The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

Energy Minimization via Graph Cuts for Semantic Place Labeling

Ehsan Fazl-Ersi, John K. Tsotsos

Abstract— **This paper presents a novel framework for semantic place labeling by formulating the problem in terms of energy minimization. A method based on graph cuts is used to minimize energy for a function of data cost and smoothness cost. While the data term aims at assigning visual observations to a set of pre-specified place categories, using appearance-based hierarchical classifiers, the smoothness term incorporates contextual evidence from neighbors to ensure that the labels vary smoothly almost everywhere while preserving discontinuities at the borders between adjacent places in the environment. Our proposed method achieved a performance of 91.85%, labeling 2,146 images from the challenging COLD database with place semantics. Correct labeling of 14.5% of images was the result of incorporating contextual information.**

I. Introduction

MODELING the robot's working environment[1] based on place[2] semantics (i.e. information that relates a location in the environment to human-understandable concepts) is a relatively new research topic in the field of robotics, with increasing interest in recent years. Semantic information about different places improves the user-interaction capabilities of a personal companion robot quite significantly, enabling it to communicate its position in a more human-friendly way (e.g., "*I am in the Kitchen*") and be instructed in a relatively natural manner (e.g., "*drive down the corridor*", "*go to the living room*"). Besides improving the interaction and communication skills of a personal companion robot, place semantics can also help the robot to better carry out other specific tasks, such as environment exploration and global localization.

In semantic place labeling, the goal is often to classify different locations of an environment into a set of pre-specified categorical semantics (e.g., "Hallway", "Office", "Kitchen", etc.), according to the observations gathered at those locations. In this paper we present a novel framework for semantic place labeling, which formulates the problem in terms of minimizing an energy function via graph cuts. The energy function is composed of data cost (i.e., the cost of assigning a semantic label to a location) and smoothness cost (the cost of assigning different semantic labels to neighboring locations). To estimate the data cost, an appearance-based place recognition method is presented, where a set of hierarchical classifiers are learned to measure the distance between observations and different semantic place categories.

The smoothness cost takes into account the contextual information to ensure that the labels vary smoothly almost everywhere while preserving discontinuities at the borders between adjacent places in the environment.

Experiments conducted on a publicly available database validate the robustness of our method in reliably classifying images into semantic place categories and indicate the efficiency of our proposed framework to significantly improve the labeling performance by efficiently incorporating contextual cues.

The remainder of this paper is organized as follows: Section II briefly reviews the-state-of-the-art methods for semantic place labelling. In Section III, we describe different steps of our method. Section IV presents the implementation details and experimental results. Finally, we conclude the paper and discuss some future work in Section V.

II. Background

One of the early semantic place labeling techniques was proposed by Martinez-Mozos et al. [10], in which a set of binary classifiers is trained to recognize specific locations (including "*room*", "*corridor*" and "*doorway*") in the environment. Each binary classifier is built by boosting simple geometric features using the AdaBoost algorithm [11], where each simple geometric feature was a numerical value, computed from the observed beams of a laser range scan, or from a polygon representation of the area covered by these observed beams. During the mapping, the robot moves around, classifies its sensor reading data into one of the semantic categories, and labels its position, according to the label of the activated class.

In a later work [8], Martínez-Mozos et al. extended their original work to incorporate the similarity constraints between neighboring points, as a form of contextual information, to update the labelling according to the contextual evidence from neighbours. they suggested the use of probabilistic relaxation labeling [12] for this purpose, which iteratively smoothes the AdaBoost classification result of each location based on the labels of the neighbors.

Friedman et al. [14] proposed an alternative approach to make use of the connectivity structure of the environment, represented by a Voronoi graph extracted from an occupancy grid map, as a source of contextual information for semantic place labeling. For each point on the Voronoi graph, observations are extracted from the occupancy map, integrating a set of spatial features (based on laser range scans) and connectivity features (e.g., number of neighbors,

E. Fazl-Ersi and J.K. Tsotsos are with the Department of Computer Science and Engineering, York University, Toronto, ON, M3J 1P3, Canada. E-Mail: {efazl, tsotsos}@cse.yorku.ca.

[1] The term "Environment" in this paper refers to the constructed indoor surroundings that provide the setting for human activities (such as home, office, etc.).

