2012 IEEE International Conference on Robotics and Automation
RiverCentre, Saint Paul, Minnesota, USA
May 14-18, 2012

Semantic map segmentation using function-based energy maximization

Kristoffer Sjöö

Abstract— **This work describes the automatic segmentation of 2-dimensional indoor maps into semantic units along lines of spatial function, such as connectivity or objects used for certain tasks. Using a conceptually simple and readily extensible energy maximization framework, segmentations similar to what a human might produce are demonstrated on several real-world datasets.**

In addition, it is shown how the system can perform reference resolution by adding corresponding potentials to the energy function, yielding a segmentation that responds to the context of the spatial reference.

I. Introduction

In the field of mobile robotics, one of the main goals is the integration of robots into the daily lives of humans, aiding us by carrying out tasks for us at home, at the workplace or in outdoor environments. There are many challenges still to overcome before this vision can become reality, however. One of them is that in order to make sure the robots do the right thing, and in the right place, means of intuitive communication between man and machine are needed – in particular, communication concerning their mutual environment.

Robots will need to parse humans' statements and requests and to formulate their own questions and reports in return, using expressions that can be understood by both human and machine. The treatment of such expressions are the subject of this paper; in particular, those describing different parts of space and which an agent might use for navigation or to carry out specific tasks.

The fundamental assumption adopted herein is that *functional* properties are key to dividing up and referring to the world, see Tversky [1]. An indoor environment is constructed intentionally with different functions compartmentalized: this room for eating, this one for sleeping, this for working; and the words we use to refer to those spaces likewise pertain to those functional distinctions. Consequently, this paper attempts to use functional aspects of space to achieve a subdivision and labeling of 2-D maps that corresponds well to human intuitions.

A. Related work

There has been a great deal of work related to the subdivision of maps into discrete units, in many different contexts. One common approach to discretizing space is by using Voronoi diagrams [2]. Another is partitioning it on the basis of the navigational actions it affords, such as in Kuipers et al. [3]. Milford et al. [4] accomplish a similar structuring using neural networks. Brunskill et al. [5] divide a 2-D map into units based on spectral clustering. Pronobis et al. [6] discuss the general problem of partitioning the world into distinct "places" based on perceptual distinctiveness and spatial relationships.

Given a discretization, the next step is to label the units in some relevant way. Diosi et al. [7] and Milford et al. [8] impose labels externally, through a user that the robot is talking to at different locations. Mozos et al. [9] classify regions using metric features, while Vasudevan et al. [10] utilize spatial relations between objects. A graph-based approach is taken by Friedman et al. [2], by performing place classification based on potentials defined on nodes in a graph, with arity up to 4, making it similar to the framework used in this paper although with a model that is more local, and learned as opposed to specified by functional criteria.

Work that examines the functional properties of space include Kuhn [11], who discusses the problem in general terms on an abstract level; and at the other end of the spectrum Dornhege and Kleiner [12], in which parts of a map are classified according to whether they afford a robot's moving through them, though not using human or linguistic concepts. Also related is Fedrizzi et al. [13] where specific places are defined on the basis of a robot's ability to manipulate objects there. Lastly, a debt is owed to Coventry and Garrod [14] who have pioneered the investigation of functional aspects of spatial relations in language.

This work is also concerned with mapping linguistic expressions to portions of space, although in a limited way. Related work has been done e.g. by Kollar et al. [15], who also use an energy optimization method to determine referents for an expression, and Mandel et al. [16], who choose the referent from among Voronoi nodes using fuzzy functions. Both of the above deal with route descriptions, and not with labeling or segmenting maps. Zender et al. [17] also deal with determining spatial entities referred to by a speaker, by finding the lowest common context in a hierarchy. There too, the set of potential referents is assumed to be given.

B. Contributions

In this paper, a method is presented by which separate, basic, common-sense criteria of a functional nature, such as may be found in a dictionary, can be combined in a single energy maximization and yield an intuitively reasonable subdivision and labeling of a map. Furthermore it is demonstrated how the same energy maximization can be used to find the referents of a linguistic expression, through

The author is with the Centre for Autonomous Systems at the Royal Institute of Technology (KTH), Stockholm, Sweden. This work was supported by the SSF through its Centre for Autonomous Systems (CAS), and by the EU FP7 project CogX and the Swedish Research Council, contract 621-2006-4520

978-1-4673-1405-3/12/$31.00 ©2012 IEEE

translating it into an energy potential in a straightforward way.

C. Structure

This paper is structured as follows: in Section II the reasoning behind using functionality as the basis for spatial segmentation is explained; Section III outlines the energy maximization framework and the solution algorithm. Experiments on various datasets are described in Section IV and Section V presents their outcomes. Section VI summarizes the paper and discusses future work.

II. Functional properties of space

The fundamental concept of this work is the idea that *function* is key to the way humans understand space, and thus also key to any successful robotic representation intended to interact with humans and human-designed environments.

