A Robust Algorithm: Find an Unknown Person via Referring Grounding

Xiping Wang, Feng Wu*, Dongcai Lu, and Xiaoping Chen

Multi-Agent Systems Lab, School of Computer Science and Technology,
University of Science and Technology of China, Hefei, 230027, Anhui, China
wxiping@mail.ustc.edu.cn, wufeng02@ustc.edu.cn, ludc@mail.ustc.edu.cn,
xpchen@ustc.edu.cn

Abstract. We propose a simple but robust method to recognize an unknown person described in natural language. In this case, a robot is given a verbal description about a person whom the robot is required to recognize. This task is challenging since humans and robots have significantly mismatched perceptual capabilities (e.g., recognizing the color of a coat). Without assuming that all linguistic descriptions and perceptual data are correct, we use a probabilistic model to ground the target person. In particular, the acceptability of color descriptions is modeled based on visual similarity and the confusion matrix of the color classifier which make the system more robust to illumination. Two groups of experiments were conducted. Our experimental results demonstrate that our system is robust to both perception and description errors.

1 Introduction

Human-robot collaboration is still a challenging task, especially for service robots. In RoboCup@Home competition, there are many tasks related to humans, e.g., giving a drink to a person, following a man, guiding a woman, etc. In many cases, it is highly likely that the target person in the instructions is described ambiguously. Usually in the completion, a robot is required to find an unknown person through recognizing some special motions of the person, such as waving hand. A very common situation is that there are several people who may be the target person, e.g., when customs coming to a restaurant or guests visiting your house. Face recognition cannot be used in this scene because it needs to remember the person's face in advance. Robots need some simple but effective and intuitional features to identify the person.

Natural language is a flexible modality for human-robot interaction [3, 13]. It is a natural way to identify a person by giving some visual feature descriptions through natural language. A referring expression is a linguistic product used to discriminate a specific object from the rest of the world. The robot needs to identify referents in the environment that are specified by its human partner. The ability to ground referring expressions is important for conversational agents

* Corresponding Author.

aimed at real-world interaction. With the ability of grounding referring expressions to objects in the environment, robots can accomplish more complex tasks through human-robot collaboration. Kunze et al. [10] proposed an approach for searching for objects on the basis of Qualitative Spatial Relations (QSRs). Huo et al. [9] use natural spatial language to control a robot performing a fetch task in an indoor environment. Similarly, this paper aims to help the robot to recognize an unknown person via referring grounding. Two kinds of cues are used. They are QSRs and color.

Imagine such a scenario. Bill invited some friends come to his house. Lucy is one of the guests. Bill is busy in preparing for dinner and he may tell the robot: "bring ice tea to Lucy from the fridge. She sits beside the cupboard in the drawing-room and wears a red coat". The example sentences above contain two kinds of features about Lucy, color and spatial relations. Grounding the meaning of the descriptive words in natural language by mapping them to its perception will enable the robot to identify the specific physical person referred to. Actually, the robot may encounter some uncertainties. First, humans and robots have mismatched capabilities of perceiving the shared environment [12]. That is to say, the perceived results of the robot may be different from that of the human being. The robot is required to to give an assessment of the acceptability of the information given by people based on its perception. There are two situations that may occur when evaluating the acceptability of a speaker's description about color. At first, there is ambiguity in naming colors even for people. One thinks a coat is orange, others may think it is yellow. Secondly, due to many scene accidental events such as unknown illuminant, presence of shadows and camera configuration, such as exposure, white balance, etc. people and color classifier often give different predications as to an object. For example, when illuminant becomes dark, the color classifier tends to judge the blue object as grey. However, blue and gray are not as visually similar as orange and yellow. Secondly, different from still objects, humans are likely to move which result in the previous spatial relation description becoming wrong. Furthermore, due to the limitations of perceptual algorithms, sensing results are not totally reliable. In the cases, the robot must select correct descriptions based on some strategy.