[2] The term "Place" in this paper refers to any specific area within the environment which has a spatial meaning for the user of a companion robot (such as "Bedroom", "Kitchen", "Corridor", etc.).

type of neighbors, shape information, etc.). The Voronoi graph is then converted into a Conditional Random Field graphical model to facilitate learning and inference.

A major drawback of the works of Martinez-Mozos et al. and Friedman et al. stems from the use of laser range scans as observations, which allows these methods to recognize only a certain type of place (e.g., they are not able to distinguish between places with similar geometric structure). Furthermore, the use of probabilistic relaxation labeling and Voronoi graphs for incorporating contextual cues, is resticted to constant and small neighbourhood radius for all locations (e.g., 2-8 neighboring locations). For larger or more adaptive neighborhood radii, these techniques are computationally very expensive.

Rottmann et al. [13] proposed another method which combines laser range features with visual features to enable the robot to support a greater variety of place semantics. Motivated by the fact that typical objects appear at different places with different probabilities, they defined visual features as the number of instances of certain categories of objects (including *"monitor"*, *"coffee machine"*, *"office cupboard"*, *"face"* and *"pedestrian"*) in the environment. For this purpose, a fast object detector was built for each of the considered object categories, using [14]. The visual features along with laser features are used to classify each location in the environment visited by the robot. Rottmann et al. further proposed to filter the classification results based on spatial dependencies between the semantic classes using a Hidden Markov Model (HMM) framework [15].

Using HMMs for incorporating contextual information, makes the classification of places in an environment dependent on the actual path followed by the robot. Furthermore, similar to [8] and [14], HMMs (as used by Rottmann et al.) are restricted to using narrow connectivity information (in this framework, the labeling of the current location is only dependent on the labeling of the previous location).

Similar to the work of Rottman et al., many of the recent approaches to semantic place labeling (e.g., [16] and [17]), are based on the occurrence statistics of different objects in different places. While this strategy relies on a successful object detection, which is still an open problem, it often leads to ambiguities arisen from common objects (e.g., chairs, lamps) located in different places (e.g., bedroom, kitchen). Some methods (e.g., [18]) tried to resolve these ambiguities by incorporating the locations of the objects as well. However, using spatial information for classifying places makes the models sensitive to changes in room decoration (i.e., variations in the spatial locations of the objects). Furthermore, although it has long been recognized that contextual information is crucial for data labelling tasks, the majority of the existing techniques, with some exceptions (as mentioned earlier), focus on semantic place labelling based on observation classification and ignore the importance of contextual evidence.

III. METHOD

Semantic place labeling involves assigning a conceptual place label to locations that the robot visits while exploring a new environment. An important constraint is that the labels should vary smoothly almost everywhere while preserving discontinuities at the borders between adjacent places in the environment.

In this paper we formulate the problem of semantic place labeling in terms of energy minimization, where the task is to find a labeling that minimizes the following standard energy function:

$$E(f) = \sum_{p \in P} D_p(f_p) + \sum_{p,q \in N} V_{p,q}(f_p, f_q) \quad (1)$$

In the above equation, f is a labeling, which maps a set of locations P to a set of labels L. $N \subset P \times P$ is a neighboring system on locations, and D_p and $V_{p,q}$ represent the data cost (i.e., the cost of assigning label f_p to location p) and smoothness cost (i.e., the cost of assigning labels f_p and f_q to neighboring locations p and q), respectively.

Minimizing energy functions like E is very hard. General-purpose optimization techniques (e.g., simulated annealing) have traditionally been used to minimize such energy functions. However, these techniques are extremely slow, requiring exponential time in practice. In this work, a graph cut method [1], based on the expansion move algorithm, is used that efficiently finds a labeling that corresponds to a local (or the global) minimum of E. More specifically, a specialized graph is first created such that the minimum cut on this graph also minimizes the energy function. Then the expansion move algorithm cycles through all the labels a, allowing any set of locations p to change their labels to a. For each a-expansion move, the energy of the resulted labeling is computed and the move with the lowest energy is selected. If the selected a-expansion move has lower energy than the current labeling, it becomes the current labeling and the process is iterated. When there is no a-expansion move for any label a to decrease the energy, then the algorithm terminates with a labeling that is a local minimum of E (refer to [1] for more details on energy minimization via graph cuts, with respect to the expansion move).