As an example, consider the concept of a kitchen. For a robot to be able to follow orders from humans in a home environment, it will be necessary for it to understand what the word means. A typical approach is to have a human "tag" points in space with the fact that a region is a kitchen [8]. The tag might be attached to a single point, or a region, segmented out by some independent process – such as using laser scans to detect doorways and grouping places on each side of the doorway into different regions [9]. The tagging might be replaced by using machine learning to train models of different regions' appearance.

However, what makes a kitchen a kitchen at a fundamental level is not its appearance, nor a person calling it "kitchen", but the fact that it is used to prepare (and store and consume) food. An appearance-based model might fail if the kitchen is of a novel layout or unfamiliar design, and an algorithm that uses doorways as cues might fail for a studio apartment, where there is no such clear boundary between "kitchen" and "living room". But if a robot can be made to recognize the potential for the *function* of a kitchen, e.g. food preparation, this will improve its ability to generalize and its capacity to communicate effectively with humans.

The semantic labels humans use for space may also vary depending on context. In the case of the aforementioned studio apartment, sometimes "kitchen" will be used to refer to the part of it that houses the sink and oven, while sometimes "room" will be used of the entire room including the kitchen area. This context-sensitivity is an additional necessary feature of a robot's system for spatial understanding.

In the following section, a framework is presented that attempts to incorporate both functional segmentation and context-sensitive reference resolution.

III. Framework for functional labeling of space

The problem is the following: given a 2-dimensional map of an environment, including a pre-segmentation of it into a number of small units, "places", find a combination of clusters of places and labels for these clusters such that each label appropriately describes the functional features of the associated place cluster. The map that is given may contain various additional information, such as occupancy data, paths existing between places, and objects associated with places.

A. Basic definitions

The set of all places in the map is termed $\mathcal{P}$. A *region* $\mathcal{R}$ is a set of places: $\mathcal{R} = \{p \in \mathcal{P}\}$.

A *label* L is a linguistic symbol corresponding to a region's perceived functional purpose. Labels used in this paper are "room", "corridor", "entrance", "kitchen", "office".

A *relational label* is a label that additionally refers to another region by its definition. Of the above, "entrance" is relational; an entrance is always an entrance *to* something.

A *labeling* $\mathcal{L}$ is a set of 3-tuples, each consisting of a region $\mathcal{R}_i$, a label for that region L_i, and a relational index k_i indicating which other region the label relates to if it is relational. The regions are subject to the constraint that each place in $\mathcal{P}$ is in exactly one region:

$$\mathcal{L} = \{\langle \mathcal{R}_i, L_i, k_i \rangle\}, \begin{cases} \bigcup \mathcal{R}_i = \mathcal{P} \\ \mathcal{R}_i \cap \mathcal{R}_j = \emptyset, \forall i \neq j \\ 1 \leq k_i \leq |\mathcal{L}| \end{cases}$$

The number of regions (and hence tuples) is variable (but cannot exceed $|\mathcal{P}|$). Note that it is quite possible that many regions in a labeling are assigned the same label; hence a labeling can differentiate between e.g. one corridor and another in the same map.

B. Energy function

Every 3-tuple in a labeling has an associated energy, representing how well that particular label describes that particular group of places. A higher energy means a better fit.

$$E(\langle \mathcal{R}_i, L_i, k_i \rangle) = f(\mathcal{R}_i, L_i, k_i, \mathcal{L}) \in [0, |\mathcal{R}_i|] \qquad (1)$$

Note that the energy depends on the entire labeling, in general. (It also depends on the map; however, that is considered a constant here and left out of the notation.) Because the number and size of regions can vary arbitrarily, in order to avoid any bias for large or small regions the label energies should be proportional to the size of the region, other things being equal, and the average energy per place be within $[0, 1]$ – so that the energy for each region is bounded by $|\mathcal{R}_i|$.

The energy function is the sum of the energies of each region in the labeling:

$$E(\mathcal{L}) = \sum_i E\left(\langle \mathcal{R}_i, L_i, k_i \rangle\right) \qquad (2)$$

The energies assigned to a label for a given region should correspond to the degree to which that region possesses the functional features that define that label. Features are combined in a weighted sum, where the weights may be negative:

$$\begin{aligned} & E(\langle \mathcal{R}_i, L_i, k_i \rangle) = \\ = \quad & \max\left\{0, \sum_k w_l(L_i)\phi_l(\langle \mathcal{R}_i, L_i, k_i \rangle)\right\} \end{aligned} \qquad (3)$$

where ϕ_l is the value of the l^{th} feature, and $w_l(L_i)$ is the weight assigned that feature for label L_i. For example, the food preparation feature has a positive weight for the kitchen label. The label energy will be in $[0, |\mathcal{R}_i|]$.

C. Labels

Below is a list of the labels used for the experiments in this paper, followed by the formulation of the functional features used. The description is headed by a commonsense definition taken from The Oxford English Dictionary (OED) Online [18].

