A Generalized Model for Multimodal Perception

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

Abstract

In order for autonomous robots and humans to effectively collaborate on a task, robots need to be able to perceive their environments in a way that is accurate and consistent with their human teammates. To develop such cohesive perception, robots further need to be able to digest human teammates' descriptions of an environment to combine those with what they have perceived through computer vision systems. In this context, we develop a graphical model for fusing object recognition results using two different modalities–computer vision and verbal descriptions. In this paper, we specifically focus on three types of verbal descriptions, namely, egocentric positions, relative positions using a landmark, and numeric constraints. We develop a Conditional Random Fields (CRF) based approach to fuse visual and verbal modalities where we model n-ary relations (or descriptions) as factor functions. We hypothesize that human descriptions of an environment will improve robot's recognition if the information can be properly fused. To verify our hypothesis, we apply our model to the object recognition problem and evaluate our approach on NYU Depth V2 dataset and Visual Genome dataset. We report the results on sets of experiments demonstrating the significant advantage of multimodal perception, and discuss potential real world applications of our approach.

Introduction

In order for a human-robot team to effectively perform collaborative tasks in complex environments, it is essential for the team to build accurate and cohesive perception of the environments. Robots in general perceive their environment via on-board sensors, *e.g.*, by using computer vision based approaches on camera images or 3-D point clouds. In a human-robot team setting (Hoffman and Breazeal 2007; Wang, Pynadath, and Hill 2016), in order to develop cohesive team perception, robots also need to be able to digest human teammates' descriptions of the shared environment to combine those with what they have perceived through computer vision systems. In this context, we address the following research question: How can we develop a perception system that can fuse information coming in different modalities such as images from camera sensors and verbal/textual descriptions from human teammates so that the perceived world model is consistent across team members?

To tackle this problem, we develop a graphical model for multimodal perception. Specifically, in this paper, we focus on fusing object recognition results using two different modalities–computer vision and textual descriptions. Based on the computer vision recognition results with errors, as the human make the command or descriptions to describe the environments, we aim to adjust the labels of objects (bounding boxes) based on the embedded relations. In this paper, we specifically focus on three types of descriptions, namely, egocentric positions, relative positions using a landmark, and numeric constraints. A Random Walk based approach on fusing vision with language performs well on understanding descriptions using binary spatial relations (Shiang et al. 2017); however, this approach is difficult to generalize to support more general n-ary relations. To overcome this limitation, we develop a Conditional Random Fields (CRF) (Chavez-Garcia et al. 2013; Sutton and McCallum 2012) based approach where we model n-ary relations as factor functions.

We hypothesize that human descriptions of an environment will improve robot's recognition if the information can be properly fused. To verify our hypothesis, we evaluate our approach on an indoor object recognition problem. Our experimental results show that team perception using our approach significantly improves the recognition performance, by up to 11.44% for the case of objects of interest.

Related work

Although the idea of human-in-the-loop perception system is not new, existing approaches are designed for specific problems or focus on the analysis of human robot interaction. For example, counting the number of objects in complex scenes by asking users to label each object in an image (Sarma et al. 2015), and detecting part of objects by asking humans to click the position of an image according to given questions that have been generated by maximizing the information gain (Wah et al. 2011). Notably, Russakovsky et al. focus on the problem of actively engaging in human computation to improve object detection (Russakovsky, Li, and Fei-Fei 2015). They formulated active annotation problem as a Markov Decision Process (MDP) (Sutton and Barto 1998) to ask an optimal question such as "Is this a chair?" according to a trade-off between the utility and the cost of asking a question. By answering such questions, human

Copyright © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

workers directly perform the labeling role to actively assist computer vision. Conversely, our proposed method does not make any assumption on the strong tie between computer vision systems and humans who provide descriptions, and thus focuses on utilizing information in the descriptions that humans have generated independently of the computer vision results. Our approach can be applied to many practical real-world problems, *e.g.*, the use of social media in disaster response, where data in different modalities can come from unrelated sources. We note that the system architecture doesn't prevent our approach from recruiting a directly interactive approach when crowdsourcing is feasible.

