Human-Robot Collaborative Scene Mapping from Relational Descriptions

Eloy Retamino Carrión and Alberto Sanfeliu

Institut de Robótica i Informática Industrial, CSIC-UPC, Llorens i Artigas 4-6, 08028 Barcelona, Spain

Abstract. In this article we propose a method for cooperatively building a scene map between a human and a robot by using a spatial relational model employed by the robot to interpret human descriptions of the scene. The description will consist in a set of spatial relations between the objects in the scene. The scene map will contain the position of these objects. For this end we propose a model based on the generation of scalar fields of applicability for each of the available relations.
The method can be summarized as follows. In first place a person will come into the room and describe the scene to the robot, including in the description semantic information about the objects which the robot can't get from its sensors. From the description the robot will form the "scene mental map". In second place the robot will sense the scene with a 2D range laser building the "scene sensed map". The objects positions in the mental map will be used to guide the sensing process. In a third step the robot will fuse the two maps, linking the semantic information about the described objects to the corresponding sensed ones. The resulting map is called the "scene enriched map".

1 Introduction

In many everyday life tasks and situations we make use of spatial relations for resolving references (eg. "The scissors are in the left drawer"), explaining a wished objects layout (eg. "put the plant close to the left corner") or focusing someone's attention into a certain region (eg. "look behind the table"). In our work a person will use this kind of relational language for describing a scene to a robot in order to help it in the mapping of the scene and to improve the map itself. The scene will consist in a set of objects distributed inside a room which dimensions are known by the robot.

We find three possible situations which can occur in this kind of collaborative mapping process: (1) the person and the robot are together while mapping the scene, (2) the robot first senses the scene and afterward the person describes it, (3) the person describes the scene and afterward the robot senses it.

In our work we will face the third situation. The person will first describe the whole scene using the aforementioned spatial relations. From the description the robot will build a map of the scene, composed of a representation of the objects located at their respective positions. This "mental map" has two purposes in the

scene mapping process. In first place, it will be used for guiding the robot sensing process. That is, after interpreting the description the robot will go into the room and seek the described objects in the positions they have in the mental map. In second place, during the description the person may add semantic information about the objects which the robot could need for fulfilling its duties. For example, a post-delivering robot who has to leave a letter at Mike's desk should know who is the owner of each desk in its map. As the described objects are matched with their corresponding sensed objects, this semantic information can be trespassed to the latter.

In the rest of the article we will distinguish between four different scene maps. (1) The "human mental map", the map the person forms in his mind from viewing the scene. (2) The "robot mental map", the map the robot builds from interpreting the human description of the scene. (3) The "robot sensed map", the map the robot builds from sensing the scene. (4) The "enriched map", built from the fusion of the robot sensed and mental maps. We call this last map enriched because it mixes the accuracy of the robot sensors with the semantic information provided by the person in his description.

We would like to remark that the generation of the "robot mental map" can facilitate by itself the human robot interaction in many different cooperative tasks apart from mapping an scene. Reorganizing the objects in a room, defining an strategy for searching an object or exploring an area. For all of these tasks the robot has to interpret the "mental map" the person has in mind and to verify it using its own sensors.

The contributions of the present work are: (1) A model for interpreting spatial relations. (2) A method built upon the former to let the robot to map a scene in collaboration with a human. In the rest of the article we will, present the related work, the relational model, the method and the results of the experiments conducted for testing it.

2 Related Work

There are numerous attempts of interpreting qualitative spatial relations. In [1] Moratz et Al. define a computational model for the projective spatial relations similar in spirit to the one developed here. In their work each relation defines a canonical direction and depending on the position of the referenced object with respect to this direction the relation became true or false. In their experiments they tried to understand how people express spatial knowledge by asking the subjects to tell the robot to go towards one of the objects in the scene (ie. to uniquely determine that object from the others in the scene). The drawback of their work is that the conditions for a spatial relation to be fulfilled are boolean which is against the intrinsic vagueness of these relations.

In [2] Stopp et Al. use a computational model of the topological (near) and projective (front, behind, left, right, above, bellow) relations for accessing to a robotic arm through natural language. Their model [3], as the one presented in here, is based on the concept of the continuous decay of the applicability of

the spatial relations as we separate from an ideal condition which defines the relation.

In the same direction, Kelleher et Al. [4] introduce the fact, also contemplated here, that the applicability of the projective relations decay with the distance to the referent object. They also state that the "size" of the generated fields is proportional to the size of the referent object.

