Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Vision-Language Fusion for Object Recognition

Sz-Rung Shiang, Stephanie Rosenthal, Anatole Gershman, Jaime Carbonell, Jean Oh
School of Computer Science, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, Pennsylvania, 15213
sshiang@andrew.cmu.edu, {srosenth, anatoleg, jgc, jeanoh}@cs.cmu.edu

Abstract

While recent advances in computer vision have caused object recognition rates to spike, there is still much room for improvement. In this paper, we develop an algorithm to improve object recognition by integrating human-generated contextual information with vision algorithms. Specifically, we examine how interactive systems such as robots can utilize two types of context information–verbal descriptions of an environment and human-labeled datasets. We propose a re-ranking schema, MultiRank, for object recognition that can efficiently combine such information with the computer vision results. In our experiments, we achieve up to 9.4% and 16.6% accuracy improvements using the oracle and the detected bounding boxes, respectively, over the vision-only recognizers. We conclude that our algorithm has the ability to make a significant impact on object recognition in robotics and beyond.

1 Introduction

The growth of social media and crowdsourcing platforms has opened access to compound descriptions for images in addition to simple labels, *e.g.*, textual descriptions associated with an image posted on social media may contain contextual information beyond the labels of image regions. An ability to digest this type of context-rich information in a perception system can be extremely useful in problem domains such as disaster response where humanitarian volunteers assess damages by looking through a plethora of images of an affected area and textual descriptions from social media (Hörler 2014). In addition, robots interacting with humans via natural language would also need such an ability to integrate what has been seen and what has been told. In this context, our research broadly addresses a problem of fusing information available in various modalities including vision and language to provide enhanced overall perception. Toward this general goal, this paper is specifically focused on fusing information from three types of sources–*i.e.*, computer vision, textual descriptions and Web data mining–for the object recognition problem.

For example, consider a robot that can recognize certain objects using a computer vision system, albeit with an imperfect accuracy. Suppose that a human issues a command to the robot *"Pick up the cup on the dishwasher."* In order to perform the given command, the robot needs to recognize those objects mentioned in the command, namely *cup* and *dishwasher*, in its environment. Here, we investigate how to improve the recognition performance after receiving the command, by utilizing the information embedded in the command itself, *e.g.*, a spatial relation between the two objects. We also take advantage of additional information mined from the Web, where extra human labeling is not required during run time. For example, from user-tagged images on the Web, we can learn that certain objects tend to co-occur frequently, *e.g.*, *dishwasher* and *stove*.

For the learning and the interpretation of spatial relations, we use an approach described in (Boularias et al. 2015) that has been extensively evaluated on ground robots for semantic navigation in unknown outdoor environments (Oh et al. 2015; 2016); the details are omitted here due to space limitation. The main focus of this paper is on efficiently integrating several types of information for better overall perception, reporting how much improvement can be achieved specifically on the object recognition task.

We take a probabilistic approach to fuse such information. We introduce **MultiRank**, an information fusion algorithm that uses the label probabilities for bounding boxes obtained from computer vision (CV) as priors and computes posteriors based on object co-occurrence statistics and verbal descriptions. We create a multi-layered graph from the bounding boxes to represent general co-occurrence relationships between labels and also spatial relations specific to each image. We then use a Random Walk algorithm (Hsu, Kennedy, and Chang 2007) to solve for the object label for each bounding box. We show that our algorithm increases the accuracy of the vision-only algorithm by 9.41% and 16.67% in the oracle (ground-truth) and detected (Lin, Fidler, and Urtasun 2013) bounding box cases, respectively, on the NYU Depth V2 datasets (Silberman et al. 2012). For the objects that are mentioned in commands (that are thus more relevant to the task), further improvement is observed; the accuracy improves by 15.24% and 17.46% in the oracle and detected bounding box cases, respectively. We conclude that our vision-language fusion approach for incorporating contextual information from humans significantly improves the performance of object recognition over the vision-only

algorithms. While our experiments are carried out on an indoor image dataset, our approach could easily be applied to more practical applications such as aforementioned disaster scenarios.

In the rest of this paper, after reviewing related works, we describe our approach for leveraging verbal descriptions in Section 3 and human-labeled datasets in Section 4. We present our MultiRank algorithm for fusing this information along with computer vision results in Section 5 and we report the experimental results in Section 6.

2 Related Work

The last decade has seen steep progress in computer vision based object detection due to deep learning algorithms. The mean average precision is reaching 70% on PASCAL data sets (Ren et al. 2015). While this is a significant improvement from previous state-of-the-art approaches such as Deformable Part Model (Felzenszwalb et al. 2010) whose results were below 50%, further improvement is anticipated especially when object recognition is used to produce actionable results. For instance, a recognition system with 0.7 precision and 0.6 recall may mean a failure rate of 0.58 at a task performance level accounting for 40% miss and another 30% false positive errors ($0.4 + 0.6 * 0.3 = 0.58$).

