can be achieved if some of the task labels are mutually *independent*, for e.g. object attributes and scenes, which allows us to solve the inference problem via a dynamical programming approach using Hidden Markov Models (HMM). Fig. 4(b) shows the HMM that combines P_L and P_V in a straightforward manner: the emissions correspond to P_V from visual processes (scenes, objects and attributes) and the transition probabilities P_L are obtained from the language model.

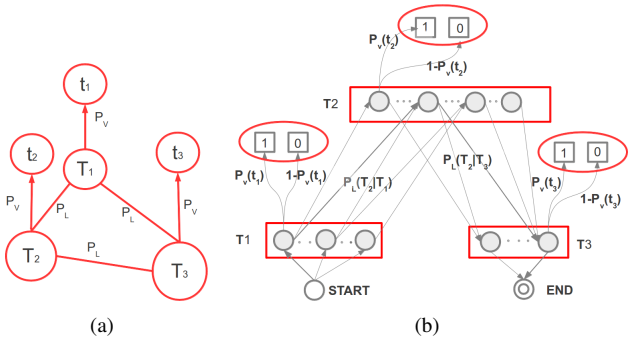


Fig. 4. (a): General case for a three labels task: P_V denotes probabilities from visual detection and P_L denotes correlation probabilities from corpus mining. (b): HMM model for three specific task labels: Scene-Object-Attribute are denoted as T_1, T_2, T_3 respectively.

IV. Experiments

A. Tools and Actions

In this set of experiments, we validate our proposed framework introduced in sec. III using a scenario where our robot observes humans making sushi. We collect the UMD Sushi-Making Dataset where we have computed the language models from initial seed labels predicted by our Knowledge Base.

1) The UMD Sushi-Making Dataset: The UMD Sushi-Making Dataset[1] [28] consists of 12 actions, performed by 4 actors using 10 different kitchen tools. This results in 48 video sequences each of around 1000 frames (30 seconds long). Other well known datasets such as the KTH, Weizmann or Human-EVA datasets [25], [11], [27] do not involve hand-tools. The dataset by Messing et al. [19] has only 4 actions with tool use. The CMU Kitchen Dataset [15] has many tool interactions for 18 subjects making 5 recipes, but many of the actions are blocked from view due to the placements of the 4 static cameras. Our Sushi-Making dataset provides a clear view of the actions and tools. The 12 actions are: cleaning, cutting, drinking, flipping, peeling, picking (up), pouring, pressing, sprinkling, stirring, tossing, turning. The tools are: tissue, knife, cup, rolling-mat, fruit-peeler, water-pitcher, spoon, shaker, spatula, mixing-bowl.

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

As was discussed in sec. III-B, some of the actions such as `picking` or `flipping` are extremely general and are easily confused. We made this choice to ensure that the language prediction P_L is *not* perfect and to show that our approach works even under noisy data.

2) Active tool detection strategy: We pursue an active strategy for detecting the relevant tools (denoted by T_1) in the video as illustrated in Fig. 5. This approach has two important benefits. By focusing our processing only on the relevant regions of the video frame, we dramatically reduce the chance that the tool detector will misfire. At the same time, by detecting the hand locations, we obtain the action trajectory, which is used to describe the action as shown in the next section.