A. Data Cost

In semantic place labeling, a natural choice for data term is to employ a multi-class place recognition framework, where a set of trained classifiers, each representing a semantic place category (e.g., "Office", "Corridor"), measures the distance between observations and different place classes. In our method, we use a variation of the place recognition system we proposed in [2], to estimate the cost of assigning different semantic labels to observations gathered at different locations in the environment.

Images are initially represented using the Scale Invariant Feature Transform (SIFT) technique [3]. In contrast to the traditional use of SIFT, we compute SIFT descriptors on a

regular dense grid over the entire image (see Figure 1 for an example). As shown in [4], dense features are more suitable for scene classification, particularly because they can capture uniform regions such as walls, floor surfaces, etc.

Similar to the bag-of-features framework (e.g., [5]), the extracted features from training images (i.e., SIFT descriptors of 16x16 pixel patches computed over a grid with spacing of 8 pixels) are quantized into a compact set of *visual words*, built automatically during training, using a clustering method (e.g., agglomerative clustering). However, unlike the original bag-of-features methods, which represent an image by aggregating the image features into a single global histogram, we use a spatial pyramid approach, similar to [6], to also take into account the spatial layout of the image features. This technique works by partitioning the image into increasingly fine sub-regions and computing histograms of local features found inside each sub-region. Therefore, while resolution at which the features are extracted remains fixed, the spatial resolution at which they are aggregated varies at each level (refer to [6] for more details).

Each visual word corresponding to a bin of the spatial histograms in the pyramid representation can act as a binary classifier, firing when its value (i.e., the number of image features that fall in that spatial bin) is above a threshold, and not firing otherwise:

$$f_n(I, \theta_n) = \begin{cases} 1, & if\ h_n > \theta_n \\ 0, & otherwise \end{cases} \quad (2)$$

Here f_n is a binary variable and θ_n is an optimal threshold associated with h_n. Given a collection of binary features, a classifier for each place can be learned by selecting and combining the appropriate features that can best separate the positive and negative training samples of that place[3]. To this aim, a binary variable $C(I)$ is used to represent the class, where $C(I) = 1$ if the image I belongs to the class, and 0 otherwise.

The discriminative value of each feature is measured by the amount of mutual information it can deliver about the class [7]:

$$I(f_n;\ C) = H(C) - H(C|f_n) \quad (3)$$

In the above equation, $I(f_n;\ C)$ is the mutual information between feature f_n and class C, and H denotes entropy. Informative feature selection starts by identifying the feature with the highest mutual information score. It then proceeds by iteratively searching for the next informative feature, f_r, that delivers the maximal amount of additional information with respect to each of the previously selected features:

$$f_r = \arg \max_{f_k \in K_r} \min_{f_j \in S_r} \left(I(f_k, f_j; C) - I(f_j; C) \right) \quad (4)$$

[3] Positive images for each place category are those taken from that place, and negative images are simply the positive images of other place categories.

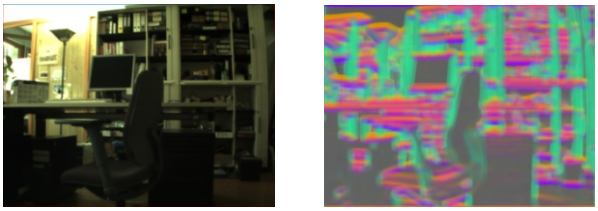


Figure 1. Visualization of SIFT descriptors of 16x16 patches, computed on a regular dense grid. The visualization is obtained by mapping the first three principal components of each 128 dimension SIFT descriptor into the principal components of the RGB color space. Note that in our experiments SIFT descriptors are computed over a grid with spacing of 8 pixels.

Here K_r and S_r are the set of features not yet selected, and the set of features already selected at iteration r, respectively. The feature selection process ends when the increment in mutual information gained by selecting a new feature is less than a certain threshold (experimentally set to 0.02), or until the number of selected features reaches a certain limit (experimentally set to 30).

Features selected so far (referred to as *top-level* features) are often strong enough to discriminate the positive and negative training images of an environment with 100% accuracy. However, it is unrealistic to expect all (or even the majority) of these features to act similarly in the test images. Therefore, for each top-level feature, a set of *child* features that provide similar information as their parents, complementary to information provided by other top-level features are selected. These child features act as 'back-up' features for their parent, standing in for the parent feature if for some reason it is missing.

