Weakly Supervised Labeling of Dominant Image Regions in Indoor Sequences

A.C. Murillo[a⋆] J. Košecká[b] B. Micusik[b] C. Sagüés[a] J.J. Guerrero[a]

[a] DIIS-I3A, University of Zaragoza, Spain.
[b] Dpt. Computer Science, George Mason University, Fairfax, USA

Abstract. The capability of associating semantic concepts with available sensory data is an important component of environment understanding. In this work we describe an approach for annotation of dominant image regions of uniform appearance, which are typically encountered indoors, such as doors, walls and floors. One of the main challenges behind correct classification of these regions requires handling large changes in the appearance as a function of lighting conditions. Instead of using large amount of training data taken under different illumination conditions, we propose an online updating of the model learned from a small number of training examples in the initial frame. We follow a two stage classification strategy: first we estimate the probabilities of individual regions belonging to each class based on appearance only; in the second stage we use Markov Random Fields (MRF) to exploit spatial layout of the scene and improve classification results. The appearance model learned in the first frame is updated in subsequent frames using the confidences obtained by the two stage classification strategy. We demonstrate our approach on two sequences of indoor environments.

1 Introduction

The focus of research in robot perception has been in the past predominantly on metric environment representations and robot localization. The environment models were typically described in terms of simple geometric features, such as points, lines and planes. More recent works on topological representations and place recognition proposed to endow environment models with some semantic labels, such as rooms, doors, corridors [1, 2]. These type of annotations can be used for enhancing human robot interaction, enable more robust localization or can provide priors for object detection and recognition.

In this work we describe an approach for annotation of dominant image regions of uniform appearance, which are typically encountered in indoor environments, such as doors, walls and floors. In this setting the correct classification of these regions requires handling large changes in the appearance as a function of lighting conditions. This in turn requires large amount of training data or careful design of image representations/descriptors which are invariant to these

⋆ Corresponding author: acm@unizar.es

changes [3]. In order to overcome the difficulties of collecting large amount of training data, we propose in this work a strategy for on-line updating of the model learned from a small number of labeled training examples in the first frame of the sequence. First, we learn the probabilities of regions belonging to individual classes based on appearance only from these labeled examples. Markov Random Fields (MRF) are then used to exploit spatial layout of the scene to improve classification results. The subsequent video frames of the image sequence are then classified by updating the initial models using high confidence regions as well as regions which can be tracked successfully. We demonstrate our approach on two indoor sequences with different variations in appearance and resolution.

Related work. Interesting results in robotic settings have been obtained regarding semantic classification of locations [1] and place recognition [4]. More recently there has been a surge of interest in endowing the acquired maps with additional semantic information, which would better facilitate environment understanding and human-robot interaction. Several approaches have been explored and varied based on the type of sensors and classification approach used. In [5] authors tackle the problem of obtaining a model of the environment defined by instantiations of objects of predefined classes (e.g., doors, walls) given range data and color images from an omni-directional camera. In [6] authors used relational Markov Networks to learn classifiers from segment-based representations. Based on laser data, the regions of the environment are classified as walls, doors and other. More recently in the robotic context the semantic labeling techniques have been extended to multi-sensor approaches where both visual data and laser data are used for multi-class object/scene label recognition [7, 8]. Issues of model updating using only visual information and range data have been used effectively in [9] for road/no-road classification, and a proposal for on-line model update using only visual data for place recognition was presented in [10]. Multi-class terrain classification for outdoor setting has been demonstrated in [11]. The problem of building cognitive maps of indoor environments has also been studied by [2].

In computer vision, several multi-class image segmentation techniques were proposed which aim at concurrent multi-class object recognition and strive to classify all pixels in an image [12–15]. Most of the techniques proceed by modeling local appearance signals associated with pixels or small regions and encode preferences for smoothness, which are typically encoded by constructing a (conditional) Markov Random Field over image pixels or image regions. Most of the experimental evaluations in this setting is done using static images only. Issues of on-line learning of object category models have been explored in [16].

2 Our Approach

The main ingredients in the semantic labeling problem are the choice of labels (number of classes) and a classification strategy to discriminate between different classes. Figures 1(a)-(b) show a reference image from our experiments and the type of labels we seek to assign.

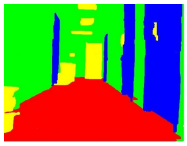
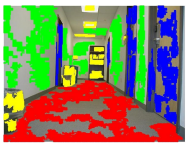
(a)  (b)  (c)  (d) 

Fig. 1. (a) Initial frame (b) ground truth labels: doors (blue), floor (red), wall (green), others (yellow) (c) extracted superpixels (d) superpixels used as training examples.