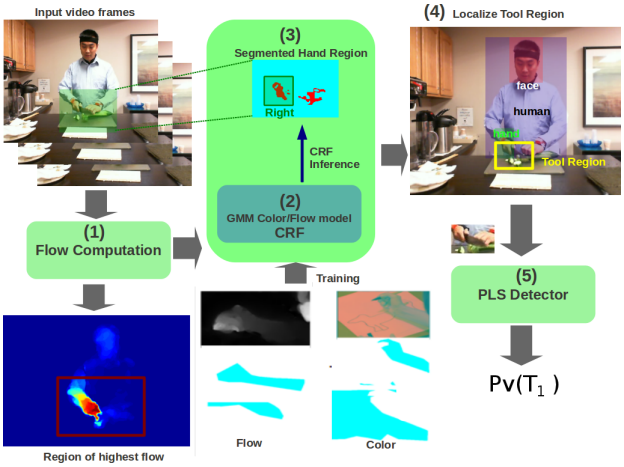


Fig. 5. Overview of the tool detection strategy: (1) Optical flow [2] is first computed from the input video frames. (2) We train a CRF segmentation model [24] based on optical flow + skin color. (3) Guided by the flow computations, we segment out hand-like regions (and removed faces if necessary) to 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 [26] (5), we compute a detection score $P_V(t_1)$, the probability that a tool $t_1 \in T_1$ exists given the video.

3) Action Recognition: Action labels are denoted as T_2 in this dataset. Tracking the hand regions in the video provides us with two sets of (left and right) hand trajectories as shown in Fig. 1. We then construct for every video a feature vector F_d that encodes the hand trajectories. F_d encodes the frequency and velocity components. Frequency is encoded by using the first 4 real components of the Fourier transforms of the position space in x- and y- direction which gives a 16-dim vector over both hands. Velocity is encoded by averaging the difference in hand positions between two adjacent frames $\langle \delta x \rangle$, $\langle \delta y \rangle$ which gives a 4-dim vector. These features are then combined to yield a 20-dim vector F_d. A SVM classifier is trained over these feature vectors to obtain the recognition score $P_V(t_2)$.

4) Results: A 4-fold cross validation was performed over the 48 videos of the Sushi-Making dataset in order to evaluate the effectiveness of our proposed approach. We first obtained the recognition accuracy using PLS (for the 10 tools) and Action Features + SVM (for the 12 actions) alone and used them as a baseline to highlight the improvement in recognition accuracies when eq. (3) and eq. (4) were applied over various values of $\lambda \in [0, 0.5]$ as shown in Fig. 6(a). Using corpus statistics, we obtained a relative improvement of 6% in recognition accuracy in both action and tool recognition compared to their baselines which had a combined average of 4.5%. As λ represents our confidence on the accuracy of the corpus-statistics versus the visual detections, different values of λ are expected to have different effects on the labels (tools or actions) considered. In this case, the language model is more biased towards tools and gives the largest improvement when λ is larger, compared to actions which have the opposite effect. Such seemingly divergent results can be explained from the inherent bias of language itself, where tools (and objects in general) have stronger correlations to specific (and limited) actions while many similar actions can be performed by numerous different tools, which results in a weaker (and hence more limited) effect on the action recognition accuracy.

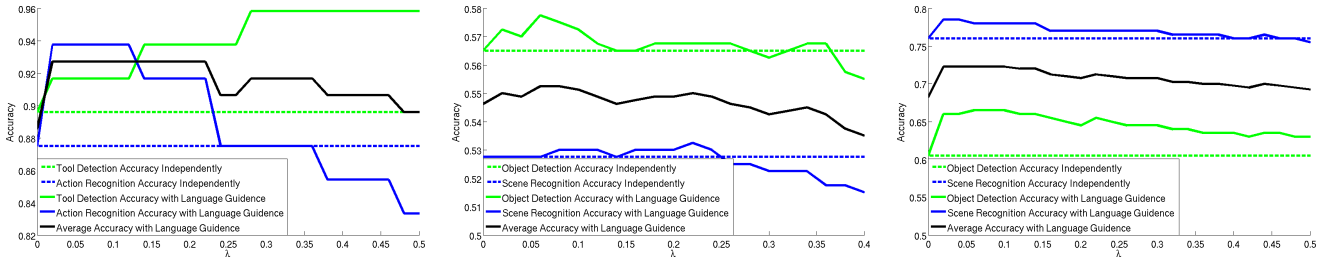
B. Objects and Scenes

In this set of experiments, we further evaluated our proposed framework for the scenario in which our robot is observing natural scenes with objects. We use a general large scale image dataset as testbed. As the images were taken from numerous domains, we found that it was easier to use the given ground truth labels of scenes and objects for this task, and we focus on showing the usefulness of the proposed approach in improving object and scene recognition accuracies.