To tackle those problems above, we propose a set of computational mechanisms that correspond to the most commonly used descriptive strategies to evaluate the compatibilities between the referring expressions and numeric attribute values from robot's perception. The acceptability of color expression is evaluated both from visual similarity and the confusion matrix of the color classifier. Based on these mechanisms, we use a probabilistic model to determine the criteria for selecting the correct combination from all the descriptions to uniquely identify the referent. We tested the system under various conditions. Experiments show that even in the case that color classification is incorrect or the target person is mis-detected, our system can give the correct or approximate correct grounding results. Two groups of experiments are designed under different illuminations. The results indicate that the system is robust both to the perception errors and description errors. Note that our algorithm can be incorporated with other methods for more complex robot decision-making and planning [1, 21]. In the following sections, we first give a brief discussion of the related work and then give an overview of our system and describe our probabilistic model and word grounding module. Then we designed two groups of experiments that correspond to several typical situations that may occur in reality to evaluate the robustness of our method. Finally, we conclude with our contribution and future work.

2 Related Work

There are two main problems during situated dialog. The speaker will encounter the problem of Referring Expression Generation (REG) when he intends to describe one object. Given Referring Expressions (REs), the listener needs to ground these REs to figure out the referent. The problem in this paper belongs to the latter. Computational approaches for referring grounding often consist of two key components [7].

The first component which can be called word grounding models [11] addresses formalisms that connect linguistic terms to the numerical features captured in the robot's representation of the perceived environment. In this paper, word grounding modules give assessment of the acceptability of QSRs and color expression. There are many lectures that focus on searching task using QSRs by relating an unknown object to a known object which can be called landmark [9, 10]. The landmark in [12, 11] also can be unknown objects. This makes the problem more complex. The robot need to grounding multi referents. However, for a service robot, it has already constructed a 2D map of the house in advance. The position of the stationary furniture (i.e., bookshelf, cupboards) which does not tend to move in everyday usage is known for the robot. It is easy and effective to select these furniture as landmark.

Different from QSRs, it is more complicated to describe color for machines due to scene accidental events. Therefore it is hard to evaluate the acceptability of color description.In [17], a perceptually based color-naming metric was designed to measure the compatibility between a color name and an arbitrary numeric color values in different color spaces (RGB and HSV). Based on this metric, [15] proposed to evaluate the acceptability of color description based on the deviation from its prototype in HSL space. This method can handle the visual similarity between similar colors such as yellow and orange. But it ignores the effects of scene accidental events such as illuminant. Although none of the color classifiers can solve the illuminant problem, the trend of error can be predicted. We can test the classifier on a large data set and obtain its confusion matrix. If 20% ground truth with a blue label are mistaken for gray, when the speaker described a coat as blue but the classifier predicted it as grey, the acceptability of blue should be enlarged even though blue and gray is not similar in color space. So we evaluate the acceptability of color expression based on the similarity in color space and the confusion matrix of the color classifier.

The second component extracts all the linguistic terms from referring expressions and combines their grounding models together to identify referents.

Gorniak and Roy [7] address the interpretation of combinatory spatial REs with incremental filters, filter out a set of potential referents of each property using perceptual data. This is computationally efficient. However, these methods are based on the assumption that all these referring expressions and perceptual data are correct. When grounding a person, the assumption is not established due to the uncertainties discussed previously. Mast et al. [16] propose a probabilistic framework base on the discriminatory power (DP) to measure how likely expressions as a whole are to distinguish the target object from the distracter object for REG. In this paper, we choose this as a criterion to select correct expressions from all the descriptions then ground the target person. At the same time, the grounding results can be adjusted with the feedback of the human partner.