To identify the child features, rather than using all the positive and negative training samples, only those that are successfully classified by the parent feature are used. Therefore, the goal is to find a combination of features that can (almost) perfectly mimic the action of the higher-level features. This can be done by applying the same information maximization procedure that was used to find top-level features.

Note that while in our previous work, parent features and their children were selected from the same level of complexity and resolution (i.e., image frame), in this work, we take advantage of the spatial pyramid representation of images, selecting the child features from increasingly finer resolutions. Therefore, while the top-level informative features are still selected from the coarsest level of resolution, capturing the *holistic* statistical properties of the images, the children or backup features are selected from a finer level of resolution, capturing more local and precise statistical characteristics. Parent and child features *together* can then provide a substantial level of robustness against appearance variations.

Feature hierarchies are built up to a pre-defined level (equal to the number of levels in the pyramid representation). Features with no children are then labeled as *atomic* features.

In the feature hierarchy, the response of each non-atomic node, f_n, indicated by s_n, is computed based on the combination of its children responses, and its own binary response (as computed by Eq. 2):

$$s_n = \left((w_0 + \sum_{i=1}^{m} w_i \, s_{ni}) > 0 \right) \bigwedge f_n(I, \theta_n) \tag{5}$$

Here, s_{ni} is the binary response of the i^{th} child of the node, m is the number of children, and w_0 and w_i are the bias and weights of the combination, respectively. Once the hierarchy is built, w_0 and w_i are computed for every non-atomic parent node, f_n, using the following equations:

$$w_i = \frac{1}{|T_p|} \sum_{j \in T_p} s_{ni}(j) - \frac{1}{|T_n|} \sum_{j \in T_n} s_{ni}(j) \tag{6}$$

$$w_0 = \frac{1}{2} \left(\frac{1}{|T_p|} \sum_{j \in T_p} \sum_{i=1}^{m} w_i \, s_{ni}(j) + \frac{1}{|T_n|} \sum_{j \in T_n} \sum_{i=1}^{m} w_i \, s_{ni}(j) \right) \tag{7}$$

In the above equations, T_p and T_n are the positive and negative training set associated with the parent node, respectively.

To determine the final response of the classifier, a *root* node is assumed for the hierarchy where the top-level features are considered as its children. The bias, w_0, and the weights, w_i, of the top-level features are then computed using Eq. 6 and Eq. 7. The response of the root node, corresponding to the entire class, is then computed using the following equation (which is derived from Eq. 5)[4]:

$$s_r = w_0 + \sum_{i=1}^{m} w_i \, s_{ri} \tag{8}$$

B. Smoothness Cost

The smoothness term in the energy function in Eq. 1 is responsible for incorporating statistic cues using contextual evidence from neighbors. Smoothness cost in our method is computed according to the following equation:

$$V_{p,q} = \begin{cases} 0, & if \; f_p = f_q \\ d(u_p, u_q), & otherwise \end{cases} \tag{9}$$

where, u_r is a vector representing the cost of assigning label f_r to location r, for all $f_r \in L$, and d is the Euclidean distance.

Considering that the semantic labeling of a location is only affected by contextual cues from its close neighbors, in our method, rather than connecting every two locations in $V_{p,q}$, we only establish neighboring links between locations that are within the same *segments*. In our system, using a simple door detection algorithm, segments are defined as partitions in the robot's trajectory that are bounded by two consequent doors (or one door and the starting or ending of the trajectory). This way, not only the contextual evidence form close neighbors are taken into account, but also the method is not dependent on some constant neighbor radius as used in [8] and [9].

The door detection algorithm used in our method is similar to the one used [17] to find narrow opening based on laser range scans. More sophisticated door detection algorithms (including vision based ones) with better accuracy could be used; however, as it is shown in our experiments, the accuracy of door detection does not significantly affect the performance of our system.

