Enhancing Object Detection by Using Probabilistic Spatial-Semantic Knowledge

Malgorzata Goldhoorn[1*], Ronny Hartanto[2]

[1] Department of Computer Science, Robotics Group, University of Bremen, 28359 Bremen, Germany.
[2] Faculty of Technology and Bionics, Rhine-Waal, University of Applied Sciences, 47533 Kleve, Germany.

* Corresponding author. Tel.: +48421178454845; email: malgorzata.goldhoorn@informatik.uni-bremen.de
Manuscript submitted July 24, 2015; accepted December 29, 2015.
doi: 10.17706/jcp.12.1.68-75

Abstract: Autonomous mobile robots that act in human living environment and perform complex tasks there must be able to obtain relevant information from the surrounding and reason about this knowledge. The system's ability to recognize objects and assign them semantic description is a vital capability for the robots in order to carry out such a complex task. Dealing with objects in the environment is not a trivial problem but rather a challenging one and detection systems relay only on information extracted from their noisy sensors are often not sufficient to recognize objects properly. Therefore, this paper presents a new approach for object recognition, in which probabilistic spatial-semantic knowledge is applied to improve the recognition result.

Key words: Object recognition, probabilistic methods, spatial reasoning, spatial relations.

1. Introduction

Robotic systems, such as assistive or service robots, able to perform actions in human living environments should support humans in their everyday life. Locating objects while executing daily tasks such as "fetch and carry", however, requires robust perception skills, where objects must be detected even if the result of the perception system is uncertain. Only if the objects have been recognized correctly can the robots perform the given task properly. Moreover, the semantic perception and indoor scene labelling are fundamental problems [1] and not yet solved. This paper presents a new approach resulting from an amalgamation of a common feature-based perception system with spatial context about object co-occurrences and relations to gain reliable object classification. To achieve this, *Probabilistic Qualitative Spatial Relations (PQSR)* are modelled with *Spatial Potential Fields (SPF)*. Using PQSR, probability distribution of several spatial relations can be obtained. In turn the SPF can be used as a method to model the PQSR. The objective of this paper is to present how such relations are defined and used to boost the object recognition process. The idea behind this approach lies in the improvement of the classification results by applying probabilistic spatial-semantic knowledge about objects' arrangement and relations. The spatial context information is used to predict the appropriate object class.

2. Related Work

Not only because of the increasing availability of low-cost RGB-D sensors like the Microsoft Kinect or the Primesense's cameras but also especially due to the fact that robots need this capability to perform

complex tasks in human living environments [2], [3] has object recognition become into one of a very active topic in robotics recently. One current work in this field [4] presented a contextually guided semantic labelling and search algorithm. In this method a graphical model with geometrical features and contextual relations between objects are used. The model is trained using a maximum-margin learning approach. The authors use merged point clouds from indoor environments obtained with a Kinect RGB-D sensor. To acquire a better view of occluded objects in the scene active object recognition is performed. Aydemir *et al.* [5] described an approach for active visual search. In this process topological relations between objects are used to create a potential search action and find the object. The authors equip the robot with so called PDF (probability distribution function) as initial information for the search process. Similar to this work in [4] the next-best view algorithm is applied to deal with occlusion. Other recent work which covers the challenges of object labelling in indoor environments using RGB-D data has been presented in [6] .The authors develop and evaluate a method for scene labelling that combines RGB-D features and contextual models by using MRFs (Markov Random Fields) and segmentation. To
gain the RGB-D features like: gradient, color and a surface normal, kernel descriptors are used. Ali *et al.* [7] proposed the use of a context model to improve the detection result of other related objects in the scene. In this approach object co-occurrence information is used to perform recognition of the object category. In the work presented by Xiong *et al.* [8] planar patches extracted from a point cloud are labelled with geometric labels such as wall, floor and celling. To model geometrical relationships such as orthogonal, parallel, adjacent and coplanar between objects, Conditional Random Field (CRF) is used.