In the majority of previous approaches the learning and testing stages were completely decoupled. First a labeled data set is used to train the classifier, followed by the testing stage on the subset of the perceptual data withheld in the training stage. Since we strive for pixel accurate labeling, the process of collecting the training data is quite time consuming as it requires assigning labels to individual regions of large set of images. Although several databases with labels are available for analysis of static images, as used in [12], the semantic categories are typically not well applicable for indoor environments. In our problem domain we would like to annotate indoor video sequences by assigning one of the following labels $\chi = \{door, wall, floor, other\}$ to each pixel. This type of regions do not fall into an 'objet category', but often constitute large areas of the image and hence can be regarded as background. Some of these regions (e.g, floor, door) can serve as important cues for navigation: floor/wall labels can determine what is drivable and doors can serve as important way-points for making navigation decisions. Besides, we would like to explore the possibility of learning simple appearance models of a small number of categories from a few labeled examples and then update the models as additional examples are found based on confidences of their classification. Since in a robotic setting we naturally work with visual data streams, the model update process will greatly benefit from the temporal coherence of the perceptual data. The goal is to obtain a fully segmented and labeled video sequence from a minimal supervision at the beginning of the sequence.

2.1 Segmentation and labeling

The set of images was segmented into superpixels using a color based segmentation algorithm proposed in [17]. Figure 1(c) shows an example of a reference image segmented at the finest level tried ($\sigma = 0.5$; $k = 500$; $min\_size = 20$) where σ is the initial smoothing of the image, k is the maximum number of segments and minimal segment size is 20. Other segmentation algorithms would work as well. Superpixels will subsequently constitute the elementary regions to be classified. For each superpixel, we compute the color moments in RGB, HSV and/or Lab spaces (in the following experiments we use only means as the higher order moments were not beneficial in our setting). The number of pixels and position (centroid) are also computed for each superpixel to help in the model update process. More complex features can be designed to allow better generalization to different environments.

Initial Frame Classification. We first describe a method for classification of a single frame of the video sequence. Given a small number of training examples (labeled superpixels) in the first frame, we first assign probabilities over different classes to every superpixel in this frame.

We have explored two approaches, generative and discriminative, to learn the labels model and to classify the individual superpixels. In the discriminative setting we have tried the GentleBoost classifier as described in [18]. The generative model is trained by fitting a Gaussian mixture model (GMM) to each class label, using the standard expectation-maximization (EM) algorithm. In this paper we only describe in detail the generative model approach and its results since this method was superior in our preliminary tests compared to the discriminative approach.

MRF formulation. To account for the spatial relationships between superpixels, we define a Markov Random Field, whose graph structure is induced by superpixels and their neighborhood relationships. The final annotation is formulated as the maximum a posteriori (MAP) estimate of the MRF via the equivalent MAX-SUM labeling problem. Given a graph consisting of a discrete set $\mathcal{T}$ of vertices, each vertex $t \in \mathcal{T}$ is assigned a label $x_t \in \mathcal{X}$. Let the elements $g_t(x_t)$ and $g_{tt'}(x_t, x_{t'})$ express qualities given to label x_t in vertex t and pairwise qualities on edges between nodes x_t, $x_{t'}$ between two vertices t and t', respectively. A MAX-SUM *labeling* is a mapping that assigns a single label x_t to each vertex, which maximizes the following sum of unary and binary functions of discrete variables:

$$\mathbf{x}^* = \underset{\mathbf{x} \in \mathcal{X}^{|\mathcal{T}|}}{\operatorname{argmax}} \Big[\sum_t g_t(x_t) + \sum_{\{t,t'\}} g_{tt'}(x_t, x_{t'}) \Big]. \tag{1}$$

For better understanding of the symbols in the MAX-SUM formulation, we refer the reader to Figure 1 in [19]. Recently, very efficient and fast algorithms for solving the MAX-SUM problem through linear programming relaxation and its Lagrangian dual have been reviewed in [20, 19]. Although finding a global optimum of Eq. 1 is not guaranteed, as the problem is NP-hard, it has been shown that often the optimal solution, or one very close to it, can be reliably achieved.