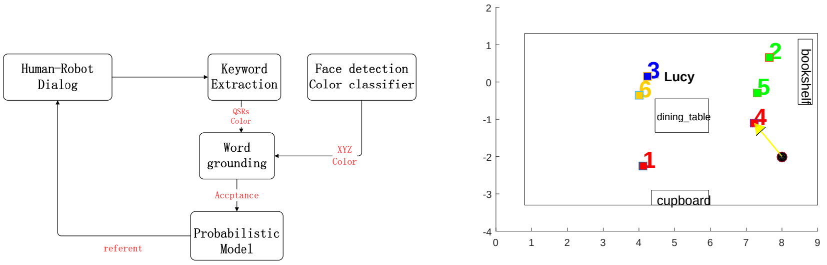
3 A Probabilistic Framework for Referring Grounding

3.1 System Overview

The setup of our experimental system is shown in Fig. 1(b). It is a simplified map of the dining-room in our lab. The robot is represented by a black dot. People in the room are represented by squares with respective colors. The robot is ordered to find Lucy whom it has never seen. It can ask people in the scene about the characteristics of Lucy until it confirms who is Lucy. The overall architecture is shown in Fig. 1(a). Through human-robot dialog, two kinds of cues, spatial relations and colors are supplied to a robot. As all the spatial relation (e.g., nearby, far from), landmarks (e.g. dining-table, TV) and color (e.g., red, grey) are known both for the robot and the speaker, we can find relationship prepositions and corresponding accusatives and color nouns in sentences by keyword matching. The robot perceives the surrounding environment and obtains the coordinates of people and their clothes colors. All the information is forwarded to word grounding module which evaluates the compatibilities between the word expressions from the dialog and the numeric attribute values or labels from perception. The probabilistic model will select the expressions that best match the robot's perception and figure out the referent.

3.2 Perception

Microsoft Kinect is used to obtain aligned RGB-D images. 2D laser scanners is used for localization and navigation. First a 2D occupancy grid map[5] [8] is generated from the laser scanning results and odometry data. We annotate the different structures such as rooms, doors, furniture and other interested objects according to their semantic information in the the grid map. The map constructed the world coordinate system of the home environment. The depth image is transformed into the world coordinate frame via tf [6]. We use HAAR [20] face detector to detect human face area in the RGB image. The corresponding area in the depth image is segmented to obtain human's location in world coordinate. It can be inferred that the area below the face area with the distance of equal



(a) (b)

Fig. 1. (a) The overall architecture. (b) An example scene.

height of the face corresponds to the human clothes. In order to reduce the influence of illumination, we use grey world [2] to correct color cast to a certain degree. A fuzzy color model which is learned from uncalibrated data obtained from "Google Image" in [19] is used as the color classifier.

3.3 Probabilistic Model

Suppose during the conversation, the speaker (No.4 in Fig 1(b)) uses three sentences to describe Lucy to the robot. "Lucy stands beside the dining-table in a blue coat.","She is far from the cupboard.","She is beside the bookshelf." It is obvious that the first sentence and the last sentence are contradictory. The robot has to find one strategy to figure out which one is credible. The characteristics described in the first two sentences clearly distinguish No.3 from others. The characteristic described in the last sentence is unable to distinguish between No.2 and No.5. As a listener, the robot is more convinced that the No. 3 who is beside the dining-table in a blue coat is Lucy rather than the ones beside the bookshelf, as the discriminatory power (DP) of this description combination is greater. DP can be modeled by a probabilistic model $P(x|D)$ which means given a description D and a person x, the probability that the listener identifies the goal person x correctly. $P(x|D)$ can be formulated in the Bayesian framework:

$$P(x|D) = \frac{P(D|x)P(x)}{P(D)} \tag{1}$$

where $P(D|x)$ is the probability that given the target person x, such a description D would be accepted by humans. The description related to the referent are sets of feature descriptors f_i(color or QSR). $D := \{f_1, f_2, \ldots, f_n\}$. For each person and feature the word grounding module gives a number within $[0, 1]$ that measures the respective feature appropriateness. Obviously, the probability of accepting that "Lucy wears a red coat" is independent of the probability of accepting "She is beside the dining-table". Therefore, it is reasonable to assume that the acceptability of different features is stochastically independent.