3. Feature-Based Perception System

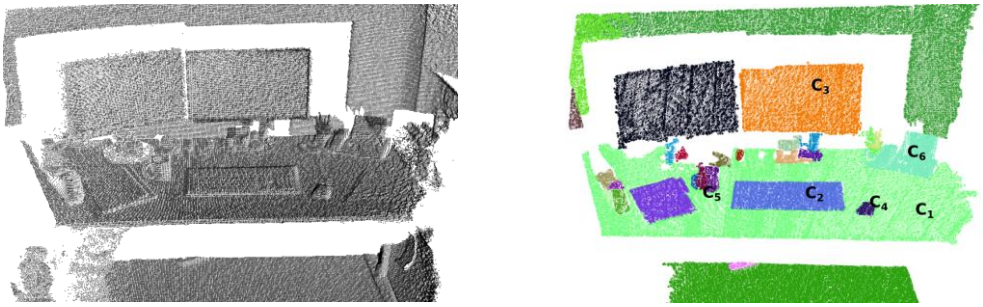
A probabilistic feature-based perception method is used to recognize objects from the environment. This system based on the approach presented by Eich *et al*. [9], in which an optimized region growing algorithm is applied to recognized regions from a point cloud. This paper describes an extended object perception system which goes beyond the previously mentioned method. We extended the object perception system by incorporating a probabilistic reasoning approach, in which spatial objects' features extracted from the data are used for objects' classification. The perception system predicts the probability distribution for each object and its class. The feature-based object categorization method works as follows:

1) The point cloud is segmented into object clusters (potential objects) according to given starting parameters
2) The spatial features of each object cluster are extracted and evaluated
3) For each potential object segment:
   a) compute its similarity to each object class
   b) assign a probability for the cluster to each known object class based on the similarity according to the spatial features of the cluster
   c) ensue the probability distribution of the given object class
4) Finally, the reasoning method returns the recognized objects for all segmented clusters

Fig. 1(a) shows a raw point cloud obtained from a Kinect camera and the segmentation result. Given a possible object class K, feature vector Φ and factor f, the probability for the cluster being a certain object class with regards to its measured features $x=(x_1,\ldots,x_n)$ can be calculated as follows:

$$P(K|\Phi)=\frac{\sum_{i=1}^{n}P(K\mid x_i)\cdot f_i}{\sum_{i=1}^{n}f_i}. \quad (1)$$

Table 1 shows an exemplary result of the feature-based classification method from one office desk. In this table, the columns denote the clusters with their ground truth values. The rows list the possible object classes whose values denotes the probability for a cluster belonging to a given object class. From the Table 1 it can be seen that some objects like, for example, *mouse* has been classified either as a *mouse* or a *mug*. This is caused by the uncertainty in the perception system.



(a) (b)

Fig. 1. Result of feature-based perception system. (a) An exemplary point cloud, (b) Result of the segmentation.

Table 1. An Exemplary Result of the Perception System with Predicted and Real Object Class

	Ground Truth (Segment ID)					
Predicted Object Class	Table (C_1)	Keyboard (C_2)	Monitor (C_3)	Mouse (C_4)	Mug (C_5)	Phone (C_6)
Table	0.3987	0.1506	0.0578	0.1592	0.1047	0.1053
Keyboard	0.0477	0.2605	0.0985	0.0475	0.0474	0.1842
Monitor	0.0188	0.1049	0.4611	0.0524	0.212	0.1321
Mouse	0.0468	0.0492	0.0356	0.3429	0.3581	0.0706
Mug	0.0713	0.0691	0.0331	0.1469	0.1293	0.0678
Phone	0.0384	0.167	0.1697	0.1408	0.2256	0.2232