As far as we know, all the previous works related with the interpretation of qualitative spatial relations face the case of single relations between objects which positions are deterministically known, therefore not considering possible uncertainties. Just in [3] Gapp defines the way to perform single compositions between relations. In the work presented here however, is required to interpret the description of a whole scene without the support of any perceptual information, ie. to "imagine" the whole scene. This requirement led to the development of a general framework for composing spatial relations, estimating probability distributions for objects which positions are known just by spatial relations with other objects and the interpretation of spatial relations between these "imagined objects". Regarding this last point, in [5] Mavridis et Al. let a robot to imagine objects on top of a table at positions expressed by spatial relations, but they are more focused in the maintenance of a 3D representation of the environment from the robot sensors and in the interpretation of spatial relations between sensed objects. In their work the variety of spatial relations is very limited. Also they can't be composed neither be expressed between imagined objects.

Regarding the existing approximations for building maps from relational descriptions, they result quite simplistic, or lets say, the accuracy of the map neither the interpretation of the spatial relations itself weren't the main objective of the research. In [6] Coyne et Al., develop a model for interpreting narratives (ie. for generating a 3D representation from them), but they are more focused in the aesthetics of the representation and in the natural language processing. Though their work contemplates an extensive vocabulary including many spatial relations, the model for the latter seems to be too deterministic. That is, the spatial relations seems to define fixed distances between the objects no matter the context (though they don't provide any details at all about this model).

The process of building a map from a description presents several similarities with the well known in robotics SLAM problem [7,8], though it presents important differences which forbid to undertake it using the same approaches. In the former a person uses a set of qualitative spatial relations for expressing to the robot the positions of the objects in a scene and his own. Since the point of view from which the spatial relations are referred influences their interpretation, the robot must infer the position of the objects and the person's at the same time. Up to here the similarities. The person's position is expressed also through relations with other objects introduced in the description, that is, there is no explicit odometry information. This makes a distinction with most of the situations faced in robotics, for example [9]. More than that, as it will be seen, the objects probability distributions extracted from the spatial relations are far from being gaussians, which makes inapplicable any EKF Slam algorithm. As

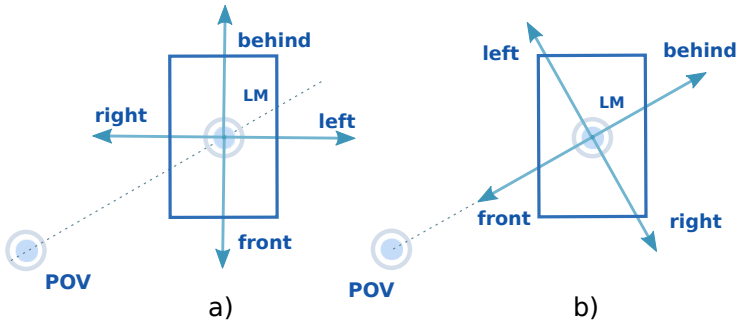

Fig. 1. Canonical directions for the projective relations in the intrinsic (a) and extrinsic (b) cases.

said above, the approach taken here for building the "mental maps" will be to extract probability distributions for the positions of the referenced objects in the given spatial relations and to reduce the uncertainty by directly composing the relations given for each object.

Finally, as it's been said, the mental map built from the description is used by the robot to guide the sensing process, ie. to direct its sensors to the positions where according to this map the actual objects more probably are. In the same direction, [10] Aydemir et Al. use background knowledge and the interpretation of spatial relations to perform indirect object searching. For example, if the robot has to look for a cup in its environment and it knows that cups are usually on top of tables; it looks first for a table, computes the region determined by the expression "on the table" and looks for the cup inside that region. Their model though, just contemplates two spatial relations: in, on. Logically they don't consider compositions and the spatial relations can't be directly used by a person for communicating spatial knowledge.

3 Relational Description Model

In this section we specify the model used by the robot to interpret the scene descriptions. The approximation taken for this interpretation is based of the generation of scalar "Fields of Applicability" (FOAs from now) for each of them. These scalar fields represent the distribution of the applicability of a certain spatial relation in every point of space. The characteristics of each FOA will depend on the spatial relation which it represents and on the pose and geometry of the objects involved in the relation (eg. the table in "on the left of the table").

3.1 Preliminary Concepts

In any grounded spatial relation there are several objects implied, each of them fulfilling a different function. Before going with the generation of the FOAs is important to define them.

In our syntax, the *Point of View* (POV) is the object which states the position from which the relation is expressed (usually corresponding to the person referring the relation). The *Landmark* (LM) will be the object used as referent (eg. the table in "the chair on the right of the table"). Finally, we will designate the *Trajector* (TR) as the referenced object (eg. the chair in the previous sentence). The FOAs will be generated using the POV and the LM. The applicability of a potential TR for a relation will be measured by evaluating the corresponding field on its center of gravity.

We will consider two types of relations, being the FOAs generated by each type closely related: topological and projective. The *topological relations* ("near", "far" and "close to" in our model) are proportional to the distance between the LM and the TR. The *projective* ones ("left", "right", "front", "behind") define a canonical direction. In this case the applicability will decay with the angular deviation of the vector $\overrightarrow{LM, TR}$ from that direction.