$$P(D|x) = P(f_1|x) \cdot P(f_2|x) \cdot \ldots \cdot P(f_n|x) \tag{2}$$

$P(D)$ gives the probability that the description D suits an arbitrarily chosen person in the scene. Suppose there are M individuals in the scene. $P(X_i)$ is the probability of randomly choosing the person X_i in the scene. For simplicity, $P(X_i) = 1/M$. If multiple people in the scene suit the description D, the probability of correctly identifying the referent will reduce. Therefore, the following formula can be obtained.

$$P(D) = 1/M \sum_{i=1}^{M} P(D|X_i) \tag{3}$$

If the target person x exists, that is $x \in X$ (he has been detected by the robot), DP of the combination of the descriptions associated with him is certainly maximal. Based on these, the strategy for the robot to identify the credible descriptions can be obtained.

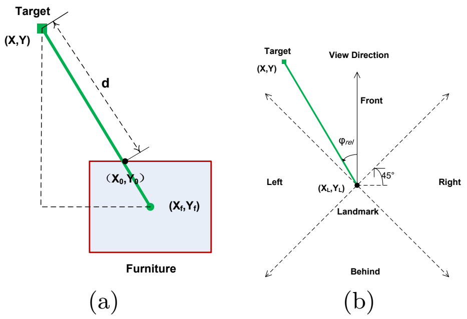
$$(D^*, x^*) = arg\ maxP(X_i|D) \tag{4}$$

D^* and x^* stand for the correct description combination for the referent and the person most likely to be the referent respectively. However, it is impossible to ensure that the robot can detect all the people in the scene. If No.3 in Fig. 1(b) has not been detected by the robot, the optimal solution will be meaningless. Under this circumstance, the acceptability of the correct description will be very low. Worse still, in the presence of incorrect descriptions, it is difficult for the robot to be aware of its mis-detection. Suppose the robot infers that No.6 is Lucy. After the robot says out its grounding result, according to conversation habits in general, the speaker will describe Lucy again to distinguish Lucy from No.6. The correct descriptions will be repeated. It is sure that the repeated descriptions are credible. If the acceptability of this description is too low (below a threshold), the robot will realize that its grounding result is wrong and it hasn't detected Lucy. Then the robot will perceive the surrounding environment again. What's more, another hint can be obtained from the speaker's feedback, there is no need to perceive the whole environment again. The position where Lucy most likely stands can be inferred from the QSRs. We sample the unoccupied pose P_u in the map and consider the pose p^* which satisfies formula 5 as the position where the referent most likely stands. $QSRs^*$ is the combination of spatial relation expressions which is most consistent with speaker's feedback.

$$(QSRs^*, p^*) = arg\ maxP(P_u|QSRs) \tag{5}$$

3.4 Word Grounding Module

Functionally, word grounding module evaluates the compatibilities between the word expressions from the dialog and the numeric attribute value from sensors. Theoretically, this is the problem of designing membership functions for fuzzy concepts. There are several membership function generation techniques, such as methods based on subjective perception, heuristic methods, methods based on



(a) (b)

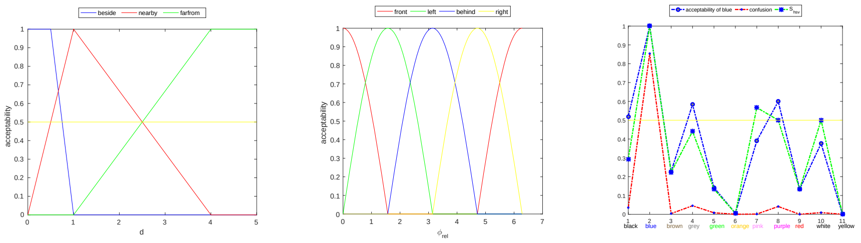
Fig. 2. (a) The distance between the target person and the furniture. The furniture is represented by a rectangle.(b) The relative angle of the target with respect to the landmark and the the partition of the four directional relations.