- Room: *"A compartment within a building enclosed by walls or partitions, floor and ceiling, esp. (freq. with distinguishing word) one set aside for a specified purpose; (with possessive) a person's private chamber or office within a house, workplace, etc. [...]"* The functional aspects focused on in the following are the *enclosure* of a room and the *specified purpose* associated with it (the ownership angle is beyond the scope of this paper as it entails social considerations besides purely spatial ones). Enclosure affords a room protection from outside disturbances and influences, and helps an agent form a definite boundary when speaking or thinking about a region.

  The room also supports some purpose or task for agents who are in it, which it will typically do through some object or set of objects located in the room, with which an agent interacts. If a portion of the room is obscured from the rest then any objects in that portion will not contribute to the perceived function of the room; thus regions which are mutually visible will tend to belong to the same room and vice versa. This is encapsulated in a feature that will be referred to as *perceptual convexity*, meaning that each place in the room is visible from the others.
- Corridor: *"A main passage in a large building, upon which in its course many apartments open."* Here, the functional aspect implied is *connecting*, i.e. a corridor serves as a main route of communication between different parts of the map.
- Kitchen: *"That room or part of a house in which food is cooked; a place fitted with the apparatus for cooking."* The focus is here on the function of *cooking*, as supported by specific objects. Having room-like features is also of relevance, although not stated as an absolute requirement.
- Office: *"A room, set of rooms, or building used as a place of business for non-manual work; a room or department for clerical or administrative work. [...]"* In this case the function is that of *work*, specifically non-manual work. Again, room attributes appear as non-essential aspects of the term.
- Entrance: *"That by which anything is entered, whether open or closed; a door, gate, avenue, passage; the mouth (of a river). Also, the point at which anything enters or is entered."* Evidently *entering* is the key aspect here.

D. Features

The above labels make use of the following set of function-related features:

1) Enclosed: The "enclosed" feature rewards labelings where room-labeled regions are compact and largely delineated by walls. It is defined as a function of the region's total outer boundary length, B_{total}, and the length $B_{external}$ of any boundaries it shares with other regions:

$$\phi_{encl} = |\mathcal{R}| \left(1 - \frac{B_{external}(\mathcal{R})}{B_{total}(\mathcal{R})}\right) \tag{4}$$

The $|\mathcal{R}|$ factor ensures the energy grows as the size of the region.

2) Perceptually convex: The measure of perceptual convexity within a region is

$$\phi_{perc} = \frac{\sum_{\{p,p'\} \in \mathcal{R} \times \mathcal{R}} Vis(p, p')}{|\mathcal{R}| - 1} \tag{5}$$

where

$$Vis(p, p') = \begin{cases} 1, & \text{if } p \text{ and } p' \text{ are visible from each other} \\ 0, & \text{otherwise (and where } p = p') \end{cases}$$

Again, the $|\mathcal{R}| - 1$ term is in order to normalize the energy to the order of the size of the region.

3) Connecting: The connecting function of corridors is evaluated in the following way: For each place in the prospective corridor, count the number of paths in the map that pass through that place, and sum this number over all places in the region. "Path" here means the shortest path between any pair of places in the map as a whole. This way, places that are crossed by many paths in the map contribute strongly to the connecting function of a region, while "dead ends" do not contribute at all. The feature can be expressed:

$$\phi_{conn} = \sum_{\substack{p \in \mathcal{R} \\ \langle p^{from}, p^{to} \rangle \in \mathcal{P} \times \mathcal{P}}} \frac{C(p, p^{from}, p^{to})}{C_{max}} \tag{6}$$

where

$$C(p, p^{from}, p^{to}) = \begin{cases} 1, & \text{if } p \neq p^{from},\ p \neq p^{to} \\ & \text{and } p \text{ is on the shortest} \\ & \text{path between } p^{from} \text{ and } p^{to} \\ 0, & \text{otherwise} \end{cases}$$

C_{max} is a normalizing constant equal to the highest value, for any single p, of $\sum_{\langle p^{from}, p^{to} \rangle} C(p, p^{from}, p^{to})$. Note that this formulation makes ϕ_{conn} independent of the labeling, and so it does not need to be recalculated during the energy maximization process.

4) Entering: The entering feature is similarly defined to the connecting feature, but focuses on a specific "entered" region, i.e. the region specified by the relational index k_i. Only paths leading from elsewhere in the map and ending inside that region are counted, and each path only counts once for the whole region (not once per place it passes):

$$\phi_{ent,k_i} = \sum_{\substack{p^{to} \in \mathcal{R}_{k_i} \\ p^{from} \in \mathcal{P} \setminus \mathcal{R}_i}} \frac{C(\mathcal{R}_i, p^{from}, p^{to})}{|\mathcal{R}_{k_i}||\mathcal{P} \setminus \mathcal{R}_i|} \tag{7}$$

where $C(\mathcal{R}_i, p^{from}, p^{to})$ is 1 if *any* place in $\mathcal{R}_i$ is on the shortest path between p^{from} and p^{to}, and 0 otherwise.