1) SUN 20 scenes dataset: We evaluated the proposed approach using a subset of the SUN 20 scenes dataset [3]. The dataset comprises of 20 scenes and 127 objects from which we selected 1000 images. The large number of scenes and objects over a large variety of domains make this an extremely challenging dataset to evaluate the effectiveness of our approach in more general situations.

2) UIUC Pascal sentence dataset: In addition, we performed evaluations of our framework using the UIUC Pascal sentence dataset, first introduced in [10]. It contains 1000 images taken from a subset of the Pascal-VOC 2008 challenge image dataset, which are hand annotated with sentences that describe the image. The ground truth labels for objects and scenes are extracted from these sentences using the Berkeley Parser [23]. We divided the images into 8 distinct scenes [29] with 20 object classes defined in the Pascal-VOC 2008 challenge.

3) Results: For the SUN 20 scenes dataset, we randomly divided the 1000 images into a training set of 600 images with the remaining 400 as the testing set. Results are summarized in Fig. 6(b). We first extracted GIST features and used a SVM classifier over the 20 scene classes to obtain the baseline scene recognition accuracy. For objects, we chose the top 50 object classes from the original 127 which yielded the best detection scores over the training set, and determined the existence of the object in the test set by comparing it with annotated ground truth labels to obtain the baseline object recognition accuracies. The same parameters and trained



(a) (b) (c)

Fig. 6. Experimental results: (a) Action and Tool recognition *vs* individual recognition baselines on the UMD Sushi-Making dataset. (b) Scene and Object recognition *vs* individual recognition baselines on the SUN 20 scenes dataset. (c) Scene and Object recognition *vs* individual recognition baselines on the UIUC Pascal sentences dataset.

object models provided by the authors of the dataset were used in all experiments. We repeat the same experimental procedure with the UIUC Pascal sentence dataset (Fig. 6(c)), using the same train-test splits, GIST+SVM classifiers over 8 scene classes and pre-trained object models over the 20 object classes provided by the authors.

We evaluated our proposed framework and compared it to the baseline accuracies in the two datasets by varying λ. For the SUN 20 scenes dataset, we obtain a relative improvement of 2.6% (objects) and 1% (scenes) and an overall improvement of 1.4% (objects + scenes) over the baselines. The results are more significant in the UIUC Pascal sentence dataset where the relative improvements range from 10% (objects) to 3% (scenes) with an overall improvement of 8.1%. Although the task for these two datasets are the same (objects+scenes), the vastly different improvements from our approach highlights the need to define the *domain* of the task prior to computing the relevant corpus statistics. In the SUN 20 scenes dataset, the large number of scenes and object classes meant that the corpus statistics is *diluted* over all object/scene classes, limiting the improvement of our approach over the baselines. This effect is mitigated in the UIUC Pascal sentences dataset which has fewer object and scene classes. This further emphasized the importance of predicting the relevant task labels given the domain knowledge (sec. III-A) in order to maximize the benefits of using language. Finally, by comparing the results task in sec. IV-A.4, we note that the divergence due to λ between objects and scenes labels is not as significant. This indicates that objects and scenes have stronger correlations compared to actions and tools, which makes intuitive sense.