Human-in-the-loop approaches in perception are not new (Pavlick and Callison-Burch 2015; Sarma et al. 2015; Salisbury, Stein, and Ramchurn 2015), as human perception excels at solving complex problems such as recognizing objects in the blurred images (Deng, Krause, and Fei-Fei 2013) and classifying celestial bodies (Kamar, Hacker, and Horvitz 2012). Despite the benefits of human-in-the-loop methods, it is less intuitive to ask humans to label objects directly in the applications (Russakovsky, Li, and Fei-Fei 2015). Using natural language is a more intuitive way for humans to communicate with a system, *e.g.*, describing an environment in a written or verbal format (Siddiquie and Gupta 2010) or commanding a robot to perform a task (Kong et al. 2014).

In the computer vision community, recent works indicate that contextual cues, *e.g.*, auxiliary information about the scene (Aboutalib 2010; Divvala et al. 2009), can help improve recognition results when the local pixel information is not reliable. Existing approaches utilize co-occurrence of objects in the training data set (Oliva and Torralba 2007; Lawson, Hiatt, and Trafton 2014) and spatial relations between those objects (Divvala et al. 2009; Choi et al. 2010).

In (Choi et al. 2010), co-occurrence and spatial priors are jointly learned in their context model to improve object recognition. Furthermore, (Mottaghi et al. 2014) made use of both global and local spatial relations, which improves on PASCAL dataset from 26.6% to 30.8% using 33 static context variables. The contextual features used in these approaches are extracted from images directly. Instead, we focus on human-generated information that can be mined from the Web offline or that can be acquired via interacting with people in a shared environment (Kaiser et al. 2014). The idea of incorporating additional modalities to improve perception has been recently studied in a simple game setting (Thomason et al. 2016) where they demonstrated the improvement in F1-score up to .354 from their vision-only system's score of .196.

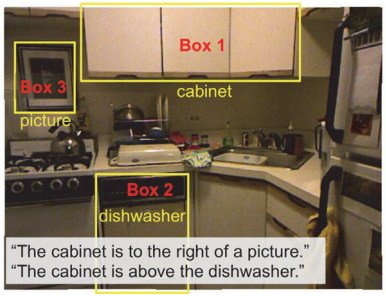

Figure 1: An example scene with verbal descriptions.

Our approach for fusing contextual information is related to graph-based ranking algorithms such as PageRank (Page et al. 1999) and Random Walks (Ipsen and Kirkland 2005; Hsu, Kennedy, and Chang 2007), which have been proposed as a re-ranking schema given some first-pass ranking results (*i.e.*, the output of a computer vision algorithm). The Random Walk algorithm fuses contextual knowledge into a graph structure, and then re-ranks scores based on both the first-pass results and the context. Layered graphs similar to ours have been successfully used in speech-related applications (Lee et al. 2014; Chen, Wang, and Rudnicky 2015). In this technique, each modality's recognition scores are modeled as nodes in one layer of the graph. The nodes are connected between the layers, and scores propagate across the layers for re-ranking. This schema is an efficient way to jointly model multiple types of heterogeneous information. In (Chen and Metze 2013), an intermediate layer is added between two layers of different modal information, and the scores of one layer are updated from another layer through projection from (and to) this centric layer. In these prior works, the graph edges reinforce strong nodes by increasing the strength of neighbors both within and between layers. By contrast, in our proposed graph designed for object recognition, nodes within a layer compete for strength based on the constraint that there exists only one correct label per object, at the same time reinforcing the nodes in other layers that are strongly linked.

3 Leveraging Verbal Descriptions

When compared to algorithms today, humans exhibit superior perception skills. Effortlessly, people can instantly segment a complex scene into a set of disconnected objects, recognize familiar objects, and classify newly seen objects into known categories. To benefit from human inputs, the systems must be able to parse and understand people's descriptions of objects in the environment. We note that speech recognition and natural language parsing are outside the scope of this paper. Instead, we use a structured language for describing spatial relations to focus on how the semantic meanings of a verbal description can be interpreted and understood by our system. The relations used in our experiments are: *left, right, above, below*, and *on*. We use the camera location as the default frame of origin, if not specified,

when reasoning about a spatial relation.

We use the following simple grammar for describing a binary spatial relation:

```
<relation> (<subject>, <object>)
```

in which the subject has the particular relation to the reference object of the description. For instance, the verbal descriptions in Figure 1 can be written as:

```
right(cabinet, picture)
above(cabinet, dishwasher).
```

4 Mining Human-Labeled Datasets

In addition to human descriptions of the environment, we explore online data sources that could provide common sense information such as objects that commonly exist in certain environments, *e.g.*, a dishwasher and a stove are found commonly in a kitchen environment. In this work, we specifically look for image databases that can provide object labels.