5) Food-preparing: The potential of food preparation is here modeled as a function of the distance to objects needed for the task. Two objects are taken as determinants: "refrigerator" and "stove"[1] The value falls off as a sigmoid with the navigation distance (not the straight-line distance):

$$\phi_{food} = \sum_{p \in \mathcal{R}} \left(\alpha \frac{1+C}{e^{d_1(p)/B} + C} + \beta \frac{1+C}{e^{d_2(p)/B} + C} \right) \quad (8)$$

where B and C are constants determining the shape of the sigmoid, and the d_1 is whichever distance (stove or refrigerator) is smaller, d_2 the larger. This formulation allows a non-zero value even if one object is missing entirely.

6) Working: The working feature is treated analogously to the food-preparing feature, except that there is only one object, "desk" and so only one corresponding term in Eq. 8.

7) Penalty features: As was noted above, the "office" and "kitchen" labels may *optionally* have the features of a room as well. However, merely adding the room-like features (Enclosed and Perceptually convex) with positive weights to e.g. the office label will have the undesirable result that a region that is very room-like will, as a result, seem substantially office-like even if it has *no* Working feature whatsoever.

Therefore two "penalty features" are introduced: "Not enclosed" and "Not perceptually convex." Whereas the ϕ encode "goodness of fit," these penalty features represent "mismatch" and are simply $|\mathcal{R}| - \phi_{encl}$ and $|\mathcal{R}| - \phi_{perc}$, respectively (recall that $\phi \in [0, |\mathcal{R}|]$). By adding these penalty features with a negative weight, regions that are not room-like merely incur a limited penalty to their energy, while room-like regions do not receive any undue "bonus" to their "office" or "kitchen" energy.

E. Label energy functions

The weights in Eq. 3 are values between -1 and 1; the functional criteria defining a label have weights that sum to 1, and functions that conflict with them have weights below zero. If the label's energy sums to less than 0, it is set to 0. Table I shows the weights used in the following experiments.

In addition a constant energy penalty is added for each region, depending on label; below, all labels use a penalty of 0 except the corridor label, for which it is 0.1 per region – this small bias causes corridors to preferentially cluster many places into a single corridor rather than several smaller ones (otherwise the total energy is indifferent to size).

The values below were selected manually, though no fine-tuning was performed. The weights would be a suitable object for learning in future work.

[1]This should only be regarded as an illustration; which objects support the "food-preparing" and "working" functionalities, and others, should properly be grounded in either the robot's or humans' actual food-preparing activities and the requirements thereof; this is future work.

	Room	Corr.	Kitchen	Off.	Entr.
Enclosed	3/4				
Perc. convex	1/4				
Not enclosed			−3/8	−3/8	
Not perc. convex			−1/8	−1/8	
Connecting	-1	1	-1	-1	-1/5
Entering					1
Food		-1	1	-1	-1/5
Working		-1	-1	1	-1/5

TABLE I: Label feature weights used in segmentation experiments

F. Referring expression matching

Maximizing the energy described above serves to produce a context-less labeling of the map. In the following it is explained how a spatial referring expression, such as "the room next to the corridor", can be matched to a part of the map using the same framework.

A *description* $\mathcal{D}$ consists of a set of *attributes* and an n-tuple of regions taken from a labeling, each called an *operand*. n is called the arity of the description. Attributes are similar to labels, but may be defined on more than one region. Each attribute is associated with some subset of the descriptions' n-tuple.

Example: A description of arity 2 might have 3 attributes:

1) Region 1 should be labeled "Corridor" (unary)
2) Region 1 and region 2 should be neighbors (binary)
3) Region 2 should be a room (unary)

This description encodes: "find a room that is next to a Corridor".

Attributes each evaluate to a number $a_i \in [0, 1]$, and their geometric mean is taken as the "fit" of the description:

$$F(\mathcal{D}) = \sqrt[n]{a_1 \dots a_n} \in [0, 1] \quad (9)$$

The energy of the description is the product of its fit and the energy of the corresponding labeling:

$$E(\mathcal{D}) = \gamma F(\mathcal{D}) E(\mathcal{L}) \quad (10)$$

where γ determines how strongly the description influences the labeling. This energy is added to that of the labeling itself, and when this sum is maximized it will tend to assign the n-tuple to regions from the labeling which possess all the attributes – which may involve influencing the labeling such that there exists a match, e.g. by reinterpreting two otherwise separate rooms as a single large room. This effect is desirable, because the description implicitly injects information that the unbiased labeling does not have access to about e.g. how a human user conceptualizes different parts of the map. γ will in general depend on the application and the linguistic context; $\gamma = 0.1$ is used in this paper.