IV. Experiments

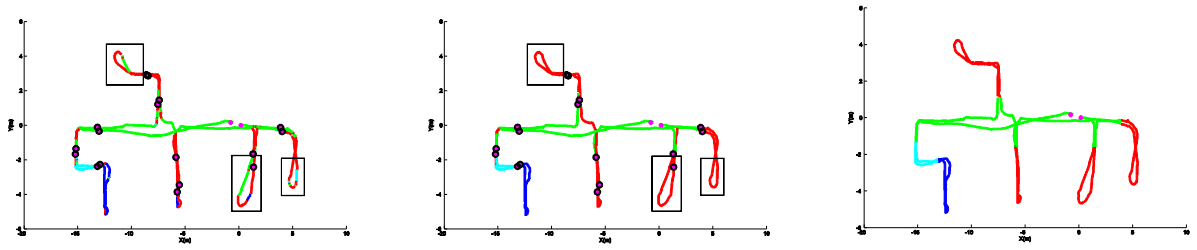
In order to evaluate the performance of our method, we use a publicly available database, called COLD (CoSy Localization Database, [19]), consisting of data sequences acquired in indoor laboratory environments with several types of places (e.g., "Corridor", "Office", "Bathroom", "Kitchen", etc.). For each place, multiple image sequences were captured over several days, under various illumination conditions.

In our experiments we use two image sequences acquired from two non-overlapping parts of Freiburg site, included in the COLD database. Similar experimental setups have been used by other researchers to evaluate semantic place labeling methods, e.g., [8] and [13], where observations gathered from part of an environment are used for training and the rest for testing. In our experiments, no scenes from the testing places are available in the training data, the lighting conditions is different between training and testing, and for the "Office" category, the method is trained with scenes from two different office places, while being evaluated on five different office places.

Our training image sequence consists of scenes from 5 places, including "Corridor", "1-Person Office", "2-Persons Office", "Bathroom", and "Stairs Area". Using the images of this sequence, 4 classifiers are learned to recognize scenes from "Corridor", "Office", "Bathroom" and "Stairs". While, for the "Office" classifier, the training data is obtained by sampling from the images of the "1-Person Office" and "2-Persons Office" in the training sequence, the remaining classifiers are trained using sample images from corresponding places in the sequence.

Overall, 800 images are used to train the classifiers, 200 images for each class. The dimensionality of the images (resized to 240x352) and the grid spacing used for dense SIFT computation (8 pixels), lead to a total of 1247 features

[4] Note that the response of each classifier shows the similarity of the observation to the class. Therefore, to be used in Eq. 1, the class responses are converted to positive distances.



(a) (b) (c)

Figure 2. The performance of our system for 2,146 test images, overlaid on robot's trajectory. (a) shows the performance of the system without incorporating the contextual information (using only the data term in Eq. 1). (b) shows the performance of the system also incorporating the contextual information (using both the data term and the smoothness term in Eq. 1). (c) shows the ground truth labeling. Comparing the areas within corresponding black rectangles in (a) and (b) indicates how the use of contextual information improves the labeling performance. In these images, red, green, cyan and blue show "Office", "Corridor", "Stairs Area" and "Bathroom", respectively. Circles overlaid on robot's trajectory indicate the locations that a door is detected. Complete demo indicating the performance on the test sequence: www.cse.yorku.ca/~efazl/iros10.avi

extracted from each image. To build the vocabulary of visual words, a random subset of 150 features from 200 training images (50 images for each class) were used, resulted in a pool of 30,000 features. Applying agglomerative clustering on the pool of features, the largest 200 clusters were selected, and for each, the centre was computed and stored as a visual word in the vocabulary. Given the vocabulary of visual words, the spatial pyramid representation for each training image was computed (up to the third level) and used to learn the hierarchical classifiers.

Our testing image sequence consists of scenes from 8 different places, namely "Corridor", "1-Person Office", "2-Persons Office 1", "2-Persons Office 2", "Large Office", "Printing Room", "Bathroom", and "Stairs Area". We expect our method to classify scenes from "1-Person Office", "2-Persons Office 1", "2-Persons Office 2", "Large Office" and "Printing Room" as "Office", while the classification of the remaining scenes should match the ground truth provided with the database (e.g., scenes from "Corridor" should be classified as "Corridor"). Overall, the testing sequence consists of 2,146 images.