In addition to active learning, multimodal information brings benefits to improve the system using single modality. Kong et al. proposed Markov Random Fields model for coreference of human descriptions for scenes (Kong et al. 2014). Different from the coreference task of pure text, they used multimodal features, such as depth and object positions, to reinforce the results of both textual coreference and visual grounding; for example, for the objects referring to the textual terms which are regarded as in the same coreference cluster, they are more likely to be the same one, and vice versa. Another work that is relevant to ours is (Thomason et al. 2016) where users describe objects using attributes, *e.g.*, distance, when playing I-SPY game. Because their approach was designed for the game specifically, a direct comparison was not possible. Their experimental results also compare the multimodal approach against their vision-only system; F1-measures are 0.196 and 0.354 for vision-only and multimodal approaches, respectively.

For fusing information in visual and textual modalities, Chavez-Garcia et al. (Chavez-Garcia et al. 2013) proposed Markov Random Fields to rerank the image retrieval results according to pseudo-relevance feedback, which contains both image and language. A Random Walk based approach known as MultiRank was proposed in (Shiang et al. 2017) where they used two types of contextual information generated by humans: Spatial relations specified in a human commander's commands and object co-occurence statistics mined from an online image data source. This approach has proven effective in improving object recognition results; however, because a spatial relation is encoded in the edges on their graph representation, the types of spatial relations that can be supported are limited to binary relations. Since this approach shares a high-level framework as our approach, we use this work as our baseline.

Graphical Model for Fusion

Problem definition

The target perception problem in this paper is object recognition. An object recognition problem in general can be defined as: Given an image input, segment the image into a set of regions and assign a class label for each region. Multimodal object recognition relaxes the input type to support additional modalities. For instance, we consider text data that human teammates and/or crowdsourcing can provide to describe the same scene. As illustrated in Figure 1, the data need to be interpreted and understood using specific techniques for each modality, *e.g.*, an image data is analyzed using computer vision techniques whereas text data is processed using natural language processing and language understanding techniques. Our target problem here is to fuse the information from both sources in a principled way to generate the recognition output that is cohesive across the modalities.

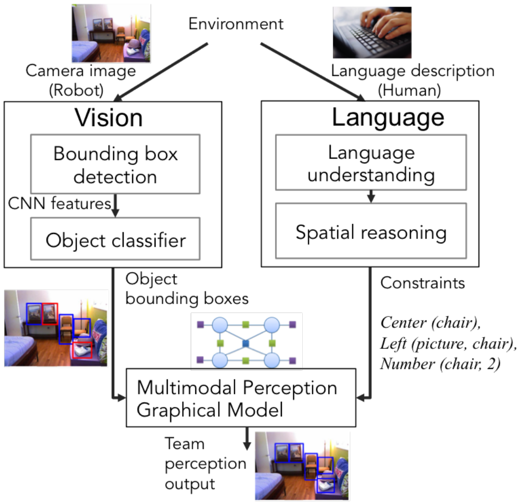

Figure 1: An architecture for multimodal team perception.

In this problem, since the final output format from the recognition system is the same as that of the computer vision systems, we formulate this problem as object recognition conditioned on additional information. Formally, given set R of regions (such as bounding boxes) and set C of object class labels, let $\phi : R \rightarrow C$ denote an assignment of all regions to class labels, Φ be a set of all possible assignments, O denote a set of random variables each of which representing the probability of the assignment being accurate, and $M = m_1, ...m_n$ be a set of input data representing n modalities. Then, the object recognition problem is finding the assignment that maximizes this distribution as follows:

$$\arg\max_{\phi\in\Phi} p(O|\mathrm{M}, \phi).$$

Background: Conditional Random Fields Conditional Random Fields is a probabilistic framework for structured predictions (Lafferty, McCallum, and Pereira 2001). Let G denote a factor graph over input and output (or label) variables X and Y. Probability distribution (X, Y) is conditional random fields if the conditional distribution $p(y|x)$ factorizes for every observed value x in X for Graph G, *i.e.*, the distribution can be written as a product of factors (or local functions) (Sutton and McCallum 2010). In general, the local functions in a CRF are defined as a weighted sum of feature functions as follows:

$$P(Y|X) = \frac{1}{Z(x)} exp(\sum_{k=1}^{K} \theta_k f_k(x, y)) \tag{1}$$

where $f_k(x, y)$ and $\theta_k(x)$ denote the k^{th} feature function and its weight, respectively, and $Z(x)$ is the normalization