The next idea we must take in consideration is that there are more than one possible frame of reference in which a relation can be interpreted [11]. In the *intrinsic* case the frame is defined exclusively by the LM orientation (Fig 1a). The front direction is determined by its physical or semantic characteristics (eg. the side in the direction of motion in a mobile object).

In the *extrinsic* case, the frame is defined by the positions of the POV and the LM (Fig 1b), being the front direction the one going from the LM to the POV.

In the performed experiments all the employed objects are cylindrical, thus they don't have an intrinsic front. In other case the influence of each frame of reference should be decided. Generally is accepted that in case of competition, the intrinsic frame dominates [12], though the concrete "weight" of each frame is not a decided matter.

3.2 Fields Of Applicability

Their values goes from 0 to 1, being the applicability for a relation null if 0 and maximum if 1.

As the purpose of the FOAs is to specify objects positions in a 2D map, they will be defined in the euclidean plane. In the expressions where the POV, LM or TR appears, it must be understood that they refer to their projections on the XY plane.

Proximity field. This field doesn't semantically correspond to any of the mentioned relations, but it will form part of the rest of the FOAs expressions. It expresses the concept of proximity between two objects and obviously decreases with the distance between them.

The reason for using the "proximity" instead of directly the distance between objects is that the former encompasses contextual factors which must be taken into account for the correct interpretation of the spatial relations. These are: (1) the size of the involved objects and (2) the size of the scene itself. For example,

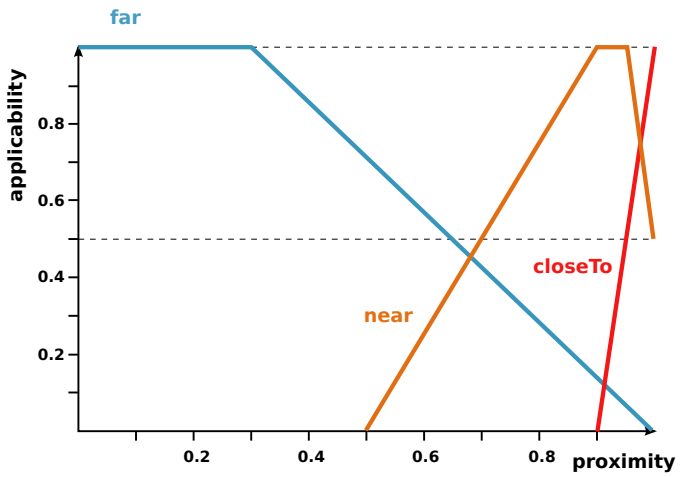

Fig. 2. Parametrization of the topological relations as a function of the proximity

the relation "near the house" clearly implies a larger area than "near the pen". By defining the "near" FOA in terms of the "proximity" instead of the distance this circumstance can be implicitly considered. In the same way, a correct interpretation of "near the house" must lead to a larger region if the considered scene is the whole city than if it's reduced just to the house neighborhood.

The former intuitive concepts are considered in the proximity expression as two constant factors which affect the proximity rate of decay with the distance (d_{max}, a_n). In turn, this decay is modeled as a linear function:

$$proximity(LM, P) = \begin{cases} 1 - \left(\frac{d}{d_{max}}\right)/a_n \ , \frac{d}{d_{max}} \leq a_n \\ 0 \ , \frac{d}{d_{max}} > a_n \end{cases}$$
$$a_n = \frac{a}{a_{max}}$$

where 'P' is a point of the plane, 'd' the euclidean distance between 'P' and the closest point of the LM and 'a' the area of the LM. 'a_n', the normalized area, expresses the fact that the proximity must decay slower for larger LMs [4]. It's defined as the area of the LM divided by the area of the larger object in the scene ('a_{max}').

'd_{max}' corresponds to the maximum distance in the scene. This factor expresses that the same distance must correspond to larger proximity values in larger scenes.

Projective relations. The values for the FOAs generated from these relations will decay with two factors. (1) the angle α between the vector $\overrightarrow{LM, P}$ and the canonical direction. (2) the distance with the LM. The latter assertion is supported by Kelleher et Al. [4], but intuitively projective relations "loose" definition

with the distance. As a extreme case, if a chair is 20 Kms far from a table it would never be said to be on the left of it, independently of the angle they form.

The canonical direction for each relation in the extrinsic and intrinsic cases are the ones in the Fig 1 [11]. Once set the canonical direction, the expression of the field for all the projective relations is:

$$projective(POV, LM, P) = \begin{cases} \left[1 - \frac{\alpha}{\alpha_{max}}\right] \cdot proximity(LM, P) \ , \alpha < \alpha_{max} \\ 0 \ , \alpha > \alpha_{max} \end{cases}$$

where $\alpha_{max} = 90^o$. As it was argued in the former subsection, the proximity was used instead of the distance between P and the LM. The decreasement with the angular deviation has been modeled as a linear function of the angle α.