Most of online image databases, such as Shutterstock or Flickr, store tag information of labels per image, *i.e.*, an image is associated with a set of object labels relevant to that image without further notion of object-level segmentation. We, therefore, focus on the *co-occurrences* of object labels in these tag lists to model the conditional likelihood of two objects occurring in the same scene; for instance, the probability of seeing a *dishwasher* should be increased if a *stove* has been seen nearby. We carefully evaluated several publicly-available image databases to find "well-labeled" images and used the same labels as in our vison-only algorithms. For each label in the label set, we download the top 500 ranked images and their tag label lists. For each label list, we look for pairs of labels that are in our label set and record them in a co-occurrence matrix. Following the same format used for verbal description, the co-occurrence information can be represented as a special relation between a pair of objects, for example,

```
co-occur (dishwasher, stove).
```

5 MultiRank Algorithm

We introduce **MultiRank**–a graph-based algorithm for fusing information. MultiRank takes as input a pre-trained CV algorithm, the image to be recognized, and parameters α, β to determine how much to trust the initial CV results versus the auxiliary non-visual context. We first run the CV algorithm on the input image to produce a set of bounding boxes around objects and the probabilities $F^{(0)}$ of each label N for each object. We then collect human inputs in a verbal form, *e.g.*, dialogs, online data, etc. Given this information, we construct the multi-layer[1] graph structure leveraging spatial relations and object co-occurrence statistics. By iteratively running Random Walk over the MultiRank graph, we compute re-ranked labels $F^{(t)}$ for each bounding box. These labels leverage both the existing CV algorithm and additional contextual information.

[1]Since a layer here represents a bounding box, we use 'boxgraph' and 'layer' interchangeably in this paper.

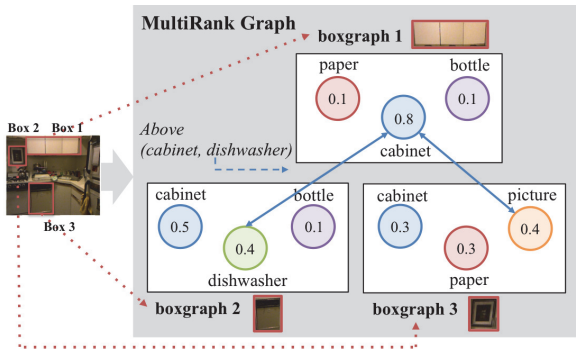

Figure 2: Illustration of an image and the bounding boxes in the corresponding MultiRank graph.

5.1 Constructing MultiRank Graph

MultiRank creates a graph that is organized as multiple smaller graphs called *boxgraphs* as shown in Figure 2. Each boxgraph represents one bounding box (one object) that was returned by the CV algorithm. The nodes within a boxgraph represent the candidate labels for that box and are assigned initial label probabilities $F^{(0)}$ from the CV algorithm. Each boxgraph can include all candidate labels or fewer select labels that have the highest probability as depicted in Figure 2. The nodes within a boxgraph are completely connected through *within-boxgraph* edges. After creating the boxgraphs, *between-boxgraph* edges are added to connect the nodes between every pair of boxgraphs, resulting in a fully connected overall graph. All edges are initially assigned weight 0 but will be assigned a weight based on human-generated context information. Formally, a MultiRank graph $G = \langle \vec{L}, E_B \rangle$ is a tuple of a vector of boxgraphs $\vec{L}$ and a between-boxgraph edge matrix E_B. Each boxgraph $L \in \vec{L}$ is a triple $L = [N, E_W, F]$ where N denotes a set of nodes; E_W, within-boxgraph edges; and F, a score vector of that boxgraph. Let O denote a set of object labels. Notationally, we say that F_l specifies the score vector F for boxgraph l; $F_l[o]$ is the F score of the node representing label $o \in O$ in boxgraph l. Similarly, $n_l[o]$ is the node representing object label o in boxgraph l, *e.g.*, in Figure 2, $n_2[\textit{dishwasher}]$ refers to the node in the center in boxgraph 2, and its score $F_2[\textit{dishwasher}]$ is 0.4.

5.2 Edge Weights as Object Relations

Iteratively, in MultiRank, the nodes (candidate labels) that are consistent with human-described relationships absorb scores from those nodes that are not, moving up in their ranks. For example, suppose that a person said "The cabinet is above the dishwasher" as depicted in Figure 2. The bounding box represented by boxgraph 2 is initially misclassified (*i.e.*, the label with the highest score is not the correct label, *dishwasher*). After receiving the human's description, however, node *dishwasher* starts absorbing scores from its neighbors in boxgraph 2 because the edge between that node and node *cabinet* in boxgraph 1 matches the description.

For each spatial relation $\phi(i, j)$ that a human provides,

e.g., above (cabinet, dishwasher), MultiRank enumerates all the possible pairs of bounding boxes or boxgraphs (l_1, l_2) that fit relation ϕ and check whether the nodes $n_{l_1}[i]$ and $n_{l_2}[j]$ representing object labels i and j in l_1 and l_2 respectively are present in the graph. In the example above, if bounding box 1 is *above* bounding box 2 and there is *cabinet* as a candidate label in boxgraph 1 and node *dishwasher* in boxgraph 2, then the node pair is added into the relations matching set $R = \{\phi(n_1[\textit{cabinet}], n_2[\textit{dishwasher}])...\}$.