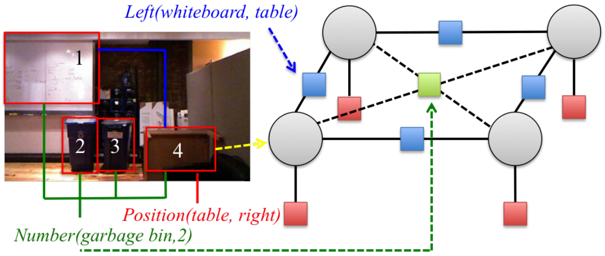

Figure 2: Our CRF model for fusing vision, egocentric (e) and relative (r) relations and numeric (n) constraints.

function defined as follows:

$$Z(x) = \sum_{y \in Y} exp(\sum_{k=1}^{K} \theta_k f_k(x, y)) \tag{2}$$

Since a node in a CRF only depends on its neighbors in its Markov Blanket (Pearl 1988), the conditional distribution can be efficiently approximated using Gibbs Sampling (Liu 1994). The prediction algorithm selects the most probable label based on the feature functions as follows:

$$y = argmax(exp(\sum_{k=1}^{K} \theta_k f_k(x, y))) \tag{3}$$

Algorithm: CRFs for information fusion To model multimodal perception as a CRF, we construct a factor graph over the input variables X, the bounding boxes from a computer vision based object recognition, and the output variables Y, a set of possible labels. We then define 4 types of factor functions: One to represent a visual modality and three to encode relation descriptions in a textual modality. Specifically, we focus on 3 kinds of relations that might be embedded in language inputs: Egocentric position, relative position, and numerical constraints, denoted by rel_e, rel_r, and rel_n, respectively. **??** shows the CRF model for multimodal team perception, illustrating different types of factors for fusing multiple modalities.

Spatial reasoning: Before we introduce the factor functions, we first define preliminaries. For each bounding box, we have the coordinates for bounding box (w_1, h_1, w_2, h_2) describing top_left and bottom_right points. The horizontal center of the bounding box is $(w_1 + w_2)/2$. Each description function returns true if its statement is satisfied, false otherwise. In this paper, we use the following predefined descriptions. For egocentric relations, we currently support three location descriptions, denoted by $\Omega = \{\text{Left, Right, Middle}\}$ where $\omega(x)$ in Ω takes a bounding box x as an input argument.

1. *Left*: The center of box x lies in the region between $(0, 0.25 \cdot |W|)$.
2. *Middle*: The center of box x lies in the region between $(0.25 \cdot |W|, 0.75 \cdot |W|)$.
3. *Right*: The center of box x lies in the region between $(0.75 \cdot |W|, |W|)$.

Definition 1 (Match (x, ω)**)** *Given bounding box x and relation function ω in Ω, if the evaluation of function $\omega(x)$ holds true then box x is matched with relation defined by ω.*

For the relative spatial relations, we currently define 5 types of relations $\Xi = \{\text{Left, Right, Above, Below, On}\}$ where $\xi(x, x')$ in Ξ takes two bounding box inputs $x = (w_{11}, h_{11}, w_{12}, h_{12})$ and $x' = (w_{21}, h_{21}, w_{22}, h_{22})$.