4. PQSR-Based Object Classification

As mentioned in 3, the feature-based perception method results in probability distribution over given object classes. These results refer only to the information extracted from the sensor data. Therefore, such method, similar to other common bottom-up approaches, has its limitations. For example, due to the occlusion some objects can not be recognized properly. Another challenge is to recognize objects which have high similarity in their spatial features, like for example a *mug* and *mouse*. Given the arrangement of the objects as a spatial information, the reasoning component can predict that an objects located to the right of the *keyboard* is likely to be a *mouse* than a *mug*.. In human environment, the likeliness of object placement in certain environment can be statistically analyzed. As a result, a spatial relation between an object with another object can be derived, e.g. a *keyboard* is usually *in front of* a *monitor*. Because this knowledge may change according to the environment, our system learns it from the real world data as described in Section 4. 2.

4.1. Probabilistic Qualitative Spatial Relations

In this work the qualitative spatial relations are extended to probability to gain more precise models of the spatial relations. Due to this extension the spatial relations can be defined more precisely by describing how accurate a given relation is. This likelihood is used as additional information to support the recognition process. Because the perception method relays *entirely* on the spatial object features extracted from sensor, which is noisy, the system needs additional knowledge to recognize the object reliably. In this approach the

following probabilistic qualitative relations: *on, above, near, leftOf, rightOf, behind, inFrontOf* are used. As some of those relations depend of the perspective from that an object is observed, we distinguish between *the binary* and *projective binary relations*. The definitions of both relations are described as follows:

Definition 1: Let $D \subseteq L \times C \times \Psi$ be a set of possible objects in the domain, where L describes a set of possible labels $C = \mathbb{R}^3 \times \mathbb{H}$ a set of possible coordinate systems, in which a given object can be located, and $\Psi \subseteq \mathbb{R}^3 \times [0;1]$ a set of three-dimensionally colored points. Then, the set of *binary probabilistic spatial relations* R_β is defined as follows:

$$R_\beta \subseteq \left\{ D^2 \to [0;1] \right\}. \tag{2}$$

Let C be a set of possible coordinate systems which refers to reference view points. Then, the set of *projective binary probabilistic relations* R_τ is defined as follows:

$$R_\tau \subseteq \left\{ D^2 \times C \to [0;1] \right\}. \tag{3}$$

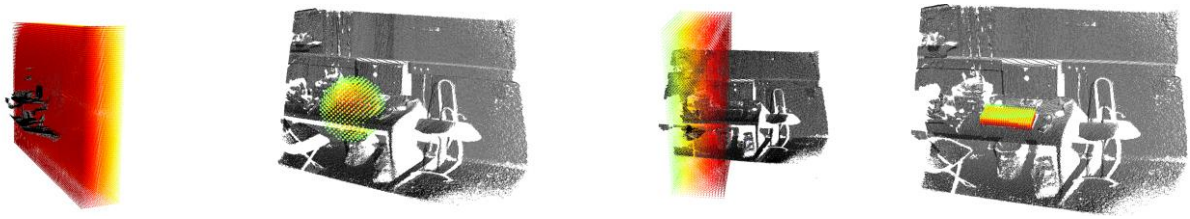
In the world model of a given environment the spatial-semantic knowledge about typical objects arrangements is represented by functions f_β and f_τ.

Definition 2: The co-occurrence probability of two objects defined as f_β (4) for binary and f_τ (5) for projective binary relations:

$$f_\beta = R \times D^2 \to [0;1] \tag{4}$$

and

$$f_\tau = R \times D^2 \times C \to [0;1]. \tag{5}$$



(a) (b) (c) (d)

Fig. 1 SPFs for several PQSRs. (a) SPF of behind monitor, (b) SPF of near keyboard, (c) SPF *Ofleft* of table, (d) SPF of on keyboard.