After matching all possible relations to the bounding boxes and nodes, the algorithm assigns weights to both within and between edges, denoted by $w \in \vec{E}_W$ and $b \in \vec{E}_B$, respectively, as follows. For each relation $\phi(n_{l_1}[i], n_{l_2}[j])$, all edges within boxgraph l_1, denoted by $w_{l_1}[o, o']; o, o' \in O$, directed *towards* $n_{l_1}[i]$ are updated as the F score of $n_{l_1}[i]$:

$$\forall k \in n_{l_1}, w_{l_1}[k, i] = F_{l_1}^{(0)}[i]. \quad (1)$$

Similarly, all between-edges in l_2 directed at $n_{l_2}[j]$ are updated to the value $F_{l_2}^{(0)}[j]$. Given the same relation $\phi(n_{l_1}[i], n_{l_2}[j])$, the two between-boxgraph edges $b_{l_1,l_2}[i, j]$ and $b_{l_2,l_1}[j, i]$ from $n_{l_1}[i]$ to $n_{l_2}[j]$ and vice versa are updated as the *source's* F score:

$$b_{l_1,l_2}[i, j] = F_{l_1}^{(0)}[i] \text{ and } b_{l_2,l_1}[j, i] = F_{l_2}^{(0)}[j] \quad (2)$$

Following our cabinet-dishwasher relation example, Figure 3 shows a bold edge between node *cabinet* in boxgraph 1 (blue node) and node *dishwasher* in boxgraph 2 (green node) as a matching relation. For within-boxgraph edges, every node in boxgraph 1 transfers 0.8–*i.e.*, the score from CV denoted by $F_1^{(0)}[\textit{cabinet}]$–of its score to $n_1[\textit{cabinet}]$. Similarly, every node in boxgraph 2 transfers 0.4 of its score to $n_2[\textit{dishwasher}]$. Next, two between-boxgraph edges are updated: $n_1[\textit{cabinet}]$ transfers 0.8 of its score to $n_2[\textit{dishwasher}]$ (edge $b_{1,2}[\textit{cabinet, dishwasher}] = 0.8$) and $n_2[\textit{dishwasher}]$ in boxgraph 2 transfers 0.4 of its score to $n_1[\textit{cabinet}]$ in boxgraph 1 (edge $b_{2,1}[\textit{dishwasher,cabinet}] = 0.4$). Weight of dotted links in the graph are set to 0.

We note that the graph generated follows the Markov assumption; therefore, the scores iteratively converge under a random walk algorithm. As in a Markov chain, the edge matrix is normalized as a probabilistic transition matrix, *i.e.*, each column sums to 1. In addition, the score vector is also normalized to sum to 1.

5.3 Iterative Convergence Using Random Walk

Intuitively, one should interpret the graph and its flow of scores as follows: The nodes without matching relations will propagate their scores to other nodes while the nodes with matching relations will preserve their scores, resulting in a graph that is biased to rank those nodes that are consistent with given relations higher than others.

The F scores flow between nodes and across edges iteratively until each converges. We update the score vector for each boxgraph F_l:

$$F_l^{(t+1)} = \alpha F_l^{(0)} + (1-\alpha) E_{W,l} \cdot \{\beta_l F_l^{(t)} + \sum_{l_2 \in R_l} \beta_{l_2} E_{B,l_2,l} \cdot F_{l_2}^{(t)}\} \quad (3)$$

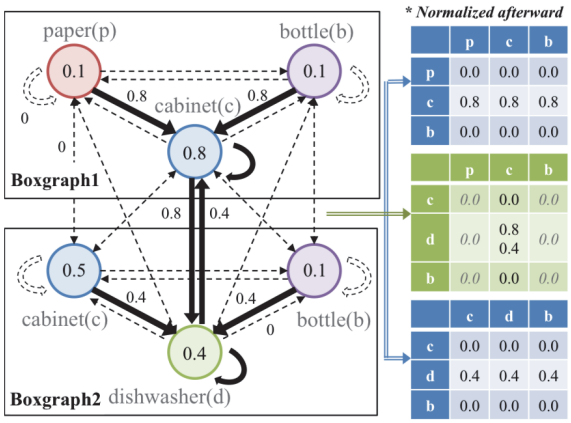

Figure 3: An example showing the edges between and within boxgraphs for a pair of boxgraphs that matches a given spatial description. The values shown here are before they have been normalized.

where $F_l^{(t)}$ is the score vector for boxgraph l at the t^{th} iteration; $E_{B,l_2,l}$, the transition matrix from boxgraph l_2 to boxgraph l; α, a parameter to balance between the initial CV probabilities and the updated probabilities that use the human input; and β, defined as weight to balance the influence of different boxgraphs as follows:

$$\beta_l = max(F_l^{(0)}) / \sum_{l_2 \in \{l, R_l\}} max(F_{l_2}^{(0)}) \quad (4)$$

where R_l are the relations that connect boxgraph l to all other boxgraphs l_i. The intuition for β is that the boxgraphs with higher CV scores should be more reliable sources for updating scores. We iteratively update the score vectors until convergence (Ipsen and Kirkland 2005). The final F scores are MultiRank's re-ranked results in an effort to improve the initial CV recognition model.