1. *Left*: The left-most point bounding box 1 (w_{11}) is smaller than the left-most point of bounding box 2 (w_{21}), the right-most point bounding box 1 (w_{12}) is smaller than the right-most point of bounding box 2 (w_{22}), and the vertical difference between two bounding boxes is smaller than $max(h_{12} - h_{11}, h_{22} - h_{21})$.
2. *Right*: Similar to relation *Left* except that the two bounding boxes are reversed in the horizontal axis.
3. *Above*: The top-most point bounding box 1 (h_{11}) is smaller than the top-most point of bounding box 2 (h_{21}), the bottom-most point bounding box 1 (h_{12}) is smaller than the bottom-most point of bounding box 2 (h_{22}), and the horizontal difference between two bounding boxes is smaller than $max(w_{12} - w_{11}, w_{22} - w_{21})$.
4. *Below*: Similar to relation *Above* except that the two bounding boxes are reversed in the vertical axis.
5. *On(Attach)*: Bounding box x is contained by the landmark bounding box x'.

Definition 2 (Match (x, x', ξ)**)** *Given bounding boxes x and x', and relation function ξ in Ξ, if the evaluation of function $\xi(x, x')$ holds true then box x and landmark x' are matched with relation defined by ξ.*

Factor functions: Let $M = \{m_{cv}, m_e, m_r, m_n\}$ represent the input modalities for vision, and the three textual modalities (we divide the textual modality into subgroups for easier reading). We define the four factors as follows:

1. Computer vision: For each bounding box x, we get a confidence score $cv(x, y)$ for each label candidate y from a computer vision based recognizer, which constitutes the computer vision feature function f_{CV} as:
$$f_{CV}(x, y) = cv(x, y), \forall m \in m_{cv}. \tag{4}$$
2. Egocentric positions: This feature encodes a type of description from the camera point of view, *e.g.*, "The cat is in the middle (of the image)." For each bounding box x, if an egocentric position relation $rel_e(y, \omega)$ matches with the spatial location of box x, the feature function is defined as:
$$f_E(x, y) = \begin{cases} cv(x, y), \ if \ match(x, \omega) \ \wedge \ rel_e(y, \omega) \in m_e \\ 0, \ otherwise \end{cases} \tag{5}$$
3. Relative positions: This feature supports descriptions using a set Ξ of relative spatial relations, *e.g.*, *left, right* or *above*. A relative position of an object is specified in relation to a landmark object, *e.g.*, "The cat is on the right of the chair," where *chair* is used as a landmark to specify the location of the *cat*. This sentence can be expressed as: rel_r('cat', 'chair', 'right'). For every pair of bounding boxes, if there is a matching spatial relations, $\xi \in \Xi$, and the object labels referred to in the description have positive probability mass, then we create a factor function

linking the variables representing the two boxes. Then, the feature function for a relative position for variable x and landmark variable x' can be written as:

$$f_R(x, x', y, y') = \begin{cases} cv(x, y)cv(x', y'),\ if\ match(x, x', \xi) \\ \wedge\ rel_r(y, y', \xi) \in m_r \\ 0,\ otherwise \end{cases} \tag{6}$$

4. Numeric constraints: A numeric constraint can specify how many instances of an object class are in an environment, *e.g.*, "There are 2 cats." This example can be expressed as $rel_n(X, \text{'cat'}) = 2$. Whereas other types of factors are localized in a subgraph that is relevant to a specific relation, a numeric constraint imposes a global constraint over all input variables.

$$f_N(X, y) = \begin{cases} -\left|Num(X, y) - rel_N(X, y)\right| \\ \quad if\ rel_N(X, y) \in m_n \\ 0,\ otherwise \end{cases} \tag{7}$$

Prediction fusion: The objective of solving the CRF for making predictions is to maximize the following distribution function:

$$P(Y|X) = \frac{1}{Z(x)} exp(\theta_{CV} \sum_{y \in Y} f_{CV}(y) + \sum_{y \in Y} \theta_E f_E(y) + \\ \sum_{x, x' \in X, y, y' \in Y} \theta_R f_R(x, x', y, y') + \sum_{y \in Y} \theta_N f_N(X, y)) \tag{8}$$