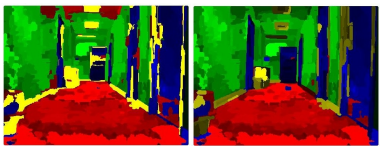
In our case the set of labels is $\chi = \{door, wall, floor, other\}$. The data term $g_t(x_t)$ label can be intuitively explained as posterior distribution over all classes for that node given the sensory data. In the generative setting we use, this is modeled as a mixture of Gaussians ($l = 5$ per class) for each label. Instead of using the entire probability distribution to determine the probability of a superpixel having particular class label χ, we find the Gaussian with the highest probability and assign that probability to the superpixel:

$$P(t|x_k) = \max_l \frac{1}{\sqrt{(2\pi)^d |\Sigma_k^l|}} \exp\left(\frac{1}{2}(\mathbf{u}_t - \mu_k^l)^T (\Sigma_k^l)^{-1} (\mathbf{u}_t - \mu_k^l)\right). \tag{2}$$

where $\mathbf{u}$ is the vector of color means computed for superpixel t. In the data term $g_t(x_t)$ the complement of the probability is being used since the data terms correspond to the penalty of an energy function being minimized. The pairwise

term $g_{tt'}(x_t, x_{t'})$ controls the mutual bond of neighboring superpixels and is in our case modeled as color mean difference between superpixels.

Figure 2 shows the initial frame classification results with the generative model without (on the left) and with the MRF smoothing (in the middle). We also use the horizon line to eliminate possibility of certain labels, e.g. the probability of being floor for all superpixels above the horizon is set to 0. The table in Fig. 2 shows the confusion matrix in the labeling with regard to the ground truth for this frame. Each row presents the actual classification for all pixels from a particular label in the ground truth, e.g., from the pixels labeled as wall in the ground truth 95% were properly labeled as wall by the automatic process, while 5% were *confused* with other labels.



	Wall	Floor	Door	Other
Wall	**0.95**	0.02	0.01	0.02
Floor	0.00	**0.98**	0.01	0.007
Door	0.02	0.07	**0.81**	0.1
Other	0.02	0.09	0.42	**0.5**

Fig. 2. Left: First frame (see Fig. 1) classification without and with MRF and horizon constraint. Here and in the remaining figures, the brighter colors in the image correspond to the higher the confidence in the classification. Right: Confusion matrix.

2.2 Model Update

Once we have obtained the initial model M_0 from the reference labels L_{ref}, and processed the first frame I_0 to get its dense labeling L_0, we propagate and update these models to the rest of the sequence. Our goal is to have the best possible model at each frame instead of obtaining a model which could be used for classification of the entire sequence. The updating process is summarized in Algorithm 1. The model M_i is obtained using the following samples: labeled examples L_{ref} in the initial frame, high confidence regions in the current frame $\tilde{L}_i$, which were obtained by the classification using the model M_{i-1}, and high confidence regions $\tilde{L}_{i-1}$ from previous frame. We first tried to update the GMM by keeping the gaussians with higher mass for each class and replacing the ones with smaller mass with new gaussians fitted to the new superpixels selected $\tilde{L}_i$. This approach had important drawbacks, since gaussians representing the appearance of smaller regions (having lower mass) tended to disappear very quickly. The approach we have used at the end to update the GMM consists of fitting from scratch the model to the set of selected labels: $L_{ref} + \tilde{L}_{i-1} + \tilde{L}_i$. Since the descriptors dimension is very low and the number of gaussians per class is small as well, this estimation process is very fast and assures constant size of the models.

The probability of a region belonging to a class depends on the distance to the GMM weighted by the percentage of regions adjacent to the corresponding superpixel in the previous frame that had that label (w_{track}). A region is considered high confidence when its probability of belonging to a class is more than 0.5 and its size is above some threshold.

Besides, in the update stage we estimate a Fundamental Matrix F between consecutive frames, from SURF [21] correspondences, and use this F to establish correspondences between superpixels. Regions whose correspondences have been established and have the same label will also be selected as high confidence regions and used for model update, even if their likelihood was below 0.5. In case F cannot be estimated due to large motions between two views we do not try to track the labels. In this case, we suppose the image divided in a 5x5 grid, and we weight the probability of each label in each of those grid cells depending on how many times a label has appeared in the sequence in that cell (w_{freq}).