Using only the data term in Eq. 1, corresponding to the classification of observations without taking into account the contextual cues, 77.35% of the test scenes are recognized correctly. Considering the change in the illumination condition and the intra-class variations, this is very good performance, demonstrating the advantages of our hierarchical classification method for dealing with variations that cannot be learned during training (e.g., illumination changes, intra-class variations, etc.). When we additionally used the smoothness term in Eq. 1, incorporating the contextual cues as well, the performance is improved to 91.85%. This substantial improvement in the performance indicates the important role of contextual cues in semantic place labeling, and validates the efficiency of our proposed framework to properly incorporate such information. In [8] and [13], performance improvement of 1.39% and 9% were reported for incorporating contextual information in semantic place labeling using probabilistic relaxation labeling and Hidden Markov Model, respectively. Although the database and classification algorithm used in our work are very different from those used in [8] and [13], comparing the results helps to put the performance of our method in taking advantage of the statistic cues, in context.

Figure 2 shows the generated semantic labels for all test scenes, with and without taking into account the contextual cues, in comparison to the ground truth. As can be seen in this figure, the use of smoothness term in Eq. 1 helps to correct many misclassifications in the light of contextual evidence from neighbors, while preserving the discontinuity at locations with low data cost (high confidence in the classification of observations, no matter if they are not consistent with contextual cues). As we will see later in this section, preserving discontinuities is very important, especially when the door detection method fails or there is no door between two neighboring places with different semantic labels.

Analyzing the confusion matrix of the performance of our method, as reported in Table 1, we found out that a substantial portion of the misclassifications are between "Office" and "Corridor" places. More specifically, the majority of misclassifications occur while the robot passes the corridor between two office places located in front of each other. In this situation, although the robot is located in the corridor, it faces an office place and therefore the captured scene is classified as "Office".

Given the robot's trajectory and laser scans for the testing sequence, the door detection algorithm used in our method detected 17 doors (including several false positives with some false negatives). To investigate how the accuracy in door detection can affect the performance of our method, we performed an experiment in which the place labeling performance is computed for scenarios where the door detection algorithm fails to find 1-9 of the 17 doors (up to 50%). In this experiment, we observed that the performance of our method gracefully degrades with the increase in false negatives of door detection. Even with more than 50% of doors missed, the method still manages to properly incorporate the contextual information and improve the initial classification results (without contextual cues). This indicates the reliability of our framework in preserving discontinuities.

	place semantic guessed by the system			
True place	*"Office"*	*"Corridor"*	*"Stairs Area "*	*"Bathroom"*
"Office"	794	19	0	0
"Corridor"	130	877	0	0
"Stairs Area"	20	2	115	0
"Bathroom"	0	0	3	175

Table 1. The confusion table for the performance of our method on the testing sequence.

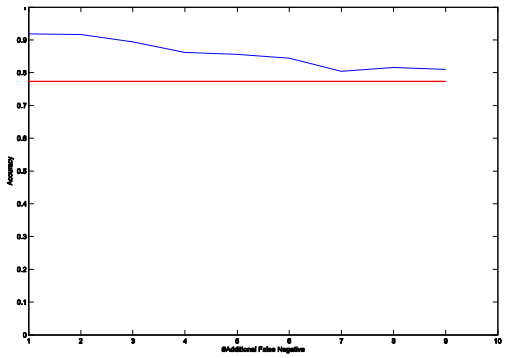


Figure 3. Place labeling performance, computed for scenarios where the door detection algorithm fails to find 1-9 of the 17 doors. For each case, the performance is averaged over 10 random trials. Even with several doors missed, the method still manages to properly incorporate the contextual information and improve the initial classification results (illustrated by a red line).

Our experiments reported in this paper, were performed on a PC with a 2.4 GHz CPU. The most time consuming process in our semantic place labeling system is the image representation (including the extraction of image features, and building the spatial pyramid), which takes around 0.8s for each image. Given the image representation, initial recognition is performed extremely fast, in just 1.5-1.7 milliseconds, depending on the number of nodes examined in each hierarch classifier. Finally complete energy minimization takes around 3.6s.

V. CONCLUSIONS

In this paper we presented a novel framework for semantic place labeling by formulating the problem in terms of energy minimization. A method based on graph cuts is used to minimize an energy function, composed of data cost and smoothness cost. While the data term aims at assigning visual observations to a set of pre-specified place categories, using appearance-based hierarchical classifiers, the smoothness term incorporates contextual evidence from neighbors to ensure that the labels vary smoothly almost everywhere while preserving discontinuities at the borders between adjacent places in the environment.