where θ_{CV}, $\theta_E, \theta_R, \theta_N$ are the weights for each feature function. Because there are only four weight parameters in our CRF model, we tune the parameters from a validation set as in (Shiang et al. 2017). In general, if the number of parameters is large, the weights of a CRF can be obtained by using quasi-Newton methods (Schraudolph, Yu, and Günter 2007). We use Gibbs Sampling to update each random variable to get an approximate solution. For each bounding box (random variable), we fix the labels of all other bounding boxes to calculate the probability using the feature functions; we then select the label with the highest probability value as the new label for that random variable. This process can be written formally as below and we iteratively update it until convergence:

$$y^* = \underset{y}{argmax}(\theta_{CV} \sum_{y \in Y} f_{CV}(y) + \sum_{y \in Y} \theta_E f_E(y) + \\ \sum_{x, x' \in X, y, y' \in Y} \theta_R f_R(x, x', y, y') + \sum_{y \in Y} \theta_N f_N(X, y)) \tag{9}$$

Experiments

We replicated the experimental setup of our baseline approach, MultiRank, as described in (Shiang et al. 2017). To validate our method, we verify it on NYU Depth v2 dataset (Silberman et al. 2012) and Visual Genome dataset (Krishna et al. 2016). We describe the setup here.

Data preparation: In NYU Depth v2 dataset, there are 1449 indoor scenes that includes over 800 kinds of objects in cluttered environments. We collected textual descriptions for 40 randomly selected images from the set (35 for testing, 5 for validation). Because natural language processing is not our main focus in this work, we collected the descriptions in a structured language that can be parsed readily. We also used another set of relations from Sentence3D dataset (Kong et al. 2014), where there are descriptions and extracted relations for each image. We filtered out all the relations with subjects or objects not in our label set; therefore there are overall 59 relations and 8 images without any relations. In the Visual Genome dataset, we used those images that include the following set of spatial relations: {`left, right, up, down, below, above, on, attached, near, next, around`} and that include at least 5 objects of the 20 object labels from the Pascal VOC 2012 set (Everingham et al.). We finally used 202 images and 1541 relations.

Evaluation metric: As in the baseline, we used accuracy as the main evaluation metric.

Vision-only algorithm: For the computer vision based object recognizer, we also adopted the same algorithms used in the baseline to reproduce the same inputs for the fusion system. A set of bounding boxes was detected using Constrained Parametric Min-Cuts (CPMC) (Carreira and Sminchisescu 2012). We then classified the bounding boxes using a Support Vector Machine (SVM) with the fc7 features from the fully connected layer of AlexNet pretrained on ILSVRC 2012 (Krizhevsky, Sutskever, and Hinton 2012). This *vision-only with detected bounding box* algorithm achieved accuracy of 0.4229 on the test data set. We also report the results using the ground truth bounding boxes; *vision-only with ground-truth bounding boxes* achieved accuracy of 0.6299.

The naïve fusion algorithm: The naïve fusion algorithm is a simple model where a new score for node x having label y is computed as a product of the confidence score of the bounding box and the sum of multiplications of the confidence score of other bounding boxes with match relations:

$$\text{score}(x, y) = cv(x, y) \cdot \sum_{x' \in X, x' \neq X, \xi \in \Xi} cv(x', y') \cdot match(x, x', \xi)$$

where $match(x, x', \xi)$ is a (0 or 1) binary function specifying whether the two boxes satisfy spatial relationship ξ.

Experimental Results

Figure 3 shows four examples of comparing the vision only and the multimodal perception results; the first row lists textual descriptions, the second row, the vision only results, and the third row, the multimodal results (corrections are highlighted in yellow). For analytic results, we report on sets of experiments that evaluate the performance of our CRF approach against the vision-only algorithm, the naïve approach, and the MultiRank algorithm.