4.2. Spatial Potential Fields

The *Spatial Potential Fields (SPF)* describe the model of qualitative spatial information about object's relations in a probabilistic manner and represent this information in the model of the environment. The very earlier concept of the SPFs can be found in [10]. Our idea was inspired by the potential fields method, which has been used in the area of robots navigation [11] as the obstacle avoidance approach. The SPFs are calculated by combination of spatial knowledge extracted from real-world data with the calculated intensity value of a given spatial relation for a given object (see Fig. 2). These are defined formally as follows:

Definition 3: Let $t, s \in D$ be target and reference object, R a given relation and c a view point of the

system. Then the SPFs are defined as follows:

$$SPF_{\tau}(R,t,s,c) = f_{\tau}(R,t,s,c) \cdot R(t,s,c) \tag{6}$$

and

$$SPF_{\beta}(R,t,s) = f_{\beta}(R,t,s) \cdot R(t,s). \tag{7}$$

For all defined SPFs, let t be the target and s the reference object, then t_p and s_p are the centers of the objects' masses, t_w, s_w denote the objects' widths and t_h, s_h the objects' heights, respectively. Based on this, we define the SPFs for the spatial relations listed in the following. All values for the relations' calculation are normalized, by:

$$R(t,s) = \begin{cases} 1 - v_r(t,s), & if\ 0 \le v_r(t,s) \le 1 \\ 0, & \text{otherwise} \end{cases}. \tag{8}$$

Definition 1 (Near): The probability of t being *near* to s is defined as:

$$v_{near}(t,s) = \sqrt{\left(t_p - s_p\right)^2} \Big/ \sqrt{\left(s_w s_h + t_w t_h\right)/2}. \tag{9}$$

Definition 2 (Above): Let the subscript o denote the orientation while ϕ refers to the world coordinate system. Then, the probability of t being *above* to s is defined as:

$$v_{\text{above}}(t,s) = \begin{cases} v_{\text{near}}(t,s), & \text{if above}(t,s) \\ 0, & \text{otherwise} \end{cases}, \tag{10}$$

where

$$\text{above}(t,s) = \exists \alpha,\eta,\gamma : \left(s_p + \alpha\, \vec{s}_{o_x} + \eta\, \vec{s}_{o_y} = t_p + \gamma \vec{c}_{\varphi_{o_z}}\right) \wedge \left(|\alpha| \le \frac{s_w}{2}, |\eta| \le \frac{s_h}{2}, \gamma > 0\right). \tag{11}$$

Definition 3 (On): Let λ>0 be a given threshold value for the maximal height allowed in the on relation. Then, the probability of t being *on* to s is defined as:

$$v_{on}(t,s) = \begin{cases} \dfrac{\left(s_0\left(t_p - s_p\right)\right)_z}{\lambda}, & \text{if on}(t,s) \\ 0, & \text{otherwise} \end{cases}, \tag{12}$$

where

$$on(t,s) = \lambda \ge \left(s_0\left(t_p - s_p\right)\right)_z \ge 0 \wedge \left|\left(s_0\left(t_p - s_p\right)\right)_x\right| \le \frac{s_w}{2} \wedge \left|\left(s_0\left(t_p - s_p\right)\right)_y\right| \le \frac{s_h}{2}. \tag{13}$$

Definition 4 (*LeftOf*): Let ξ >0 be a given threshold value for the maximal distance between s and t. Then, the probability of t being *leftOf* to s is defined as:

$$v_{left}(t,s,c) = \begin{cases} \dfrac{\left|\left(c_o t_p\right)_y - \left(c_o s_p\right)_y\right|}{s_w \xi}, & \text{if } \left(c_o t_p\right)_y > \left(c_o s_p\right)_y. \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

The relation *rightOf* is calculated analogously to the relation *leftOf,* with the difference that now a greater projective y-value is required to satisfy the relation. By the relations *inFrotOf* and *behind* the same formula is used as in Definition 4, just replacing y- by x-axis.

In our method, the probability values for the target objects being in the spatial relation with a reference object are learned from the real-world data.