Algorithm 1: Model Initialization and Update Process

$M_0 = \texttt{estimateGMM}\,(L_{ref})$ `/* estimate GMM from labels` L_{ref} `in` I_0 `*/`
$L_0 = \texttt{classify}(I_0, M_0)$ `/* classify pixels in` I_0 `using model` M_0 `*/`
for $i = 1$ **to** n **do**
  $w_{freq} = \text{updateW}(L_{i-1}, w_{freq})$ `/* count labels in each 5x5 grid cell */`
  $\text{F} = \texttt{robustEstimationF}(I_i, I_{i-1})$ `/* compute F */`
  if F **then**
    $L_i = \texttt{classify}(I_i, M_{i-1}, w_{track})$ `/* MRF labeling */`
    $C_i = \texttt{findSuperpixelCorrespondences}(F, I_i, I_{i-1})$
    $\tilde{L}_i = \texttt{selectExamples}(L_i, L_{i-1}, C_i)$ `/* take high-confidence labels */`
    $w_{track} = \texttt{getTrackWeights}(L_{i-1})$ `/* count neighbors labels in` L_{i-1} `*/`
    $M_i = \texttt{estimateGMM}(L_0, \tilde{L}_i, \tilde{L}_{i-1})$ `/* estimate new model */`
    $L_i = \texttt{classify}(I_i, M_i, w_{track})$
  else `/* do not update the model in this step */`
    $M_i = M_{i-1}$
    $L_i = \texttt{classify}(I_i, M_i, w_{freq})$

3 Results

This section shows examples of labeling the dominant regions in typical indoor sequences starting from a few samples in the first frame.

Test 1. The data used in this test consists of 40 frames (resolution 640x480) extracted equally along an indoor sequence. We have labeled every third frame for evaluation as ground truth. All superpixels were extracted with the algorithm from [17] and the color descriptors are the mean values of HSV and Lab spaces. In this test, the Fundamental Matrix between frames was computed from SURF [21] points correspondences.

Figure 1(a) shows the first frame of this test sequence, and Fig. 1(d) the labels used to learn the initial models. Then we classify all superpixels in this frame to get a dense labeling as shown in Fig. 2. Later we propagate the labeling to the rest of the sequence as explained in Sec. 2.2. Figure 3 shows the labeling obtained for frames along the sequence using the static initial model or the dynamically updated one with and without the MRF. We can see from these results the improvements of applying the model update and the MRF smoothing

as opposed to using the constant initial model for the whole sequence. The sequence consists of three smaller subsequences. Between each subsequence a sudden change occurs, e.g. 90^o rotation, which breaks completely the image flow between last frame from one subsequence and first frame of the second one. Rows 1,2 and 4 in Fig. 3 correspond to the end of each subsequence. These sudden changes are automatically detected because it is not possible to robustly estimate the F matrix at that points.

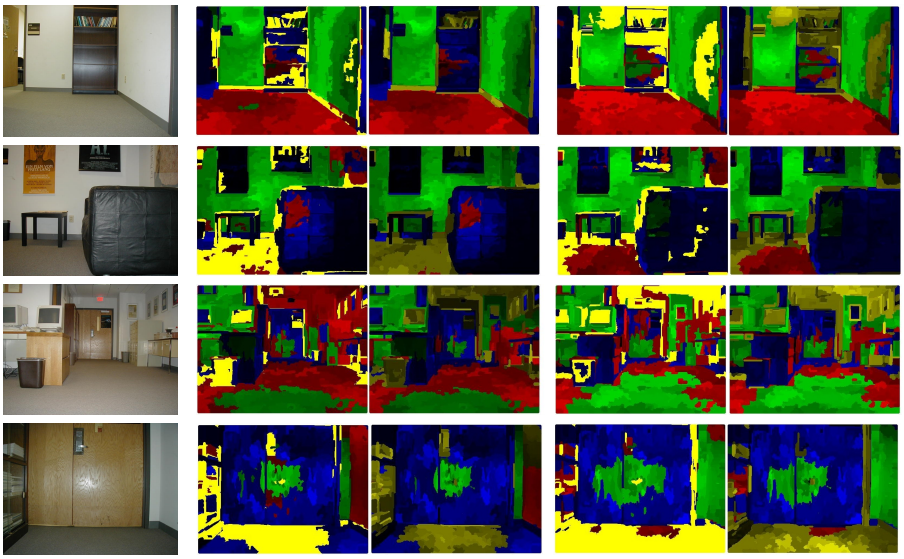
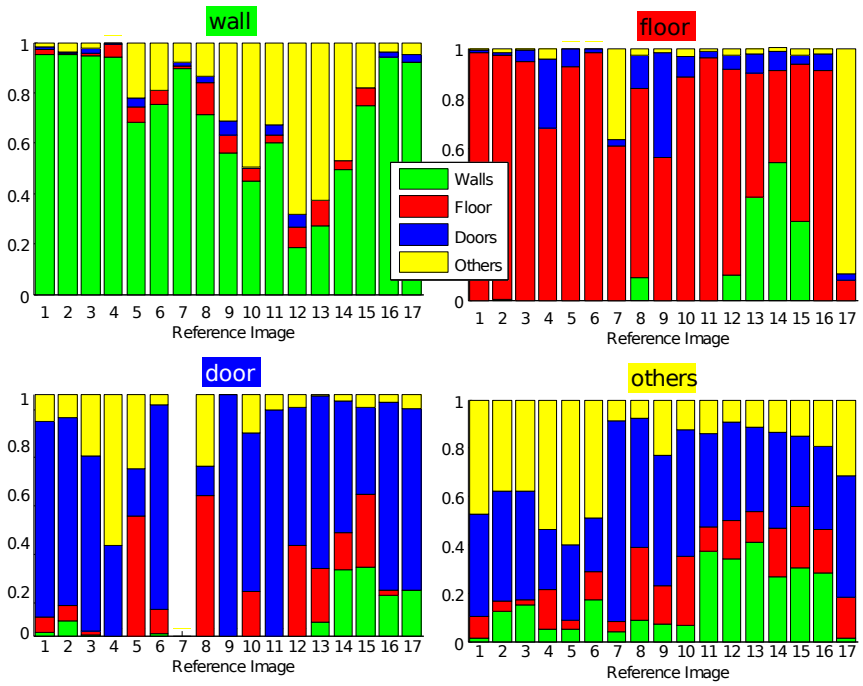