Results on relative spatial relations Table 1 show the results using the naïve, MultiRank and CRFs when textual inputs include only binary spatial descriptions, that is, an input describes an object using its relative position with respect to a landmark. An example of this type is "A cat is below the table." In this set, the vision-only object recognizer achieved

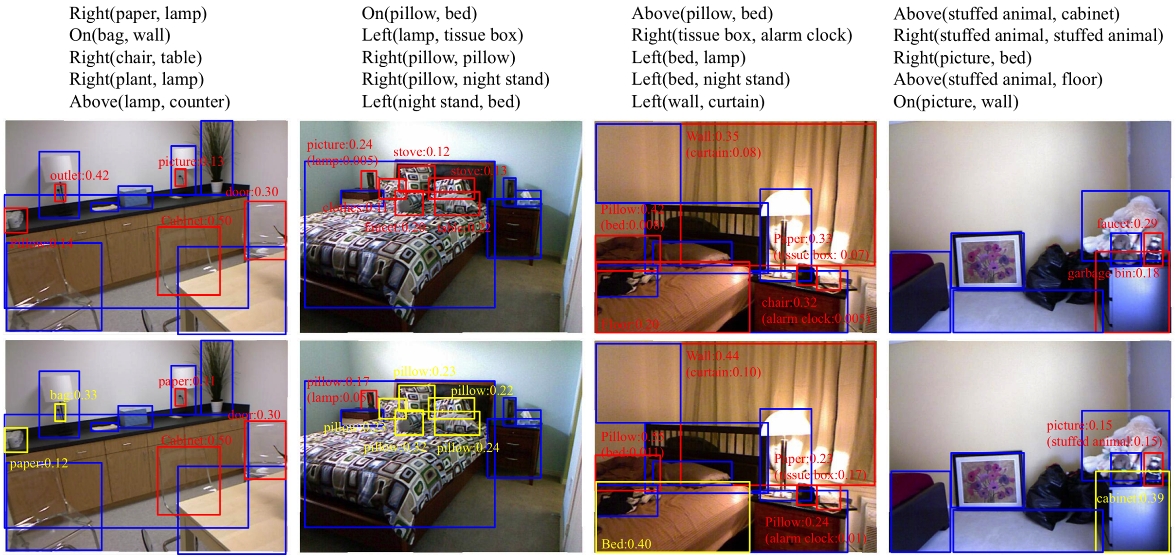

Figure 3: The examples showing the recognition performance of the vision-only (upper) and our multimodal (lower) approaches. The blue boxes are the correctly classified ones; the red boxes, misclassified ones; and the yellow boxes, those boxes that have been corrected after multimodal fusion. The numbers shown are the confidence score.

Table 1: Results of vision-only recognizer versus MultiRank and CRF using relative spatial relations with ground-truth bounding boxes. *significant t-test:* $*$=*p value* $<= 0.05$, †=*p value*$<= 0.10$.

	Naive	MultiRank	Proposed
NYU Depth v2 dataset			
vision-only	0.6299		
relative(1)	0.6303	0.6311	0.6330
relative(3)	0.6417	0.6607	0.6675
relative(5)	0.6518	0.6856	0.6789
relative(8)	0.6583	0.7142	0.7052
relative(10)	0.6688	0.7240	0.7114
relative(*)	0.6344	0.6347	0.6359
Visual Genome			
vision-only	0.4359		
relative(*)	0.4632	0.4814	0.4836

62.99% in accuracy for the ground truth bounding boxes in NYU depth dataset. Because the environments are cluttered in the images used in the experiments, the naïve model failed to effectively utilize the relations and it only achieved 3.89% improvement given 10 relations for the ground-truth bounding boxes. By contrast, both graphical models, MultiRank and CRF approaches, achieved significant improvements over the (vision-only) inputs using 3, 5, 8, 10 relations per image. MultiRank performed slightly better than CRF model, but the results were compatible. The difference here might be due to the optimization implementation of MultiRank and CRFs. Recall that, in CRFs, we use Gibbs Sampling to deal with the intractable structure and thus the results are approximate.

We also verify our results on relations generated from Sentence3D relations, which is shown in *relative(*)* row. In sentence3D dataset, there are 1-2 relations per image in average; therefore the improvement is subtle. However, it's not to say that there are not abundant information in human descriptions or our algorithm performs not good, it's because in NYU depth dataset we filtered 74 object labels from 891 labels, thus some of the relations are not subjected to the objects that we choose. The original number of relations is 135 in 35 images, which is 4 relations per image in average.