Topological relations. The topological relations have been modeled as lineal parametrizations of the proximity, in accordance with the interval of distances for which each of them is conceptually acceptable ("close to" for very short distances, "near" for mid-length distances and "far" for points very separated from the LM).

The parametrizations used in the model are the ones represented in Fig 2. They were fitted in order to improve the interpretation of the descriptions in the performed experiments.

3.3 Virtual Objects

When the person describes the scene, the robot just knows about the position of the objects from the relations made in the description. For taking this circumstance into account we introduce the concept of *virtual objects* in contrast with the sensed ones, ie., the ones acquired from the sensors.

The position of these objects, created from the description, is defined by the spatial relations in which they are the TR. For example, in "there is a chair in front of me" the position of the chair is defined by "in front of me".

As the FOAs represent the distribution of a spatial relation applicability, it's natural to express the uncertainty in a virtual object position in function of the FOAs corresponding to the relations in which it was TR.

Concretely, we define the probability density function for the center of gravity of a virtual object 'TR' of being located at the point 'P' if its position was specified by the spatial relation 'rel' as:

$$f_{TR}(P|POV = Q, LM = R) = \frac{1}{n} rel(POV, LM, P)$$
$$n = \int_P rel(POV, LM, P) dP$$

being 'rel(POV,LM,P)' the FOA corresponding to the spatial relation 'rel' and 'P,Q,R' points of space.

Three things must be noted in the former expression. (1) there is a conditional dependence with the positions of the the POV and LM. This is natural, as those positions appear in the expression of the FOAs. (2) The integral in the normalization of the FOAs should be evaluated over the region corresponding

to the interior of the room, representing that the probability for a virtual object of being inside the room in which the description is performed must be 1. (3) It's normalized and it just can take non-negative values, hence fulfilling the conditions for being a density function.

If the POV or the LM are also virtual objects, we will need to extract the marginal distribution for the TR position by making use of the law of total probability:

$$f_{TR}(P) = \frac{1}{n} \iint_{Q,R} rel(POV, LM, P) f_{POV}(Q) f_{LM}(R) dR dQ$$

The former expression should be autonomously evaluated by the robot for any relation, POV and LM when interpreting a description. As a practical workaround, we opted for discretizing the probability distribution for the virtual objects position. In this way, $f_X(P)$ turns into $p(X = P)$ and the integral into a double summation.

The discretization was made by evaluating the FOAs over a grid covering the scene region (ie. the interior of the room).

The resulting expression for the marginal distribution is:

$$p(TR = P) = \frac{1}{n} \sum_Q \sum_R rel(POV, LM, P) p(POV = Q) p(LM = R)$$
$$n = \sum_P rel(POV, LM, P)$$

where 'P,Q,R' are points of the grid.

In the worst case (POV and LM virtual), when processing a spatial relation the FOA must be evaluated for each grid point being the POV and the LM also at any grid point. This makes the algorithm to be $O(n^3)$ with the number of grid points.

As an approximated solution to the former marginal distribution expression, we can take the assumption that the POV and the LM are located at their mean positions according to their own distributions. That is, to assume that all the terms in the summation in the marginal distribution expression are zero but the one in which $p(POV = P_{POV})$ and $p(LM = P_{LM})$, where P_{POV} and P_{LM} are the mean positions for the POV and the LM.

With this approximation the algorithm for processing a spatial relation turns to be $O(n)$ with the number of grid points. In the experiments the exact and the approximate solutions were tested in order to decide if the first one deserves its higher time complexity.

Regarding how theoretically appropriate is taking the former assumption, it's partially supported by the "gricean principle" which states that when an utterance is given from a speaker to a listener both of them expect the contextually most typical interpretation of the utterance [13]. In our case this principle can be translated as that the speaker will give a reference expecting the listener to "imagine" the objects in their more typical positions, ie. in the ones corresponding to the maximums of applicability for the relations which defined their positions.

Finally, we have to consider that the spatial relations included in the model just give information about the relative positions between the implied objects, ie. they say nothing about their size or orientation. In our model this implies that as all the information about the virtual objects comes from the description there will be a complete uncertainty in these magnitudes.

In the expression of the "proximity" field it appears the distance to the LM and its area. These parameters obviously depend on the concrete geometry of the LM which in turn depends on the orientation and size (jointly with the position and the type). Thereby, some assumptions must be taken. These will be that the virtual objects are of a "standard size" (see the architecture subsection) for the computation of the area and that they are point objects for the computation of the distance.

3.4 Composition Model

In a description, more than one relation can be provided for the same virtual object in order to better specify its position. That is, they can be referred more than one relation with the same TR. Therefore it's needed to determine how this spatial information will be composed.