5.4 Edge Weights with Even More Context

The edge weights represent relationships between label probabilities for a single object and between objects. There are many more options for adding contextual information to the MultiRank graph in addition to spatial relations. In our experiments, we also used co-occurrence data from human-labeled online sources as well as the confusion matrix of the vision-only algorithm to update edge weights. We apply the co-occurrence matrix into the between-boxgraph weight assignment as illustrated in Figure 4. For example, the edges between node i in boxgraph 1 and node j in boxgraph 2 could be weighed according to the (i, j) element in the co-occurrence matrix. Similarly, confusion matrix shows which object labels are less likely to be misclassified. We update the within-boxgraph weights to represent the likelihood of an object being misclassified by the first-pass CV algorithm.

6 Experiments

We evaluate MultiRank with human-generated data against a vision-only model, one of the recent computer vision algo-

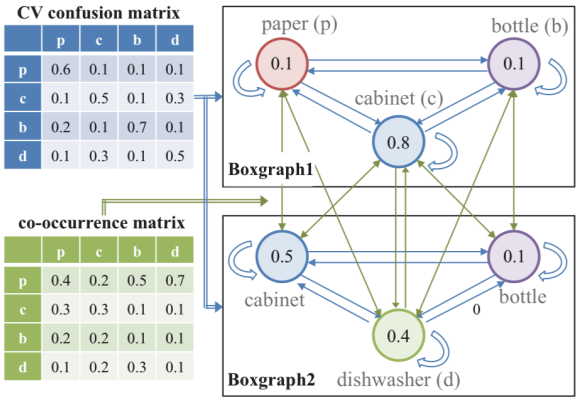

Figure 4: Illustration of weight assignment for within-boxgraph and between-boxgraph edge matrices using context information.

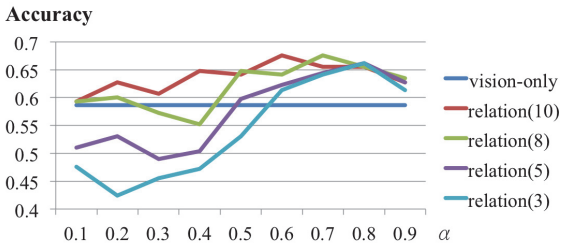

Figure 5: Accuracy under varying α values on validation set.

	top 1	top1/top2 ratio
correct	0.5662 (0.2283)	13.87 (20.81)
wrong	0.3134 (0.1441)	2.62 (2.83)

Table 1: Correlation between the CV algorithm's confidence values and the actual recognition output: The top 1 is the highest confidence score distribution (mean, standard deviation) and the top1/top2 ratio is the ratio between the highest and the second highest values.

rithms that will be described in the following sections. We validate our algorithm on the NYU depth dataset which contains many indoor scenes such as kitchens and hundreds of different objects. For evaluation metrics, we use accuracy and mean average precision (mAP) for recognition tasks, and F1-score for a detection task as in Figure 8. In the following subsections, we describe the experimental settings and report the results.

6.1 Image Dataset

The NYU depth dataset (Silberman et al. 2012) is composed of 1449 scenes and 894 kinds of objects. Images in the dataset are filled with complex environments frequently consisting of more than 20 labeled, overlapping and occluded objects. In each image, RGB values for each pixel and also depth information are provided, as are the ground truth object positions; bounding boxes are labeled with the true object names. For our experiments, we removed those labels that occurred fewer than 50 times, resulting in 74 possible labels for each bounding box. While typically this would mean that our MultiRank algorithm creates 74 nodes for each box, we reduced the number of label nodes to the top 20 for each bounding box in order to reduce noise in the prediction results.

6.2 Computer Vision (CV) Algorithm

This section describes the vision-only object recognition system used in our experiment. Object recognition can be decomposed into two subtasks: object (bounding box) detection and object classification. In the experiments, we used both the ground truth and detected bounding boxes.

Object Detection To detect bounding boxes, we used the Constrained Parametric Min-Cuts (CPMC) algorithm (Lin, Fidler, and Urtasun 2013) (Carreira and Sminchisescu 2012) on 2D and 3D information[2]. The classifier was trained on 795 images from NYU depth dataset and tested on the rest of images. We selected only foreground objects with high occurrence frequency for training the object detection model, resulting in overall 21 labels. The bounding box candidates are counted as recalled if the intersection over union (IOU) is higher than 50%, and the recall rate of the object detection is in the 70% range. For each image we extract top 30 bounding boxes according to the likelihood.

Object Classification Given a set of bounding boxes, we used a classifier to assign labels to them. We used Caffe (Jia et al. 2014) to extract the fully-connected layer, known as fc7 features (2D image features), in Alexnet pre-trained on ILSVRC 2012(Krizhevsky, Sutskever, and Hinton 2012) for each bounding box. Using the fc7 features, we trained an SVM classifier (Chang and Lin 2011) to categorize each box into 74 object labels. We note that, due to the limited number of training data, we used an SVM instead of deep learning as the classifier for our task. Using 5-fold cross validation, this vision-only model achieves an accuracy of 0.6299 and mAP 0.7240 in the ground-truth bounding box case and accuracy 0.4229 and mAP 0.2820 in the detected bounding box case.