C. Scenes, Objects and Attributes

In the last experimental scenario, we require our robot to recognize scenes, objects as well as object attributes. The importance of recognizing objects through their *attributes* has recently received attention in Computer Vision [9]. Our proposed framework lends itself naturally to the use of such high-level knowledge since mining the correlation between attributes and objects is still doable from a large corpus, under the condition that the attribute labels are commonly used in daily life. By invoking the assumption that object attributes are independent of the scenes in which they occur,

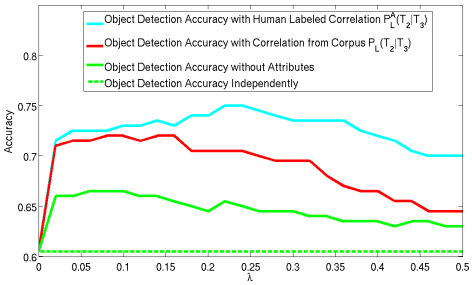


Fig. 7. Object recognition performance on the UIUC Pascal sentence dataset when object attributes are used.

we are able to model the problem using the HMM introduced in sec. III-D. We extend the experiments conducted over the UIUC Pascal sentence dataset in sec. IV-B.3 by adding human annotated attributes described in [9] as the third task label T_3 in the framework. As our focus is not on designing attributes detectors in this paper, we have allowed $P_V(t_3)$ to be the human annotated attributes. From the original list of 64 attributes, we first select the top 20 semantically meaningful attributes by hand (discarding those that have little or no relevance to the 20 object classes) and then compute a new language model $P_L(T_2|T_3)$ that relates these attributes to the object classes. By considering the human annotated attributes as a form of "attribute corpus" we compute $P_L^A(T_2|T_3)$ which represents the upper-bound on the language model that we can expect from mining the corpus. The object recognition accuracy results are summarized in Fig. 7. The first two plots (in green) are from Fig. 6(c), where no attribute information was used. By adding attributes in the framework, we obtained an additional improvement of 10% and 13.6% for $P_L(T_2|T_3)$ and $P_L^A(T_2|T_3)$, respectively compared to the case when no attributes were used. The strength of using attributes for object recognition are clear here: we are able to consistently improve upon the baseline (when no attributes are used) and when a cleaner corpus that relates objects and attributes directly are used.

V. Related Works

Our work is mostly related to Yang et al. [30], in which they showed that correlation learned from corpus can guide descriptions of natural images. Kulkarni et al. [14] proposed to generate simple image descriptions by designing various

Conditional Random Field models. Both studies focused on generating descriptions of images, while in our work, we focus on using correlation and graphical models to study fundamental recognition tasks. Additionally, advances in Natural Language Processing and Computer Vision have lead to several works that have focused on using sources of data that are readily available "in the wild" to analyze static images. The seminal work of Duygulu et al. [8] showed how nouns can provide constraints that improve image segmentation. Berg et al. [1] processed news captions to discover names associated with faces in the images, and Jie et al. [13] extended this work to associate poses detected from images with the verbs in the captions. Some studies also considered dynamic scenes. [4] studied the aligning of screen plays and videos, [16] learned and recognized simple human movement actions. These recent works had shown that exploiting co-occurring text information from scripts and captions aids in the visual labeling task. Our paper takes this further by using *generic* text obtained from the Gigaword corpus [12]. As was shown in the preceding sections, by using NLP tools, we can still derive useful correlations for multi-label recognition tasks.

VI. Conclusion

A framework has been introduced which integrates NLP tools for guiding multi-label visual recognition tasks. We applied it to three different real world tasks for our robot: 1) tools+actions recognition on the UMD Sushi-Making video dataset, 2) objects+scenes recognition on the SUN 20 scenes and UIUC Pascal sentences image datasets and 3) scenes+objects+attributes recognition on the UIUC Pascal sentences dataset. The experimental results reported support the effectiveness of our framework compared to baselines using visual processing alone. We also explored how much we should trust corpus statistics by adjusting a regularization parameter that balances our confidence on the accuracy of the corpus statistics versus the visual detections. In addition, the experiments also highlighted the need to have a pre-defined domain to constrain the corpus statistics and the usefulness of adding attributes in improving object recognition performance. In future work, we intend to explore extracting information about the spatial and temporal relations of objects in the scene (the prepositions) and about attributes of objects and verbs (adjectives, part descriptions of nouns and adverbs) from language. Using these information we can explore improving visual recognition at earlier stages in the computation. For example, we can use this information in segmentation, and for learning labels from both vision and language information, and reformulate visual recognition as a reasoning process.