Two relations with the same TR can be given with two purposes: (1) to delimit the region in which it can be located. This case corresponds to an intersection between the regions in which each of the relations is applicable. In logical terms it will be a conjunction (eg. "the ball is on the right on the table and close to the wall"). (2) to enlarge the region in which it can be located. This case corresponds to a union or disjunction (eg. "The ball is on the right of the table or on its left").

We will suppose that when two relations are referred for the same TR the intention is to concrete its position. Hence, in this case it will be performed a conjunction. In contrast, when the LM in a relation is ambiguous (eg. "near the wall" if there are more than one wall), a constructive composition (disjunction) will be performed for that relation being the LM each of the possible candidates (eg. each of the walls in the former example).

Regarding how these two compositions are performed, it can be remembered from the last paragraph that from a spatial relation in which the TR is a virtual object, it can be deduced a probability density function for its position ($f_{TR}(P)$). When two relations are given for the same TR we will have two different densities:

$$f^1_{TR}(P) = \frac{1}{n} \iint_{Q,R} rel_1(POV_1, LM_1, P) f_{POV_1}(R) f_{LM_1}(Q) dR dQ$$
$$f^2_{TR}(P) = \frac{1}{n} \iint_{Q,R} rel_2(POV_2, LM_2, P) f_{POV_2}(R) f_{LM_2}(Q) dR dQ$$

An intersection (or conjunction) will correspond to the joint probability density function, and a union (or disjunction) to the union of the two density functions.

Before concreting the expressions for the composition a reflection must be done about the independence of the composed densities. Each of them represents the probability for the same virtual object of being located at a certain point

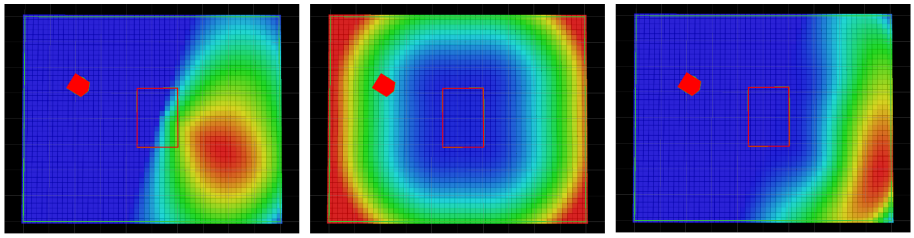

Fig. 3. Probabilities distributions for and object a) "behind the table", b) "far from the table" and c) and the conjunction of the two previous ones. Red color corresponds to the maximum value in the distribution and blue to the minimum. The table is the hollow rectangle and the POV the filled polygon.

of space given a different spatial relation. Or more precisely, they represent the probability for the relations which define the respective densities of being fulfilled when the object is located at a certain point. With the latter interpretation, and remembering the expressions of the FOAs, it can be stated that the fact of fulfilling one of the relations doesn't directly conditions the fulfilling of the other. Although there can be an indirect dependence. For example, the statement "the chair is on the left of the table" indirectly conditions the veracity of the statement "the chair is on the right of the table", but in terms of the definitions of the FOAs which represent both relations, there is no dependence.

Basing in the former argument, we will suppose the density functions corresponding to different spatial relations to be independent. This statement leads to the following expressions for the composition of two relations:

$$f^1_{TR}(P) \bigcap f^2_{TR}(P) = f^1_{TR}(P) \cdot f^2_{TR}(P);$$
$$f^1_{TR}(P) \bigcup f^2_{TR}(P) = f^1_{TR}(P) + f^2_{TR}(P) - f^1_{TR}(P) \bigcap f^2_{TR}(P);$$

In the discrete case (ie. discrete probability distributions) the expressions for the the composition are analogous to the latter. In Fig 3 is shown the distribution for a TR which is "behind the table" and "far from the table" (ie. the final distribution is the intersection of the former ones).

4 Collaborative scene mapping

In this section we overview the method used for building the scene maps described in the introduction and the architecture of the implemented system. The presented collaboration procedure tries to exploit the inherent capacities of each of the parts: the robot and the human. The robot can contribute with the accuracy of its sensors and its computational capacities for processing and composing information. On the other hand, the human have a unique facility in the segmentation and classification of objects. He also may possesses background knowledge and information about particularities or functionality of the objects in the environment to map.

In first place the person describes the scene using the contemplated set of spatial relations and the robot forms a mental map of it by interpreting those relations according to the presented model. This mental map provides information about the type of the objects in the scene and the region where they are most probably located. When sensing the scene, this information is used for sense guidance. That is, the robot looks for the actual objects in the regions corresponding to the covariance ellipses of the virtual objects in the mental map. Also, as it's been said, after sensing the scene the robot tries to match these virtual objects with the sensed ones, linking the semantic information associated to the former to the latter.