clustering, etc.[4] As for QSRs, we use the heuristic method which is designed based on rules to generate the membership functions for different spatial relations expressions. As for the color expressions, a method similar to clustering is used to design membership functions. Assume the face detector gives the coordinates of a person is (X, Y, Z) and color classifier predicts his clothes as C_M. The relationship prepositions and corresponding landmarks and color nouns extracted from the sentences is f_i, l_i and C_H. The position of the furniture (X_f, Y_f) can be obtained from the the grid map. The minimum applicability for all descriptions in this paper is set to 0.5.

Qualitative Spatial Relation QSRs consist of distance relations {nearby, beside, between, far from} and directional relations {left of, right of, front of, behind}. As to the directional relations, the landmark only can be the robot or the speaker. The distance d between the goal person and the furniture is shown in Fig. 2(a). $d = \sqrt{(X - X_0)^2 + (Y - Y_0)^2}$. The membership functions for the three binary relation(beside, nearby, far from) of the distance relations are designed in triangular shapes (Fig. 3(a)). When the distance d is close, the membership grade of beside is the highest; when d is far, the membership grade of far-from is the highest; otherwise, when d is moderate, the membership grade of nearby is the highest. The ternary relation (between) can be represented as the combination of two binary relations.

The four directional relations are showed in Fig. 2(b). Robot and speaker define the reference axis (view direction) which partitions the surrounding space. Then, the spatial relation is defined by the partition in which people lie with respect to the reference axis. Given the view direction $\boldsymbol{V} = (x, y)$, to determine the partition, we calculate the relative angle φ_{rel}:

$$\varphi_{rel} = \tan^{-1}\frac{Y - Y_L}{X - X_L} - \tan^{-1}\frac{y}{x} \tag{6}$$



(a) (b) (c)

Fig. 3. (a) The applicability of each distance relation along d. (b) The applicability of each directional relation within the interval $[0, 2\pi]$. (c) The acceptability of blue for different colors predicted by the classifier.

The membership functions for the four directional relations are designed as cosine curve shapes. For example, the membership function for FrontOf is:

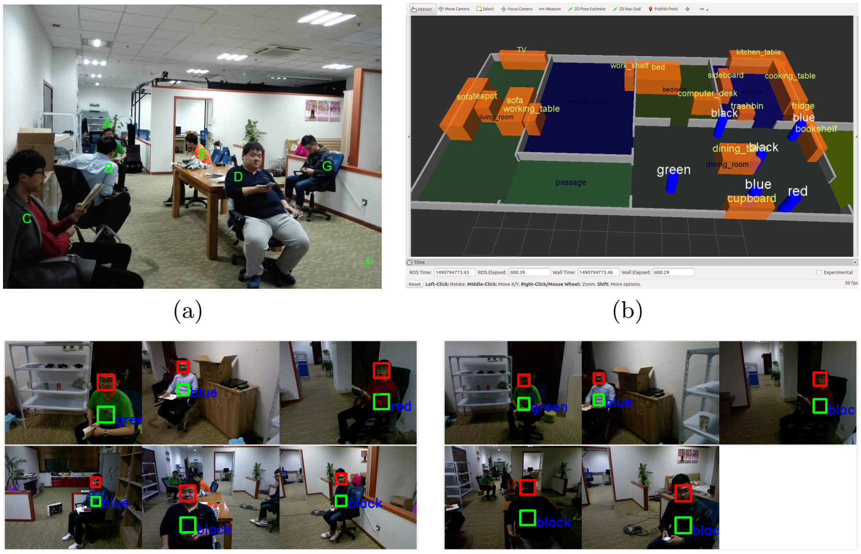
$$P(f_i = \text{FrontOf}|x) = max(cos(\varphi_{rel} - cnt_i), 0) \tag{7}$$

cnt_i for directional relations FrontOf, LeftOf, Behind, RightOf, is 0, 0.5π, π, and 1.5π respectively. From Fig. 2(b), it is intuitive that the greater the deviation from cnt_i, the smaller the possibility to accept the description f_i. The applicability of each directional relation within $[0, 2\pi]$ is show in Fig. 3(b).