Our approach combines statistic cues and observation classification into a single and fast framework, without being restricted by constant and small neighbourhood radii, or being dependent on the actual path followed by the robot.

Experiments conducted on a publicly available database validated the robustness of our method in reliably classifying images into semantic place categories and indicated the efficiency of our proposed framework to significantly improve the labeling performance by efficiently incorporating contextual cues.

REFERENCES

[1] Boykov Y., Veksler O., Zabih R. Fast Approximate Energy Minimization via Graph Cuts. IEEE Transaction on Pattern Analysis and Machine Intelligence. vol. 23, no. 11 - 2001.

[2] Fazl-Ersi E., Elder J.H. and Tsotsos J.K. Hierarchical Appearance-Based Classifiers for Qualitative Localization. Proceedings of the International Conference on Intelligent Robots and Systems. - 2009.

[3] Lowe D., Object Recognition from Local Scale-Invariant Features. Proceedings of International Conference on Computer Vision, 1999.

[4] Fei-Fei L. and Perona P. A Bayesian Hierarchical Model for Learning Natural Scene Categories. Proceedings of International Conference on Computer Vision and Pattern Recognition. - 2005.

[5] Dance C., Willamowski J., Fan L., Bray C. and Csurka G. Visual Categorization with Bags of Keypoints. ECCV International Workshop on Statistical Learning in Computer Vision . - 2004.

[6] Lazebnik S., Schmid C. and Ponce J. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. Proceedings of the Conference on Computer Vision and Pattern Recognition. - 2006. - pp. 2169–2178.

[7] Epshtein B. and Ullman S. Feature Hierarchies for Object Classification. Proceedings of the International Conference on Computer Vision. - 2005.

[8] Martínez-Mozos O. and Burgard W., Supervised Learning of Topological Maps Using Semantic Information Extracted from Range Data. Proceedings of the International Conference on Intelligent Robots and Systems, 2006, pp. 2772-2777.

[9] Friedman S., Hanna P. and Fox D., Voronoi Random Fields: Extracting Topological Structure of Indoor Environments via Place Labeling. Proceedings of the 20th International Joint Conference on Artificial Intelligence, 2007, pp. 2109-2114.

[10] Martínez-Mozos O., Stachniss C. and Burgard W., Supervised Learning of Places from Range Data using AdaBoost. Proceedings of the International Conference on Robotics and Automation, 2005.

[11] Freund Y. and Schapire R., A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Proceedings of the Second European Conference on Computational Learning Theory, 1995, pp. 23-37.

[12] Rosenfeld A., Hummel R. and Zucker S., Scene Labeling by Relaxation Operations. IEEE Transactions on Systems, Man and Cybernetics, 1976, 6 (6), pp. 420-433.

[13] Rottmann A., Martínez Mozos O., Stachniss C., Burgard W., Semantic Place Classification of Indoor Environments with Mobile Robots Using Boosting. Proceedings of the National Conference on Artificial Intelligence, 2005, pp. 1306-1311.

[14] Viola P. and Jones M., Rapid Object Detection using a Boosted Cascade of Simple Features. Proceedings of the Conference on Computer Vision and Pattern Recognition, 2001, pp. 511-518.

[15] Rabiner L., A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Reading in Speech Recognition, 1989, 77 (2), pp. 267-296.

[16] Galindo C., Saffiotti A., Coradeschi S., Buschka P. and Fernandez-Madrigal J., Multi-Hierarchical Semantic Maps for Mobile Robotics. Proceedings of the International Conference on Intelligent Robots and Systems, 2005, pp. 2278-2283.

[17] Zender H., Martínez Mozos O., Jensfeltc P., Kruijffa G. and Burgard W., Conceptual Spatial Representations for Indoor Mobile Robots. Robotics and Autonomous Systems, 2008, 56 (6), pp. 493-502.

[18] Vasudevan S., Gächter S., Nguyen V. and Siegwart R., Cognitive Maps for Mobile Robots - An Object Based Approach. Robotics and Autonomous Systems, 2007, 55 (5), pp. 359-371 .

[19] Pronobis A. and Caputo B. COLD: COsy Localization Database. International Journal of Robotics Research, 2009. - 5 : Vol. 28.