6.3 Human-generated Information

For the spatial relations, we randomly selected 40 images out of 1449 images and manually labeled each image with 10 spatial relations; 10 additional images were used for validation to tune the parameter α in Equation (3). For human-labeled online co-occurrence, we collect image labels from Shutterstock in which label lists are well curated. Using the 74 labels from NYU depth dataset, we downloaded up to 500 images matching each label. For each of the 74×500 images, we downloaded the complete human-generated label list. Then, we counted the frequency of co-occurrence of every possible pair of labels across each label list and record it in a co-occurrence matrix.

[2]We note that 3D information is used only for detection.

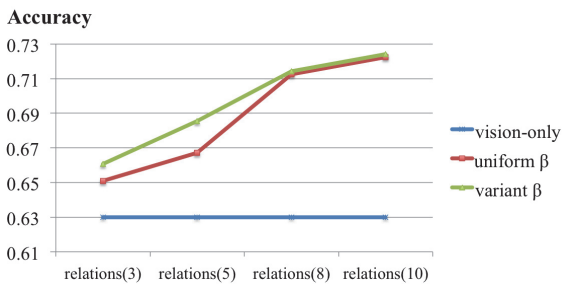

Figure 6: Different β parameter setting: *uniform* is the results using same value for every layers, while *variant* is the results using Equation (4).

6.4 MultiRank Parameters

MultiRank includes two parameters: α and β. Parameter α represents the informativeness of contextual information in the re-ranking process; if the value of α is 1 then the algorithm purely relies on CV whereas the value of $\alpha = 0$ indicates the recognition is only based on contextual information without using vision. The parameter β similarly takes the confidence score of each boxgraph into account as opposed to treating all of the boxgraphs equally using a uniform value. These parameters were tuned empirically.

Figure 5 shows that the accuracy is maximized when the CV output and the contextual information are fused at around $6:4$ ratio when 10 relations are used. Generally, the optimal value for α is decreased as more descriptions are provided, which indicates that perception based on verbal descriptions can be more effective than visual analysis.

In general, a higher confidence value from CV does not necessarily mean that the recognition outcome is more likely to be correct, *i.e.*, an algorithm may generate a wrong output with a high confidence. Parameter β is justified only if a first-pass CV algorithm's confidence score is a good indicator for the actual recognition accuracy. As shown in Table 1, our choice of CV algorithm's confidence score is positively correlated with the accuracy. In Figure 6, we compare the performance using different setting of β value. *"Uniform"* is the results using uniform value for every layers as β value, while *"variant"* is the results using the value defined as Equation (3), which is proportional to the highest confidence score among labels in each layer. This result supports our motivation for parameter β that the use of variant β (over a uniform value) during score updates defined in Equation (3) improves the results.

6.5 Experimental Results

The baseline naïve fusion algorithm: The naïve fusion algorithm is a simple model where the reranked scores for a boxgraph is computed as a weighted sum of its own label confidence score and the confidence scores of other bounding boxes with matching relations. Table 2 shows the results using naïve fusion algorithm. Even when we used 10 relations per image, only minor improvements (2.25% in accuracy and 3.06% in mAP) have been observed.

	Accuracy	mAP
vision-only	0.6299	0.7240
naïve fusion (1)	0.6309	0.7296
naïve fusion (5)	0.6435	0.7442
naïve fusion (10)	0.6527†	0.7546*
spatial relations (1)	0.6331	0.7327†
spatial relations (3)	0.6607*	0.7515*
spatial relations (5)	0.6856*	0.7691*
spatial relations (8)	0.7143*	0.7896*
spatial relations (10)	0.7240*	0.8002*
co-occurrence	0.6331	0.7288
confusion+ co-occurrence	0.6558†	0.7527*

Table 2: Results of vision-only model versus MultiRank using different human-generated information. *Significant t-test:* $*$*=p value* $<= 0.05$, $\dagger$*=p value*$<= 0.10$.

Comparisons of different information sources We report on the performance of the vision-only model and how the performance changes when the same CV algorithm is supplemented with different subsets of our human-generated information. For simplicity, we refer only to the different subsets of information even though we did use the CV results in each test. We first varied the number of spatial relations (1, 3, 5, 8, 10) that were used in our graph. The results are recorded as the average score of 3 random samples of relations. In addition to spatial relations, we also used contextual information such as the CV confusion matrix (to take the weakness of the CV algorithm into account) and the object co-occurrence statistics collected from Shutterstock. We tested different combinations of these.

Table 2 shows the results using naïve fusion and MultiRank with different kinds of contextual information. The vision-only model is the first-pass prediction results of CV recognition model using oracle bounding boxes for object detection. Comparing the vision-only to the varied number of spatial relations, the results indicate that more relations result in more improvement. Using only 1 relation, marginal improvement was observed in mAP, whereas no significant improvement in accuracy. With 3 relations, we started observing statistically significant improvement in accuracy. 9.41% accuracy and 7.62% mAP improvement was achieved using 10 relations.

Whereas the use of verbal descriptions that may require human effort during run-time, the use of the object co-occurrence and the confusion matrix can be achieved without involving people at run-time. Row of *confusion+co-occurrence* in Table 2 displays the results using 1) confusion matrix as within-boxgraph edge weights and 2) co-occurrence matrix as between-boxgraph edge weights; the accuracy was marginally improved by 2.6% in this case which is comparable to using 3 or fewer verbal descriptions per image.