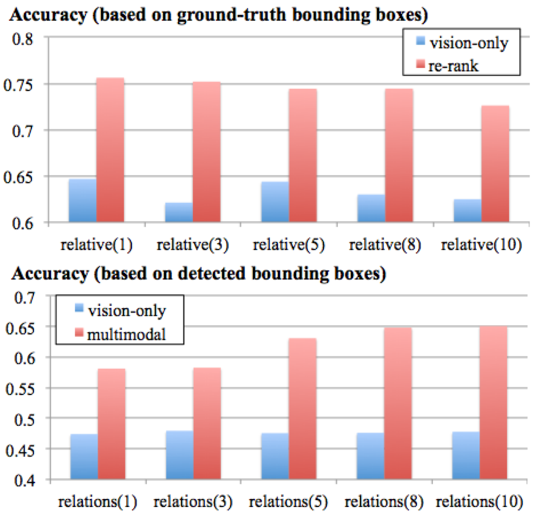

Figure 4: CRF multimodal perception performance on objects of user interest. Only the objects mentioned in the relation set are counted for evaluation.

The lower part of Table 1 is the results for Visual Genome dataset. The vision-based classifier achieve 43.59% in accuracy and the our proposed method achieves 48.36% in ac-

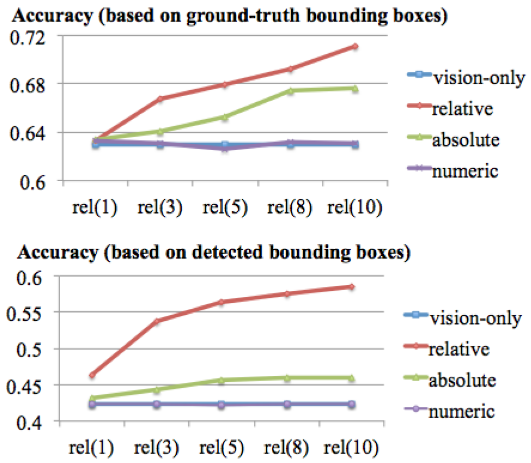

Figure 5: Results using CRF with different information.

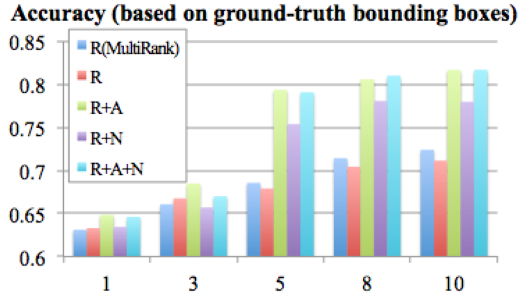

Figure 6: Results with different information combination.

curacy, which slightly performs better than MultiRank. Recall that the number of images in our test set is 202 and the overall number of relations are 1541, which means that there are about 7 relations per image. The relative improvement is about 4.77%, which is consistent with our experiments in NYU Depth v2 dataset.

CRF performance on objects of user interest In this set of experiments, we checked the performance only based on those objects that are directly referred to in the textual inputs. As illustrated in Figure 4, we gained more substantial improvements in this case when compared to the previous set in **??** . In the ground-truth bounding boxes case, we achieved 10.94%, 10.04% and 10.11% gain of accuracy using 1, 5 and 10 relations. In the detected bounding boxes case, because we began with a significantly lower vision-only inputs, the accuracy in an absolute scale was lower; however, we observed even more substantial improvement, which is 10.70%, 15.51% and 17.28% using 1, 5 and 10 relations. Based on the result, we expect multimodal perception can improve the performance at the task level, *e.g.*, robots following human directions can benefit from our approach by being able to interpret user commands, such as ”Pick up the cup on the table,” with an increased accuracy.