Attributes implemented for the experiments below are:

- Operand region A should have a specific label
- Operand region A should contain a specific place p^*
- Operand region A should be large
- Operand region A should be located toward a given direction in the map

- Operand region A should be located in a given drection relative to operand region B

IV. Experiments

This section describes experiments done using the above framework, operating on three grid maps, based on publicly available datasets obtained from the Robotics Data Set Repository (Radish) [19]: FR079, Intel and SDR (see Figure 1). The FR07 and Intel maps were obtained from Burgard et al. [20]. In order to obtain the initial oversegmentation of places $\mathcal{P}$ that the framework needs, a set of nodes and connections were added manually in the manner of an exploring robot to produce a graph similar to e.g. Mozos et al. [9]. Each free grid cell was then assigned to the closest (via free space) node, forming a place and permitting the computation of border lengths (see Sec. III-D). Objects were also assigned manually to places in two of the three maps, for illustrative purposes. The SDR map was left without objects.

A. Energy maximization

The high-level features making up the energy function make it problematic for standard graphical solving methods. For the purposes of this paper a stochastic method, simulated annealing, was found to provide adequate optimization. Simulated annealing works by taking random moves, and may move against the energy gradient in order to escape local minima, but does so at an ever-decreasing probability as time passes; see Algorithm 1. ($E(\mathcal{D})$ is substituted for $E(\mathcal{L})$ if matching a description.)

Algorithm 1 Energy maximization procedure

$T := T_{start}$
while $T > T_{end}$ **do**
  $\mathcal{L}_{new} :=$ `perturb`$(\mathcal{L}_{cur})$
  if $E(\mathcal{L}_{new}) > E(\mathcal{L}_{cur})$ **then**
    $p_{accept} := 1$
  else
    $p_{accept} := \exp\left(\frac{E(\mathcal{L}_{new}) - E(\mathcal{L}_{cur})}{T}\right)$
  end if
  if rand() $< p_{accept}$ **then**
    $\mathcal{L}_{cur} := \mathcal{L}_{new}$
  end if
  $T := T \cdot \kappa$
end while

All experiments used $T_{start} = 2$ and $T_{end} = 0.001$. The cooling-down rate, κ was set to 0.9998, leading to a step count of circa 40 000.

The `perturb` function changes the labeling using one of the following moves, picked at random:

1) *Transfer*: A donor region is picked at random, and a receiver region is picked from among the donor's neighbors. Places are transferred from the donor to the receiver until a random trigger stops it, or that entire connected component is transferred.
2) *Split*: A seed place is picked at random from the map, and another seed is picked from the neighbors of that place within the same region. The two seeds then grow competitively within the region, until a random trigger stops the process or that entire connected component is covered. Finally one of the grown seeds is picked at random to generate a new region with a random label.
3) *Relabel*: A random region is picked and given a random new label.
4) *Reassign index*: The relational index k_i of a relational label is set to a new random region
5) *Reassign description*: If a description is being used, change one of its operands to a new random region

Note that nothing in these rules keeps a region from becoming disconnected in the process. Maintaining a region's integrity comes out of the energy maximization.

After each `perturb` move above (except #5), additionally the description – if one is in use – is locally optimized by taking each of the regions that was affected by the change, and trying it in the place of each current operand in turn, to see if the description's value is improved by switching. This is done before p_{accept} is computed, and permits the description to effectively steer the labeling toward an optimum for both description and labels.

V. Results

Figure 1 shows the result of a context-less segmentation of the three maps. For the most part, the result accords with what a human might come up with. Some corridors in the upper half of the SDR map are mislabeled as rooms, probably because the many loops make for many alternative paths that "dilute" the connected property compared to the southern corridor. This might be remedied by normalizing that property more locally.

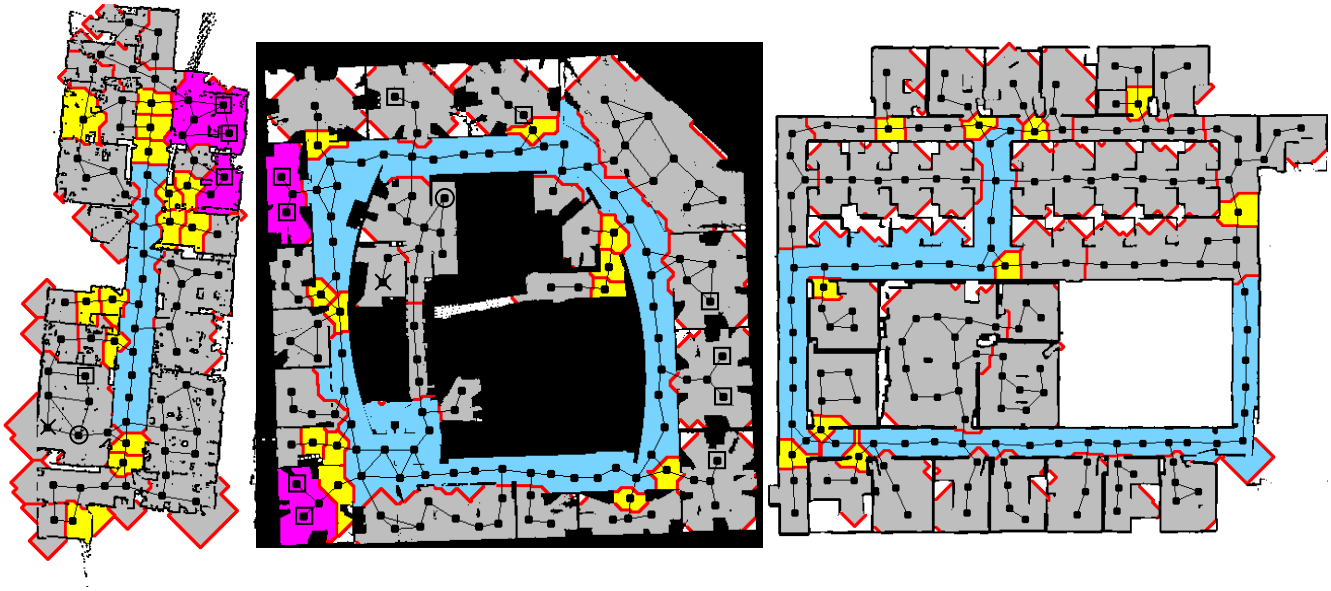
Note that this segmentation comes about purely from commonsense functional semantics, without the training of perceptual models, heuristics such as detected doorways or explicit tagging by humans.