4.1 Architecture

The system has three main components: (1) a spatial relations library which, given a relation, a POV and a LM (sensed or virtual), evaluates the corresponding FOA over a 2D grid following the model described in the previous section. (2) A simple parser which translates the descriptions given by the human. (3) A Geometric Scene Description component (GSD), which keeps a list of the scene objects and actualize the information about them when new relations are processed. The objects can be added to the GSD in two ways, being perceived by the robot sensors (sensed objects) or introducing them in the description (virtual objects).

Each object in the list is an instance of a type of object (eg. chairs) which has associated a 3D mesh used by the spatial relations library in the computation of the FOAs. Each instance has a pose, a covariance, an a scale. The scale is set as the proportion between the dimensions of the actual object and the 3D mesh corresponding to its type. The virtual objects will be considered to be of a "standard size", ie. for them the scale will be set to one.

The virtual objects in the GSD have also a grid corresponding to their position probability distribution. From this distribution is extracted a mean position and a covariance matrix used in the matching of the robot mental and sensed maps.

4.2 Robot mental map: processing the descriptions

In the first part of the mapping process the person describes the spatial layout of the scene without the robot being there. The idea is that as he provides relations for the same TR he will be imposing additional constraints over its position which will reduce its covariance.

In order to ensure a correct interpretation of the description, the person must follow some rules. These are:

1. Before making any movements he must specify them by describing his target position in the same way as he would do with any other object, ie. using the same spatial relations (eg. "I move in front of the second table"). This ensures a correct position for the POV in the interpretation of the relations.

2. He must specify his intrinsic orientation when providing relations in which he is the LM (eg. "The table is on my left"). He will do that by telling where is he looking when referring the relation (eg. "If I look to the back wall, the table is on my left"). This ensures a correct definition of the intrinsic frame in these cases.
3. The objects involved in a spatial relation must be unequivocally specified. That is done by naming them with its type and the ordinal number corresponding to their order of appearance in the description (eg. "the second chair is on the right of the first table").

The robot mental map is formed by all the virtual objects in the GSD after processing the whole description.

4.3 Enriched map: scene check and map refinement

In the next step the robot senses the scene by using its 2D range laser. For this task the robot is supposed to be able to detect and classify the objects present in the scene (at least the ones introduced in the description) and to localize itself inside the scene (ie. in a map of an empty room with the known dimensions).

As it's been said, in order to ease this process the robot uses the formerly computed mental map for guiding the sensing. This is performed by following the next method: after coming into the room, the robot places itself in the middle of it and direct the laser to the regions corresponding to those in the mental map where the virtual objects are more probably located. That is, to those regions corresponding to the covariance regions of the virtual objects.

The sensed map will be formed by all the sensed objects in the GSD after the sensing process.

After sensing the scene, the robot performs the matching of the virtual objects with the sensed ones. To achieve an assignment (called objects fusion) the system looks for three requirements:

1. The sensed and virtual object are of the same type.
2. The probability for the virtual object of being at the sensed object position is greater than a certain threshold (according to its final probability distribution). For the experiments this threshold was set to the half of the maximum in the probability distribution.
3. For each relation in the description in which the virtual object acts as LM, there is a sensed object of the TR type with an applicability greater than a threshold when the relation is evaluated using the candidate sensed object as LM. This threshold was set to 0.5. This second check helps to prevent wrong fusions in complex scenes (when there are several objects of the same type).

If the three requisites fulfill the sensed object is taken as a candidate for the fusion. From all the candidates is chosen the one closest to the mean position of the virtual object. Once a fusion is accomplished, the semantic information associated with the virtual object is added to the sensed one and the former is

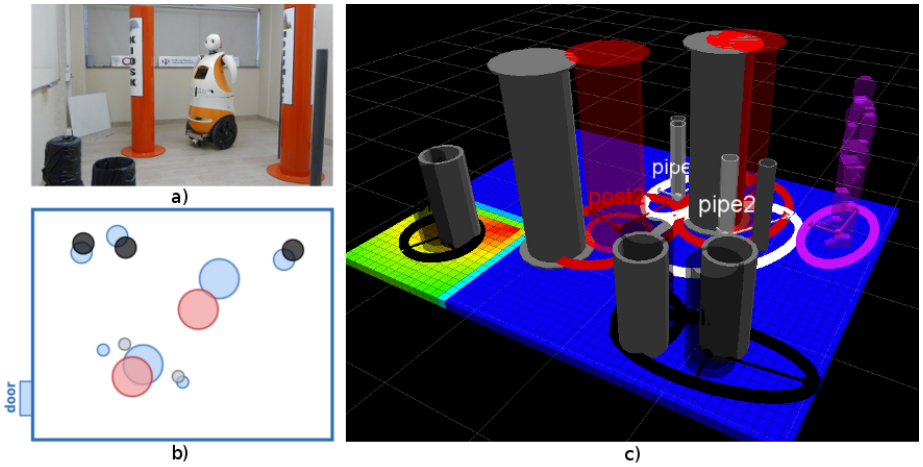

Fig. 4. a) Robot "Tibi" sensing the second scene. b) 2D view of the virtual map showed in the picture (c) contrasted with the ground truth (blue objects). c) Mental and sensed maps for the second scene used in the experiments. The sensed objects are the gray ones. The virtual ones has the color the person told in the description. They are visualized the covariance ellipses for the virtual objects.