CRF performance on information types In this experiment, we evaluated how each type of information impacts the performance of the CRF approach. Figure 5 shows the accuracy using 4 different types of information on the ground-truth and the detected bounding boxes cases. The result using relative spatial relations performed the best, achieving 8.15% improvements using 10 relations, respectively. The result using egocentric position relations performed fairly well but with a smaller amount of improvement, 4.63% improvements using 10 relations, respectively. A reason for the weaker performance in the case of egocentric description might be because the number of objects included in each description is only one. The improvement using numeric constraints was not significant. This observation indicates that the feature function may not fully capture numeric constraints to be effectively updated in the CRF where the variables are updated in a random order as opposed to being updated in an informative order where those variables that are relevant to given descriptions can be given a higher priority.

In the detected bounding box case (the second figure in Figure 5), the improvement using relative spatial relations was higher than using other types of information, and the improvement of using numeric constraints was close to 0 (Note that the two data lines of ‘numeric’ and ‘vision-only’ almost overlapped.) Unfortunately, the numeric constraint is hard to make improvement under the current bounding box scenario. For instance, because there are a lot of errors in bounding box detection, several objects can be grouped together in one bounding box. In such a case, even when all of the detected bounding boxes are accurately classified, the number of objects can still be incorrect.

Figure 6 shows how the performance changes when different types of information are combined. Letters R, E and N represent relative, egocentric and numeric constraints, respectively. In this figure, the x-axis shows the number of descriptions in each category, *e.g.*, value of 3 for the R+E+N means that 3 relative relations, 3 egocentric relations, and 3 numeric constraints were used. When multiple types of information was used together, there was further improvement in the final outcome. This shows the ability of the proposed model to achieve further improvement, compared to MultiRank which can only utilize binary relations, given various kinds of information.

Conclusion

In this paper, we present a generalized model for multimodal team perception using the CRF framework. We define feature functions to encode multimodal inputs, namely the recognition results from a computer vision system and three different types of textual descriptions about shared environments. In sets of experiments, we demonstrate that our multimodal approach significantly improves the perception performance from the vision only, single-modal system. We also show that our CRF model can generalize to support structured n-ary relations that the existing baseline model is not able to represent. The results are promising to indicate that our multimodal team perception approach can be applied to real world problems such as multimodal semantic map construction in disaster response or other general human robot systems.

References

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.

Chavez-Garcia, R. O.; Escalante, H. J.; Montes, M.; and Sucar, L. E. 2013. Multimodal Markov Random Field for Image Re-ranking based on Relevance Feedback. *ISRN Machine Vision* 2013(2013):16.

Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

Hoffman, G., and Breazeal, C. 2007. Effects of anticipatory action on human-robot teamwork efficiency, fluency, and perception of team. In *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction*, HRI '07, 1–8. New York, NY, USA: ACM.

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

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; Bernstein, M.; and Fei-Fei, L. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

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.

Lafferty, J. D.; McCallum, A.; and Pereira, F. C. N. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, 282–289. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Liu, J. 1994. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problems. *Journal of the American Statistical Association* 89(427):958–966.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

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

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.

Schraudolph, N. N.; Yu, J.; and Günter, S. 2007. A stochastic quasi-Newton method for online convex optimization. In *Proc. 11th Intl. Conf. Artificial Intelligence and Statistics (AIstats)*. San Juan, Puerto Rico: Society for Artificial Intelligence and Statistics.

Shiang, S.; Rosenthal, S.; Gershman, A.; Carbonell, J.; and Oh, J. 2017. Vision-language fusion for object recognition. In *AAAI Conference on Artificial Intelligence (AAAI)*.

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

Sutton, R. S., and Barto, A. G. 1998. *Introduction to Reinforcement Learning*. Cambridge, MA, USA: MIT Press, 1st edition.

Sutton, C., and McCallum, A. 2010. An introduction to conditional random fields. *arXiv preprint arXiv:1011.4088*.

Sutton, C., and McCallum, A. 2012. An introduction to conditional random fields. *Foundations and Trends in Machine Learning* 4(4):267373.

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)*.

Wah, C.; Branson, S.; Perona, P.; and Belongie, S. 2011. Multiclass recognition and part localization with humans in the loop. In *IEEE International Conference on Computer Vision (ICCV)*.

Wang, N.; Pynadath, D. V.; and Hill, S. G. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, HRI '16, 109–116. Piscataway, NJ, USA: IEEE Press.