Color Most computer vision works consider the eleven basic color terms of the English language: black, blue, brown, grey, green, orange, pink, purple, red, white, and yellow [18]. The visual similarity between C_H and C_M can be evaluated as the distance of their prototypes in HSV space. Color classifier can not be perfectly fit with human visual perception. The color classifier is likely to confuse some color pairs when under different illuminants. This results in its prediction differing from human's judgment. This also can be another evidence for their similarity. Therefore, the acceptability between C_H and C_M can be evaluated by the linear sum of their similarity in HSV $S(C_H, C_M)$ and the probability of confusing C_H with C_M $confusion(C_H, C_M)$ which is shown in formula 8.

$$P(C_H|C_M) = \alpha \cdot S(C_H, C_M) + \beta \cdot confusion(C_H, C_M) \tag{8}$$

Where α and β is normalized coefficients. We test the color classifier on the eBay data set [19] and get its confusion matrix. Reference to the method of clustering in HSV space in [14, p. 2], we get the similarity matrix. The acceptability of blue for different colors predicted by the classifier is shown in Fig. 3(c). The red line is the possibility of confusing blue with different colors. The classifier is most likely to confuse blue with grey and purple. The green line shows the similarity in HSV. Even though the similarity between blue and grey is lower than 0.5, due to the high possibility of confusing blue with grey, the acceptability of blue is improved above the minimum acceptability. This improves the tolerance of our system to classifier errors and make it more robust to scene accidental events.



(a) (b)

(c) (d)

Fig. 4. (a) The setup of our experiments. (b) The world model of the robot. The detection results of people are shown in cylinders. The color of the one who it thinks to be the target person will turn red. Otherwise, it keeps blue. The detection results under bright and dim illumination are shown in (c) and (d) respectively.

4 Experiments

During human-robot interaction, there are two factors that have influence on the grounding result. They are the perceptual results of computer vision algorithms and the descriptions given by its partner. Therefore, we verify the robustness of the system by the following two schemes. One is to change the illumination of the environment to affect the perception result. The other is to change the speaker or target person to change the descriptions. Therefore, we designed two groups of experiments under different light conditions. Under each light condition, for each of the speakers who have been detected by the robot, the robot is required to ground different target persons.

The setup of our experimental system is shown in Fig 4(a). There are 7 people in the room. They are marked A∼G. In the first group, all fluorescent lamps in the room are open. In the other, one lamp is turned off. The robot perceived the whole room. The detection results under bright and dim illuminations are shown in Fig. 4(c) and Fig .4(d). The robot uses natural language to describe its grounding results and display it in rviz to ensure its partner can understand it. We did 66 $(6 \cdot 6 + 5 \cdot 6)$ experiments in total. Three examples are shown in Table 1-3. At the left of the tables is the detection results by the robot. People detected by the robot are represented by squares. In the first example (under bright illuminations), the robot is required to ground Peter(F in Fig 5(a)) who