Definition 8: Formally, the probability is calculated using the average probability of all observed occurrences between the objects $t, s \in D$ being in the given relation R:

$$R_k(t,s) = \frac{\sum_{t \in D} \sum_{s \in D} R(t,s)}{|t \in D|}. \tag{15}$$

4.3. Field Intensity

Based on the learned knowledge from the data the SPF are calculated as described in Section 4.2. The SPF provide probabilistic spatial knowledge about objects relations. However, to predict the probability that a given object class is located at the given position, several SPF must be taken into account. This combination we called *Field Intensity (FI)*. This FI denotes the probability of an object belonging to a certain object class given its relation to other object and can than be used as a qualifier for the object detection system.

Definition 9: Let $t, s \in D$ be target and reference object, R a given relation and c a viewpoint of the system. The FI of given SPFs is calculated as follows:

$$FI(t,c) = \frac{FI_\beta(t) + FI_\tau(t,c)}{|R_\beta| + |R_\tau|} \tag{1}$$

where

$$FI_\tau(t,c) = \sum_{r_\tau \in R_\tau} \sum_{s \in D} SPF_\tau(r_\tau, t, s, c) \tag{2}$$

and

$$FI_\beta(t) = \sum_{r_\beta \in R_\beta} \sum_{s \in D} SPF_\beta(r_\beta, t, s). \tag{3}$$

Only after the FI value is known can the probability for the object belonging to a given object class be calculated. This is done by the multiplication of the result obtained from the feature-based perception system with the field intensity value.

5. Experiments

To validate our approach we sampled point cloud data from several offices of our institute using Kinect V2. For the experiments we gathered 26 different office scenes, which we used for training and evaluation with 6 and 20 respectively. The recorded point clouds were combined into segments by using our perception system as described in 3. The part of the data was manually labelled with ground truth information for learning purposes. In the experiments the following object classes: *Table, Keyboard, Monitor, Mouse* and *Phone* were used. The objects are arranged differently in each office, depending on their owner's preferences. Therefore, not all offices have the mentioned object classes. For each experiment we

first applied the feature-based perception system, which returns the probability distribution for each object and its class. The result of the perception system serves as an input for the PQSR-based method, in which spatial-semantic knowledge is applied to improve the classification result. The result of the feature-based perception system is shown in Table 2. The columns of the table denote the ground truth object classes and the rows denote the predicted object classes. The values correspond to the *mean* probability over all scenes. The values are not normalized in order to avoid the distorting the result. The result in the table shows that most of the objects were classified correctly; however, the assignment is not sufficient. Not all objects are recognized correctly. Because the values refer to the mean probability, there is still an incorrect assignment of object classes in some of scenes.

Table 2. Recognition Result of the Feature-Based Perception Method (Mean Probability)

Predicted	Actual					
	Table	Keyboard	Monitor	Mouse	Mug	Phone
Table	0.38716	0.16181	0.06971	0.15553	0.10228	0.10598
Keyboard	0.04646	0.28035	0.10491	0.04905	0.06509	0.15201
Monitor	0.0218	0.12549	0.44287	0.05032	0.15657	0.1332
Mouse	0.04567	0.04368	0.0306	0.33368	0.28253	0.07198
Mug	0.07075	0.0682	0.03809	0.18203	0.25229	0.0677
Phone	0.03818	0.1216	0.10574	0.11345	0.17322	0.20469

Table 3. Recognition Result after Applying the PQSRs (Mean Probability)

Predicted	Actual					
	Table	Keyboard	Monitor	Mouse	Mug	Phone
Table	0.00344	0.00152	0.00061	0.00129	0.001	0.00071
Keyboard	0.0004	0.00249	0.00104	0.0004	0.00066	0.00102
Monitor	0.00012	0.00072	0.00387	0.00024	0.00103	0.00085
Mouse	0.0004	0.00042	0.00026	0.00294	0.00227	0.00041
Mug	0.00045	0.00043	0.00029	0.00105	0.00249	0.00046
Phone	0.00019	0.0006	0.00083	0.0005	0.00113	0.0012