Some of the kitchen- or office-related objects do not give rise to "kitchen" or "office" labels. It may be "cheaper", with the weights given, to simply assign "room" (which cannot be regarded as "incorrect" either). However, if *context requires* the presence of a specific label, this can tip the balance and assign "kitchen" or "office" to those same regions – se below.

A. Description resolution

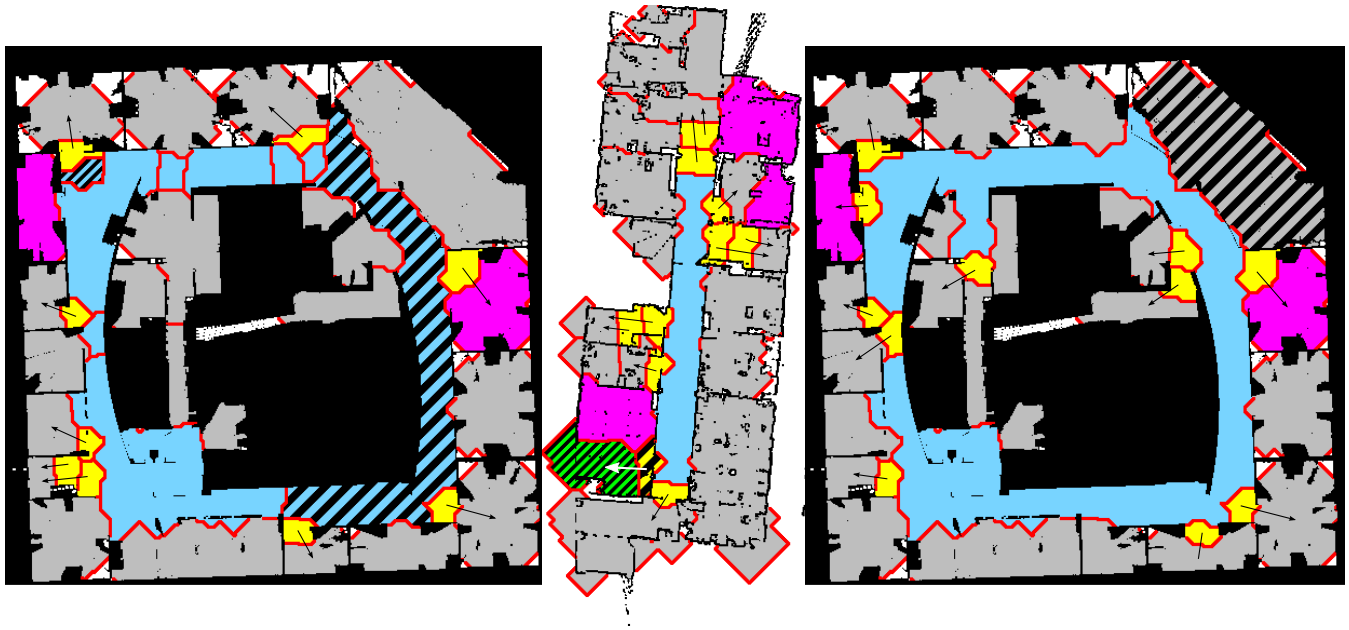
The following are some examples of reference resolution performed on the maps as described in Sec. III-F. They demonstrate that the functional framework can provide both flexibility and simplicity to spatial reference resolution. The labelings are shown in Figures 2 and 3.