Table 1. Ground Peter

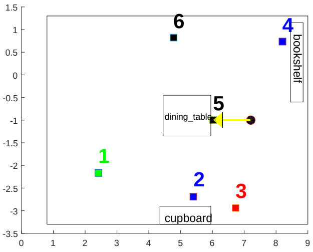
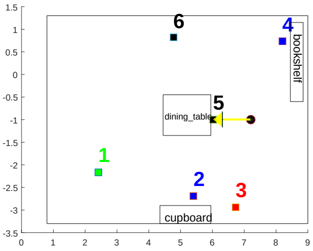
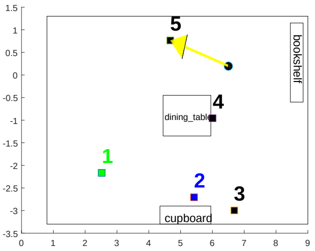
Speaker: D Ground Peter(F in Fig 4(a))	Dialog
	R: Can you tell which one is Peter? S: He wears a pink coat. (The robot misidentified blue as pink) R: I think the one is Peter. (x_3) He is nearby the cupboard and nearby the dining-table, his clothes is red. ($DP = 0.2755$, based on $P(pink\|x_3) = 0.6567$) S: No, he sits beside the bookshelf. R: I think this one is Peter. (x_4) He is beside the bookshelf. His clothes is blue. ($DP = 0.9989, P(beside\|x_4) = 0.917 > 0.5$) S: Yes

Table 2. Ground David who hasn't been detected

Speaker: D Ground David(E in Fig 4(a))	Dialog
	R: Can you tell which one is David? S: He sits beside the dining_table in an orange coat. R: I think the one is David. (x_3) He is nearby the cupboard and nearby the dining-table, his clothes is red. ($DP = 0.483$, based on $P(orange\|x_3) = 0.599$) S: No, he sits behind me and his coat is orange. R: I think I have missed him. ($DP = 0.444, P(orange\|x_6) = 0.293 < 0.5$) I think he is beside the dining_table($p^* = (3.95, -0.90)$) S: Yes

has been detected and his color has been predicted correctly. But during the conversation, the robot misidentified blue as pink. As red is most similar to pink, the robot thinks x_3 is Peter. After the speaker supplements more cues, according to the principle of maximum DP, the robot thinks x_4 is Peter. In the second example (under bright illuminations), the robot is required to ground David who hasn't been detected by the robot is shown in Table 2. Even though David has not been detected, according to QSRs, the robot calculates the most probable position where David exists. Therefore, this is a successful grounding. An example of unsuccessful grounding Jim who has been detected but whose color is misidentified is shown in Table 3. The error rate of the overall system is shown in Table 4. The description error is caused by speech misidentification or man-made mistakes. The grounding error is lower than the perception error and description error. It demonstrates that our grounding system is robust to the perception errors and description errors.

Table 3. Ground Jim whose color is misidentified

Speaker: G Ground Jim(C in Fig 4(a))	Dialog
	R: Can you tell which one is Jim? S: Jim sits beside the cupboard in a red coat. R: I think the one is David. (x_2)He is between the cupboard and the dining_table, his clothes is blue. ($DP = 0.456$, based on $P(beside\|x_2) = 0.936$) S: No, his coat is red. R: I think I have missed him. ($P(red\|x_3) = 0.294 < 0.5$) I think he is beside the cupboard($p^* = (5.15, -2.4)$) S: No

5 Conclusion

In this paper, we consider to help a robot recognize an unknown person through the descriptions consisting of spatial relations and coat color. We use a probabilistic model to model DP. The person described by the user can be grounded based on the criterion of maximizing DP. We use a word grounding module to evaluate the compatibility between the word expressions and the numeric attribute values. The acceptability of color descriptions is modeled base on visual similarity and the confusion matrix of the color classifier. This improves the tolerance of our system to classifier errors and make the system more robust to scene accidental events. For the case that the target person isn't detected by the robot, we use QSRs to infer the most likely location of the target, which is convenient for the robot to search again. Two groups of experiments are designed which indicates that our grounding system is robust to the perception errors and description errors. This work has been demonstrated on the Open Challenge test in RoboCup2016@Home. In our future work, in addition to QSR and color, more features can be added to our system. Moreover, our long-term goal is to make the robot more intelligent through the full use of the surrounding information.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (No. 61603368), the Youth Innovation Promotion Association of CAS (No. 2015373), and Natural Science Foundation of Anhui Province (No. 1608085QF134).