Results based on the detected bounding boxes In Figure 7, the performances of the vision-only system are compared between the oracle and the detected bounding

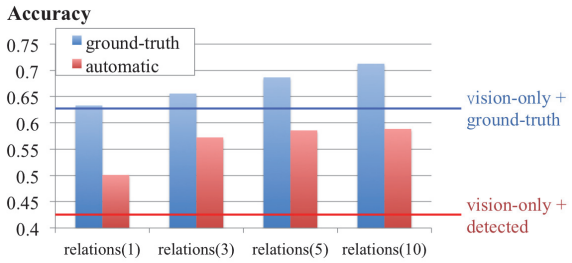

Figure 7: Comparison of MultiRank results on accuracy based on detected and ground-truth bounding boxes.

bounding box	ground-truth		detected	
Accuracy	general	error	general	error
vision-only	0.6299		0.4229	
relation(10)	0.7240	0.7241	0.5885	0.6857*
relation(5)	0.6856	0.6899	0.5857	0.6714*
relation(3)	0.6607	0.6753*	0.5723	0.6600*
relation(1)	0.6331	0.6526*	0.5009	0.5990*

Table 3: Accuracy of MultiRank using descriptions including general versus error-prone objects. *Significant t-test:* $*$=*p value* $<= 0.05$.

boxes cases. The accuracy of vision-only system drops from 62.99% to 42.28% when using the detected bounding boxes; however, the amount of improvement after incorporating descriptions is more substantial. Intuitively, the reason might be that contextual information is more informative when vision is poor.

Results focusing on error-prone bounding boxes Hitherto, we have used general spatial relations in the experiments, that is, the descriptions were selected without considering the user's interest nor the weakness of the vision-only model. We applied the relations to the misclassified objects to simulate possible scenarios where humans are asked to correct the object labels using verbal descriptions. For example, consider a *cup* on the *desk* is misclassified as *pen*. A human may correct this mistake by saying *The cup is above the desk*. By focusing on spatial relations related to the misclassified bounding boxes, we achieved further improvement as shown in Table 3. This result suggests that intelligently choosing the spatial relations be applied when possible. As the number of spatial relations gets lower, the error-specific relations achieve more improvement. The right part of Table 3 shows the results based on detected bounding boxes. Although the accuracy in an absolute scale is still lower than the one using ground-truth bounding boxes, the relative improvement is much higher, *i.e.*, 9.8% compared to 0.71% using 10 relations. This is because the detected bounding boxes are more prone to have errors in them, leaving a larger room for improvement.

Results focusing on objects of user interest In the last set of experiments, we evaluated the performance only based on the objects referred to in humans' commands. The results in Figure 8 provide a supporting evidence for our intuition

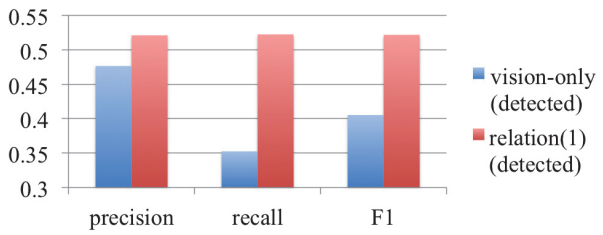

Figure 8: Performance improvement on the objects of interest after giving one relation.

that the major advantage of our approach would apply to the objects of user interest, *i.e.*, those objects that are relevant to a given task. We define the task level error as the recall error combined with the precision error among the recalled. The results show that the task level error is substantially reduced by 12% when we use only 1 relation per image and by more than 20% with 3 relations.

7 Conclusion

In this paper, we propose a graph-based perception algorithm, MultiRank, that can fuse information from computer vision approaches and other perception sources such as human labeled data available online and verbal descriptions provided by people in a shared environment. We apply the idea to the object recognition problem, and demonstrate that our approach of using human-generated data can significantly improve over the computer vision only algorithm. MultiRank has a limitation that it is difficult to generalize to support n-ary relations or aggregate constraints. Our ongoing effort builds on our current model to address this limitation, by converting the graph into a factor graph where a structural factor can be included to represent dependencies over multiple random variables in the graph.

Acknowledgments

This work was conducted in part through collaborative participation in the Robotics Consortium sponsored by the U.S Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement W911NF-10-2-0016. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Aboutalib, S. 2010. *Multiple-Cue Object Recognition for Interactionable Objects*. Ph.D. Dissertation.

Boularias, A.; Duvallet, F.; Oh, J.; and Stentz, A. 2015. Learning to ground spatial relations for outdoor robot navigation. In *IEEE Conference on Robotics and Automation (ICRA)*.

Carreira, J., and Sminchisescu, C. 2012. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(7):1312–1328.

Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2:27:1–27:27.

Chen, Y., and Metze, F. 2013. Multi-layer mutually reinforced random walk with hidden parameters for improved multi-party meeting summarization. In *INTERSPEECH 2013*.

Chen, Y.-N.; Wang, W. Y.; and Rudnicky, A. I. 2015. Jointly modeling inter-slot relations by random walk on knowledge graphs for unsupervised spoken language understanding. In *Proc. the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, CO, USA: ACL.