1) Fig. 2a: "The eastern corridor" (Operand A: Labeled "corridor"; operand B: Labeled "corridor", located east of A). The expression implies there is at least one other corridor that is less easterly.
2) Fig. 2b: "Entrance to a kitchen" (Operand A: labeled "entrance", relational index must point to B; Operand



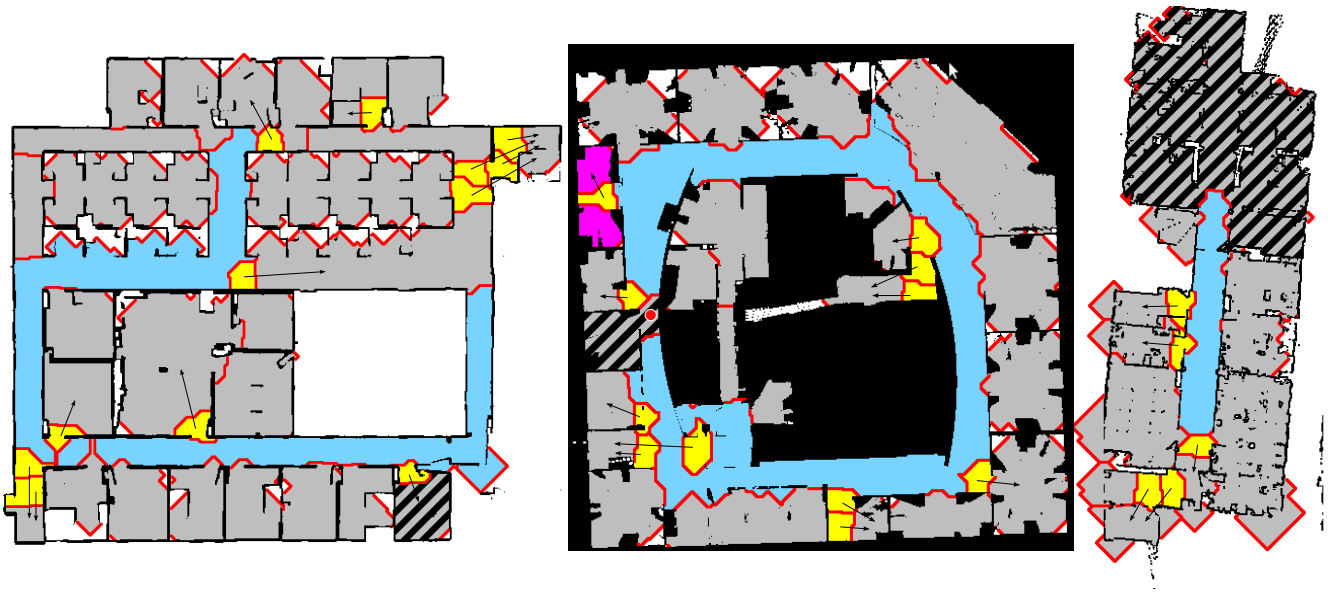
(a) FR079 map (b) Intel map (c) SDR map

Fig. 1: Labeling of regions. Grey signifies rooms, blue corridors, purple offices, and yellow entrances. Red lines delimit regions. Nodes used to create the places are also shown, with connectivity. Object annotations are shown as well: A circle represents a refrigerator object; an $\times$, a stove; a black square, a desk.



(a) "The eastern corridor" (b) "Entrance to kitchen" (c) "A big room"

Fig. 2: Fitting descriptions to map. Diagonal stripes indicate the primary operand of the description, horizontal ones the secondary when applicable. Small arrows indicate the relational indices of entrance regions.



(a) "A south-easterly room" (b) "The room at place $< p^* >$" (c) "A big room"

Fig. 3: Fitting descriptions to map, continued.

B: labeled "kitchen"). Note how the description enforces the creation of an entrance as well as the subdivision of the room (cf. Fig. 1a) into "kitchen" and "office."

3) Fig. 2c: "A big room" (Operand A: Labeled "room", large size).
4) Fig. 3a: "A south-easterly room" (Operand A: Labeled "room," south-east in the map). Here, no contrast set is used (the article "a" does not imply any) but instead it is the absolute position in the map that counts.
5) Fig. 3b: "The room at place $< p^* >$" (Operand A: labeled "room", contains p^*). Although not part of a context-less labeling (Fig. 1b), the best fit was found through extending the room into the corridor.

An example of a failed resolution is displayed in Figure 3c: "The big room". Here the search got stuck in a local maximum, where the potential seeking to increase the size of the referred room has caused too many separate rooms to merge.

VI. Conclusions

This paper has shown how a conceptually very simple – and, consequently, flexible – energy maximization approach can be used to perform segmentation of 2-D maps into units, using features taken from the functional aspects that form the core of spatial semantics. The resulting clusters correspond well to human intuitions. Additionally, it is shown how the framework can use the same mechanism to find matches for referring expressions, even adjusting the segmentation to accommodate the context implicit in those expressions.

A. Future work

The small set of different labels used in this work represents a prototype, showing that space can be segmented on functional grounds and how it might be done. In future work the set of features and labels should be expanded; in addition, the features need to be more realistically grounded in the robot's sensing and acting repertoire. Here, machine learning methods should be applied; the same goes for the different weights used. Ideally, the contextual resolution of expressions should be performed in a feedback loop during e.g. dialogue, which would proceed until the solution made sense in the context. Experiments on a real robot will also require objects to be detected using vision or other means.