deleted from the GSD. The final position for the object is the one corresponding to the sensed object.

If after trying to fuse all the virtual objects there are still any of them in the GSD the robot infers that the sensed scene doesn't correspond to the described one. As at the time being there is no mechanism to detect wrong fusions, if one virtual object is fused with a sensed object not corresponding to it, that will probably prevent the correct fusion of the rest of the objects. Anyways this never happened in the performed experiments.

The objects in the GSD after the fusion forms the enriched map. In our experiments we symbolize the semantic information given by the person with the color of the objects (as the robot is using a range laser which provides no color information). That is, with the "name" of the color of the objects as the person perceives them (eg. "red"). In the maps an RGB value is associated with each color just for visualization purposes.

5 Experiments

The conducted experiments were focused on testing the accuracy of the robot mental maps and the efficiency when matching them with the sensed ones. For that mean, we placed a set of objects inside a room and asked ten people among the researchers of the institute to describe the scene using the relations specified in section 3 and following the rules in subsection 4.2. The only information the robot had about the scene were the dimensions of the room, that there was a

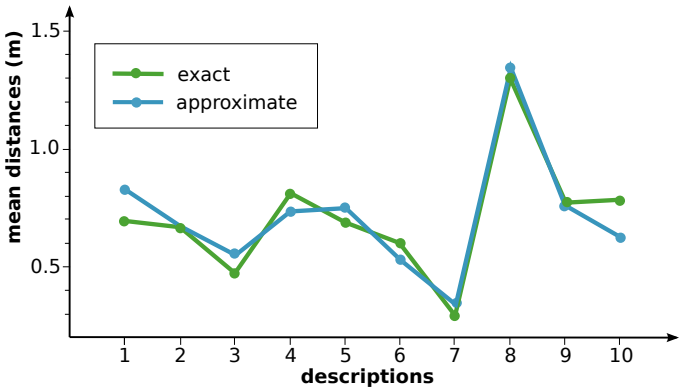

Fig. 5. Mean distances from the virtual objects to the ground truth in the ten descriptions processed for the scene2.

door and the location of the latter. The initial position of the person describing the scene was supposed to be "close to the door".

From each description the robot built a map by generating and composing each relation and afterward came into the scene, sensed it and performed the matching. The objects could be of three types: post, bin or pipe. Two scenes were tested. A simple one with one object of each type (scene1). A more complex one with several objects of each type (scene2).

As an example of the provided descriptions, next is shown the one from which was computed the most accurate mental map of the second scene (description 7 in Fig 5).The built mental map is shown in Fig 4.

"There is a red post in front of me. The post is near me. There is a white pipe behind the post. The pipe is close to the post. There is a black bin on my left. The bin is close to the left wall. There is a black bin on my left. The second bin is close to the left wall. The second bin is close to the first bin. The second bin is on the right of the first bin. There is a black bin close to the back wall. The third bin is close to the left wall. There is a white pipe near me. If I look to the third bin the second pipe is in front of me. There is a red post near the back wall. The second post is near the left wall. If I look to the third bin the second post is in front of me. The third bin is behind the second post."

The descriptions were processed using exact and approximate solutions when generating FOAs from virtual POV or LM (subsection 3.3).

The results were evaluated according to three parameters:

1. Mean distance from the objects in the robot mental map to the ground truth for all the objects in all descriptions ("distance" in table 1).
2. Mean covariance in the virtual objects position, expressed as the length of the radius of a circumference of the same area as the covariance ellipse, $r = \sqrt{area/\pi}$ ("covariance").
3. The percentage of times that the robot succeed to match the robot mental and sensed maps ("matches").

		distance (m)	covariance (m)	matches (%)
approximate approach	scene1	0.9 ± 0.2	0.6 ± 0.1	90
	scene2	0.6 ± 0.2	0.6 ± 0.1	80
exact approach	scene1	0.9 ± 0.2	0.73 ± 0.09	100
	scene2	0.6 ± 0.1	0.73 ± 0.08	90

Table 1. Mean distances, mean covariances and percentage of success in the matching of the virtual and sensed maps for the two scenes using the two approaches for generating the FOAs.