VII. Acknowledgements

The support of the European Union under the Cognitive Systems program (project POETICON++) and the National Science Foundation under the Cyberphysical Systems Program is gratefully acknowledged. Y. Yang and C. Teo are supported in part by the Qualcomm Innovation Fellowship.

References

[1] T. L. Berg, A. C. Berg, J. Edwards, and D. A. Forsyth. Who's in the picture? In *NIPS*, 2004.
[2] T. Brox, C. Bregler, and J. Malik. Large displacement optical flow. In *CVPR*, pages 41–48. IEEE, 2009.
[3] M. J. Choi, A. T. J. Lim, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. *CVPR*, 2010.
[4] T. Cour, C. Jordan, E. Miltsakaki, and B. Taskar. Movie/script: Alignment and parsing of video and text transcription. 2008.
[5] D. A. Cruse. *Lexical semantics*. Cambridge, England: University Press, 1986.
[6] A. Desolneux, L. Moisan, and J. M. Morel. *From Gestalt Theory to Image Analysis*, volume 34. 2008.
[7] J. Domke. Parameter learning with truncated message-passing. *CVPR*, 2011.
[8] 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 *ECCV*, 2002.
[9] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. *CVPR*, 2009.
[10] 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 *ECCV*, 2010.
[11] L. Gorelick, M. Blank, E. Shechtman, R. Basri, and M. Irani. Actions as space-time shapes. *PAMI*, 29(12):2247–2253, 2007.
[12] D. Graff. English gigaword. In *Linguistic Data Consortium, Philadelphia, PA*, 2003.
[13] L. Jie, B. Caputo, and V. Ferrari. Who's doing what: Joint modeling of names and verbs for simultaneous face and pose annotation. NIPS. NIPS, December 2009.
[14] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. Berg, and T. Berg. Baby talk: Understanding and generating simple image descriptions. *CVPR*, 2011.
[15] F. D. la Torre, J. Hodgins, J. Montano, S. Valcarcel, R. Forcada, and J. Macey. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. Technical report, CMU-RI-TR-08-22, Robotics Institute, Carnegie Mellon University, July 2009.
[16] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.
[17] H.-T. Lin, C.-J. Lin, and R. C. Weng. A note on platt's probabilistic outputs for support vector machines. *Mach. Learn.*, 68:267–276, October 2007.
[18] J. C. Mallery. Thinking about foreign policy: Finding an appropriate role for artificially intelligent computers. 1988.
[19] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *ICCV 09*, 2009.
[20] G. A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41, 1995.
[21] G. A. Miller and C. Fellbaum. *WordNet then and now*, volume 41 of *Language Resources and Evaluation*, pages 209–214. Springer, 2007.
[22] K. Pastra. Praxicon: a grounded, compositional & generative concept world. *the 4th International Conference in Cognitive Systems (CogSys)*, 2010.
[23] S. Petrov and D. Klein. Improved inference for unlexicalized parsing. In *Proceedings of HLT-NAACL*, 2007.
[24] C. Rother, V. Kolmogorov, and A. Blake. "grabcut": interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, 2004.
[25] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, 2004.
[26] W. Schwartz, A. Kembhavi, D. Harwood, and L. Davis. Human detection using partial least squares analysis. In *International Conference on Computer Vision*, 2009.
[27] L. Sigal, A. O. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1-2):4–27, 2010.
[28] C. L. Teo, Y. Yang, H. Daume, C. Fermuller, and Y. Aloimonos. A corpus-guided framework for robotic visual perception. *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.
[29] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *ICCV*, pages 273–280. IEEE Computer Society, 2003.
[30] Y. Yang, C. Teo, H. Daume, and Y. Aloimonos. Corpus-guided sentence generation for natural images. *EMNLP*, 2011.
[31] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, June 2010.