The simulated annealing method used for solving the energy in this paper leaves much to be desired in terms of efficiency. It might be worthwhile to explore other approaches, such as conditional random fields; however, because of the general nature of the energies used few simplifying assumptions can be made by any algorithm.

B. Acknowledgements

The data sets used to create the maps used in this paper were obtained from the Robotics Data Set Repository (Radish) [19]. Thanks go to Cyril Stachniss, Andrew Howard and Dirk Hähner for providing this data. The FR07 and Intel maps were obtained from Burgard et al. [20].

References

[1] B. Tversky, "Structures of mental spaces: How people think about space," *Environment and Behavior*, vol. 35, pp. 66–80, 2003.

[2] S. Friedman, H. Pasula, and D. Fox, "Voronoi random fields: Extracting the topological structure of indoor

environments via place labeling," in *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, Hyderabad, India, January 2007.

[3] B. J. Kuipers, J. Modayil, P. Beeson, M. MacMahon, and F. Savelli, "Local metrical and global topological maps in the hybrid spatial semantic hierarchy," 2004.

[4] M. J. Milford, G. F. Wyeth, and D. Prasser, "RatSLAM: a hippocampal model for simultaneous localization and mapping," in *Proc. of the 2004 IEEE International Conference on Robotics and Automation (ICRA)*, New Orleans, LA, USA, April 2004.

[5] E. Brunskill, T. Kollar, and N. Roy, "Topological mapping using spectral clustering and classification," in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, San Diego, CA, October 2007.

[6] A. Pronobis, K. Sjöö, A. Aydemir, A. N. Bishop, and P. Jensfelt, "A framework for robust cognitive spatial mapping," in *Proc. of the 14th IEEE International Conference on Advanced Robotics (ICAR)*, Munich, Germany, June 2009.

[7] A. Diosi, G. Taylor, and L. Kleeman, "Interactive SLAM using laser and advanced sonar," in *Proc. of the 2005 IEEE International Conference on Robotics and Automation (ICRA)*, Barcelona, Spain, April 2005.

[8] M. J. Milford, R. Schulz, D. Prasser, G. F. Wyeth, and J. Wiles, "Learning spatial concepts from RatSLAM representations," in *Robotics and Autonomous Systems*, December 2007.

[9] O. M. Mozos, P. Jensfelt, H. Zender, G.-J. M. Kruijff, and W. Burgard, "From labels to semantics: An integrated system for conceptual spatial representations of indoor environments for mobile robots," in *Proc. of the Workshop on Semantic information in robotics at the 2007 IEEE International Conference on Robotics and Automation (ICRA)*, Rome, Italy, April 2007.

[10] S. Vasudevan, S. Gächter, V. Nguyen, and R. Siegwart, "Cognitive maps for mobile robots – an object based approach," *Robotics and Autonomous Systems*, vol. 55, no. 5, pp. 359–371, 2007.

[11] W. Kuhn, "Modeling the semantics of geographic categories through conceptual integration," in *Proc. of the Second International Conference on Geographic Information Science*. Boulder, CO, USA: Springer, September 2002.

[12] C. Dornhege and A. Kleiner, "Behavior maps for online planning of obstacle negotiation and climbing on rough terrain," in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, San Diego, CA, October 2007.

[13] A. Fedrizzi, L. Mösenlechner, F. Stulp, and M. Beetz, "Transformational planning for mobile manipulation based on action-related places," in *Proc. of the 14th IEEE International Conference on Advanced Robotics (ICAR)*, Munich, Germany, June 2009.

[14] K. R. Coventry and S. Garrod, *Saying, seeing and acting: the psychological semantics of spatial prepositions*. Hove, 2003.

[15] T. Kollar, S. Tellex, and N. Roy, "A discriminative model for understanding natural language route directions," in *Proc. of the AAAI Fall Symposium "Dialog with Robots"*, Arlington, VA, USA, November 2010.

[16] C. Mandel, U. Frese, and T. Rofer, "Robot navigation based on the mapping of coarse qualitative route descriptions to route graphs," in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Beijing, China, October 2006.

[17] H. Zender, G.-J. M. Kruijff, and I. Kruijff-Korbayová, "Situated resolution and generation of spatial referring expressions for robotic assistants," in *Twenty-first International Joint Conference on Artificial Intelligence (IJCAI)*, Pasadena, CA, USA, July 2009.

[18] "The Oxford English dictionary," 2011. [Online]. Available: http://www.oed.com

[19] A. Howard and N. Roy, "The robotics data set repository (Radish)," 2003. [Online]. Available: http://radish.sourceforge.net/

[20] W. Burgard, C. Stachniss, G. Grisetti, B. Steder, R. Kümmerle, C. Dornhege, M. Ruhnke, A. Kleiner, and J. D. Tardós, "A comparison of SLAM algorithms based on a graph of relations," in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, St. Louis, MO, USA, October 2009.