The results achieved are summarized in Table 1. There is no data about the performance in the sensing process as the robot achieved to detect all the objects in every case. This success is no doubt due to the simplicity of the geometry of the chosen object. This choice was made in order to isolate the assessment of mental map and fusion processes from possible errors in the object detection.

The first thing we realized after conducting the experiments is that not all the people have the same "descriptive skills" (or the same understanding of space). As it can be seen in Fig 5 the results are very description dependent, going the mean distances from 0.29 m to 1.43 m (though almost all of them stay in the interval 0.4 - 1.0).

For using the mental map as a tool by itself the interesting numbers are those corresponding to the mean distances and covariances. But for using it as an intermediate step in the mapping process (as done in this article), the most relevant data is the percentage of matches. The accuracy of the mental map is important for guiding the sense and fusing the objects. But if after all the robot achieve to match the two maps, the final object positions will be the sensed ones. In this sense, the success percentage is around 90 %.

Regarding the options in the FOAs generation, the mean distances to the ground truth were the same when using the exact and approximate solutions in the generation of FOAs from virtual objects. Though the former option tend to generate larger covariances, which helps to improve the matches percentage. On the other hand, the larger covariances could lead to wrong fusions. But this didn't happen in the conducted experiments.

6 Conclusions

We have presented a model which let a robot to build the map of a scene in collaboration with a person. To that end, the person must describe to the robot the layout of objects in the scene using qualitative spatial relations being able to include semantic information about the objects in the description. An extension of the model which let to express more precise information about distances and orientations would no doubt improve the accuracy of this map. Although we plan to research in that direction in the future, in the present work we preferred to limit ourselves to a more ordinary language which people usually use when expressing spatial knowledge.

The mental map built from the description provides information about the type of the objects in the scene and the region where they are most probably located. For the time being just the position is used for guiding the sensors. An interesting expansion for a future work would be to study the joint use of the type and position information to improve the segmentation process. Also taking into account possible occlusions in the sense guidance would make the method more robust.

Finally, the mental and sensed maps are fused, being trespassed the semantic information in the virtual objects to the sensed ones.

Acknowledgments. This work has been partially funded by Spanish Ministry of Economy and Competitiveness under project TaskCoop DPI2010-17112.

References

1. R. Moratz, K. Fischer, and T. Tenbrink, "Cognitive modeling of spatial reference for human-robot interaction," *International Journal on Artificial Intelligence Tools (IJAIT)*, vol. 10, no. 4, pp. 589–611, 2001.
2. E. Stopp, K.-P. Gapp, G. Herzog, T. Laengle, and T. C. Lueth, "Utilizing spatial relations for natural language access to an autonomous mobile robot," 1994.
3. K.-P. Gapp, "Basic meanings of spatial relations: Computation and evaluation in 3d space," 1994.
4. J. Kelleher and J. van Genabith, "A computational model of the referential semantics of projective prepositions," in *Syntax and Semantics of Prepositions* (P. Saint-Dizier, ed.), vol. 29 of *Text, Speech and Language Technology*, pp. 211–228, Springer Netherlands, 2006.
5. N. Mavridis and D. Roy, "Grounded situation models for robots: Bridging language, perception, and action," in *In Proceedings of the AAAI-O5 workshop*, pp. 32–39, 2005.
6. R. Coyne and R. Sproat, "Wordseye: an automatic text-to-scene conversion system," in *SIGGRAPH*, pp. 487–496, 2001.
7. R. Smith, M. Self, and P. Cheeseman, "Estimating uncertain spatial relationships in robotics," in *UAI*, pp. 435–461, 1986.
8. G. Dissanayake, P. M. Newman, S. Clark, H. F. Durrant-Whyte, and M. Csorba, "A solution to the simultaneous localization and map building (slam) problem.," *IEEE Transactions on Robotics*, vol. 17, no. 3, pp. 229–241, 2001.
9. S. Thrun, W. Burgard, D. Fox, H. Hexmoor, and M. Mataric, "A probabilistic approach to concurrent mapping and localization for mobile robots," in *Machine Learning*, pp. 29–53, 1998.
10. A. Aydemir, K. Sjöö, J. Folkesson, and P. Jensfelt, "Search in the real world: Active visual object search based on spatial relations," in *to appear in Proc. of the IEEE International Conference on Robotics and Automation (ICRA'11)*, 2011.
11. G. Retz-Schmidt, "Various Views on Spatial Prepositions," *AI Magazine*, vol. 9, no. 2, pp. 95–105, 1988.
12. G. A. Miller and P. N. Johnson-Laird, *Language and perception*. Harvard, 1976.
13. H. Grice, "Logic and conversation," in *Syntax and Semantics: Speech acts* (P. Cole and J. L. Morgan, eds.), pp. 41–58, Academic Press, 1975.