Choi, M. J.; Lim, J.; Torralba, A.; and Willsky, A. 2010. Exploiting hierarchical context on a large database of object categories. In *Computer Vision and Pattern Recognition (CVPR)*.

Deng, J.; Krause, J.; and Fei-Fei, L. 2013. Fine-grained crowdsourcing for fine-grained recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Divvala, S. K.; Hoiem, D.; Hays, J.; Efros, A. A.; and Hebert, M. 2009. An empirical study of context in object detection. In *CVPR*, 1271–1278. IEEE Computer Society.

Felzenszwalb, P. F.; Girshick, R. B.; McAllester, D.; and Ramanan, D. 2010. Object Detection with Discriminatively Trained Part Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(9):1627–1645.

Hörler, R. 2014. Crowdsourcing in the humanitarian network an analysis of the literature. B.S. thesis, ETH, Zurich.

Hsu, W. H.; Kennedy, L. S.; and Chang, S.-F. 2007. Video search reranking through random walk over document-level context graph. In *Proceedings of the 15th ACM International Conference on Multimedia*, MM '07, 971–980. New York, NY, USA: ACM.

Ipsen, I. C. F., and Kirkland, S. 2005. Convergence analysis of a PageRank updating algorithm by Langville and Meyer. *SIAM Journal on Matrix Analysis and Applications* 27(4):952–967.

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

Kaiser, P.; Lewis, M.; Petrick, R. P. A.; Asfour, T.; and Steedman, M. 2014. Extracting common sense knowledge from text for robot planning. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2014)*, 3749–3756.

Kamar, E.; Hacker, S.; and Horvitz, E. 2012. Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '12.

Kong, C.; Lin, D.; Bansal, M.; Urtasun, R.; and Fidler, S. 2014. What are you talking about? text-to-image coreference. In *CVPR*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc. 1097–1105.

Lawson, W.; Hiatt, L.; and Trafton, J. 2014. Leveraging cognitive context for object recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 387–392.

Lee, H.; Shiang, S.; Yeh, C.; Chen, Y.; Huang, Y.; Kong, S.; and Lee, L. 2014. Spoken knowledge organization by semantic structuring and a prototype course lecture system for personalized learning. *IEEE/ACM Transactions on Audio, Speech & Language Processing* 22(5):881–896.

Lin, D.; Fidler, S.; and Urtasun, R. 2013. Holistic scene understanding for 3d object detection with rgbd cameras. In *The IEEE International Conference on Computer Vision (ICCV)*.

Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.-G.; Lee, S.-W.; Fidler, S.; Urtasun, R.; and Yuille, A. 2014. The role of context for object detection and semantic segmentation in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, 891–898.

Oh, J.; Suppe, A.; Duvallet, F.; Boularias, A.; Vinokurov, J.; Navarro-Serment, L.; Romero, O.; Dean, R.; Lebiere, C.; Hebert, M.; and Stentz, A. 2015. Toward mobile robots reasoning like humans. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Oh, J.; Zhu, M.; Park, S.; Howard, T.; Walter, M.; Barber, D.; Romero, O.; Suppe, A.; Navarro-Serment, L.; Duvallet, F.; Boularias, A.; Vinokurov, J.; Keegan, T.; Dean, R.; Lennon, C.; Bodt, B.; Childers, M.; Shi, J.; Daniilidis, K.; Roy, N.; Lebiere, C.; Hebert, M.; and Stentz, A. 2016. Integrated intelligence for human-robot teams. In *International Symposium on Experimental Robotics (ISER)*.

Oliva, A., and Torralba, A. 2007. The role of context in object recognition. *Trends in Cognitive Sciences* 11(12):520–527.

Page, L.; Brin, S.; Motwani, R.; and Winograd, T. 1999. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab.

Pavlick, E., and Callison-Burch, C. 2015. Extracting structured information via automatic + human computation. In *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2015, November 8-11, 2015, San Diego, California.*, 26–27.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Neural Information Processing Systems (NIPS)*.

Russakovsky, O.; Li, L.-J.; and Fei-Fei, L. 2015. Best of both worlds: human-machine collaboration for object annotation. In *CVPR*.

Salisbury, E.; Stein, S.; and Ramchurn, S. 2015. Crowdar: augmenting live video with a real-time crowd. In *HCOMP 2015: Third AAAI Conference on Human Computation and Crowdsourcing*.

Sarma, A. D.; Jain, A.; Nandi, A.; Parameswaran, A. G.; and Widom, J. 2015. Surpassing humans and computers with JELLYBEAN: crowd-vision-hybrid counting algorithms. In *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2015, November 8-11, 2015, San Diego, California.*, 178–187.

Siddiquie, B., and Gupta, A. 2010. Beyond active noun tagging: Modeling contextual interactions for multi-class active learning. In *CVPR*.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. In *ECCV*.

Thomason, J.; Sinapov, J.; Svetlik, M.; Stone, P.; and Mooney, R. 2016. Learning multi-modal grounded linguistic semantics by playing i, spy. In *International Joint Conference on Artificial Intelligence (IJCAI)*.