Table 4. The error rate of the overall system

Face Detection Error	21.4% (3/14)
Color Classification Error	27.3% (3/11)
Description Error	18.9% (56/297)
Grounding Error	12.1% (8/66)

References

1. A. Bai, F. Wu, and X. Chen. Online planning for large markov decision processes with hierarchical decomposition. *ACM Trans. on Intelligent Systems and Technology*, 6(4)(45), August 2015.
2. G. Buchsbaum. A spatial processor model for object colour perception. *Journal of the Franklin institute*, 310(1):1–26, 1980.
3. Y. Chen, F. Wu, W. Shuai, N. Wang, R. Chen, and X. Chen. Kejia robot - an attractive shopping mall guider. In *Proceedings of the 7th International Conference on Social Robotics*, pages 145–154, 2015.
4. G. Deschrijver and E. E. Kerre. On the relationship between some extensions of fuzzy set theory. *Fuzzy sets and systems*, 133(2):227–235, 2003.
5. A. Elfes. Using occupancy grids for mobile robot perception and navigation. *Computer*, 22(6):46–57, 1989.
6. T. Foote. tf: The transform library. In *Proc. of TePRA*, pages 1–6. IEEE, 2013.
7. P. Gorniak and D. Roy. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21:429–470, 2004.
8. G. Grisetti, C. Stachniss, and W. Burgard. Improved techniques for grid mapping with rao-blackwellized particle filters. *IEEE Trans. on Robotics*, 23(1):34–46, 2007.
9. Z. Huo, T. Alexenko, and M. Skubic. Using spatial language to drive a robot for an indoor environment fetch task. In *Proc. of IROS*, pages 1361–1366. IEEE, 2014.
10. L. Kunze, K. K. Doreswamy, and N. Hawes. Using qualitative spatial relations for indirect object search. In *Proc. of ICRA*, pages 163–168. IEEE, 2014.
11. C. Liu and J. Y. Chai. Learning to mediate perceptual differences in situated human-robot dialogue. In *Proc. of AAAI*, pages 2288–2294, 2015.
12. C. Liu, R. Fang, and J. Y. Chai. Towards mediating shared perceptual basis in situated dialogue. In *Proc. of Meeting of the Special Interest Group on Discourse and Dialogue*, pages 140–149, 2012.
13. D. Lu, Y. Zhou, F. Wu, Z. Zhang, and X. Chen. Integrating answer set programming with semantic dictionaries for robot task planning. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017.
14. Y. Lu, W. Gao, and J. Liu. Color matching for colored fiber blends based on the fuzzy c-mean cluster in hsv color space. In *Proc. of FSKD*, pages 452–455, 2010.
15. V. Mast, Z. Falomir, and D. Wolter. Probabilistic reference and grounding with pragr for dialogues with robots. *Journal of Experimental & Theoretical Artificial Intelligence*, 28(5):889–911, 2016.
16. V. Mast and D. Wolter. A probabilistic framework for object descriptions in indoor route instructions. In *International Conference on Spatial Information Theory*, pages 185–204, 2013.
17. A. Mojsilovic. A computational model for color naming and describing color composition of images. *IEEE Trans. on Image Processing*, 14(5):690–699, 2005.
18. J. Van De Weijer and F. S. Khan. An overview of color name applications in computer vision. In *Workshop on Computational Color Imaging*, 2015.
19. J. Van De Weijer, C. Schmid, J. Verbeek, and D. Larlus. Learning color names for real-world applications. *IEEE Trans. on Image Processing*, 18(7):1512–1523, 2009.
20. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. of CVPR*, 2001.
21. F. Wu, S. Ramchurn, and X. Chen. Coordinating human-UAV teams in disaster response. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pages 524–530, 2016.