Table 3 shows the recognition result after applying the probabilistic spatial-semantic knowledge as described in 4. This knowledge was extracted from the real office environments without the use of artificial objects arrangements and refers to the human preferences who work in in the given offices. Similar to the previous table the resulting values denote the *mean* probability over all data sets. The results in the Table 3 show, that by applying the PQSRs the object classification has been improved. For instance, the feature-based perception system classifies the *mug* as a *mouse* which is wrong. This was caused by the high similarity of the both objects' appearance. However, given the learned knowledge, *mugs* are usually placed *to the left* of a *keyboard*, where *mousses* are usually located to *the right of* a *keyboard*. Applying this spatial-semantic information can improve the correctness of the classification system, compared to just using the perception based exclusively on spatial information obtained from the sensor. Based on the shown results, it can be concluded that the application of the PQSR-based method in object recognition system can improve the overall detection result and eliminate the incorrect objects' assignment.

Acknowledgment

This work has been supported by the Graduate School SyDe, funded by the German Excellence Initiative within the University of Bremen's institutional strategy.

References

[1] Ariadna, Q., & Torralba. A. (2009). Recognizing indoor scenes. *Computer Vision and Pattern Recognition*.

[2] Galindo, C., Fernandez-Madrigal, J.-A., Gonzalez, J., & Saffiotti, A. (Nov, 2008). Robot task planning using

semantic maps. *Robotics and Autonomous Systems*.

[3] Pangercic, D., Tenorth, M., Jain, D., & Beetz, M. (2010). Combining perception and knowledge processing for everyday manipulation. *Intelligent Robots and Systems*.

[4] Anand, A., Koppula, H. S., Joachims, T., & Saxena, A. (2012). Contextually guided semantic labeling and search for three-dimensional point clouds. *The International Journal of Robotics Research.*

[5] Aydemir, A., Sjoo, K., Folkesson, J., Pronobis, A., & Jensfelt, P. (2011). Search in the real world: Active visual object search based on spatial relations. *Robotics and Automation*.

[6] Fox, D., *et al.* (2012). Rgb-(d) scene labeling: Features and algorithms. *Computer Vision and Pattern Recognition*.

[7] Ali, H., Shafait, F., Giannakidou, E., Vakali, A., Figueroa, N., Varvadoukas, T., & Mavridis, N. (2014). Contextual object category recognition for rgb-d scene labeling. *Robotics and Autonomous Systems*.

[8] Xiong, X., & Huber, D. (September 2010). Using context to create semantic 3d models of indoor environments. *Proceedings of the British Machine Vision Conference*.

[9] Eich, M., & Dabrowska, M. (2010). Semantic labeling: Classification of 3d entities based on spatial feature descriptors. *Robotics and Automation*.

[10] Goldhoorn, M., & Hartanto, R. (2014). Semantic labelling of 3d point clouds using spatial object constraints. *Computer Vision, Imaging and Computer Graphics Theory and Applications*.

[11] Borenstein, J., & Koren, Y. (1991). The vector field histogram-fast obstacle avoidance for mobile robots. *Robotics and Automation*.

Malgorzata Goldhoorn received her diplom (master) degree in computer science from the University of Bremen, Germany in 2011. She is currently a PhD candidate at the University of Bremen within the Robotics Research Group and Graduate School System Design. Her research interests are robotics and artificial intelligence with focuses on robot perception, computer vision, probability theories, knowledge representation, spatial reasoning, and machine learning.

Ronny Hartanto is a professor for computer engineering at Rhine-Waal University of Applied Sciences in Kleve, Germany. He received his Ph.D. from Osnabrueck University, Germany, in 2009. His research interests are mobile robotics and artificial intelligence with focuses on knowledge representation and reasoning, plan-based robot control, planning algorithms, and machine learning.