Fig. 3. Test 1. Labels propagation on the sequence frames (a). Results using the initial model without (b) or with MRF (c), or the dynamically updated model without (d) or with MRF (e).

The bar plots on top of Fig. 4 present a summary of the results in the classification for all frames with ground truth available. There, each subplot corresponds to one label, and each bar to one of the reference images with ground truth available. Each subplot shows which percentage of the area with a certain label in the ground truth was classified with each of the possible labels, e.g., first bar at first subplot shows that most wall area in the ground truth of first reference image was correctly classified as wall, while a small percentage was confused with the other three possible labels. We should note that the process seems able to recover from mistakes, e.g., in the third example in Fig. 3 big areas are incorrectly labeled, however, after a few steps, the quality of the labeling in the last frame of the sequence (row 4 in the same figure) has improved a lot. We can see in the subplots this evolution along the sequence: at some point the classifications get quite low results, but it does not propagate them too far,

thanks to the fact that the models always keep robust hints from the intial frame or those from robust correspondences. The majority of wall and door labels are correct, while the other two labels, floor and others, present worse results. However, we should notice that one of the more common mistakes in these classification is *other* labels classified as *door*, and this is due to the fact that most *other* labels correspond to furniture that is made of the same material than doors. Other issue to point regarding these results is that most confusions of the three background labels were with the *other* label.

The Table at the bottom of Fig. 4 shows the average confusion matrix, using the dynamically updated model, for all samples in the tree subsequences or only in the first one. These ratios have the same meaning as in Table in Figure 2 and are the average values for the whole sequence. As could be expected, the performance decreases after each of those sudden changes, we observe better rates if we evaluate only the first subsequence than if we evaluate the whole sequence together.



First sub-sequence (until image 7)

	Wall	Floor	Door	Other
Wall	**0.94**	0.006	0.05	0.01
Floor	0.20	**0.5**	0.3	0.006
Door	0.01	0.001	**0.98**	0.008
Other	0.08	0.005	0.5	**0.4**

All sub-sequences

	Wall	Floor	Door	Other
Wall	**0.70**	0.002	0.27	0.04
Floor	0.20	**0.24**	0.55	0.01
Door	0.06	0.001	**0.93**	0.01
Other	0.16	0.003	0.62	**0.22**

Fig. 4. Test 1. Top: classification results for each label. Bottom: average confusion matrix for all reference images.

Test 2. In this second test, we only show the automatic labeling obtained because we did not have ground truth to get more precise performance measurements as in previous test. We use two subsequences (resolution 320 x 240) from the public dataset LabelMe [1]. In this case the segmentation is done into smaller superpixels. See Fig. 5 to get a qualitative idea of the results obtained for these two sequences. We should notice the similarly good behaviour with much lower quality images, e.g., last example in test2A is quite blurred.

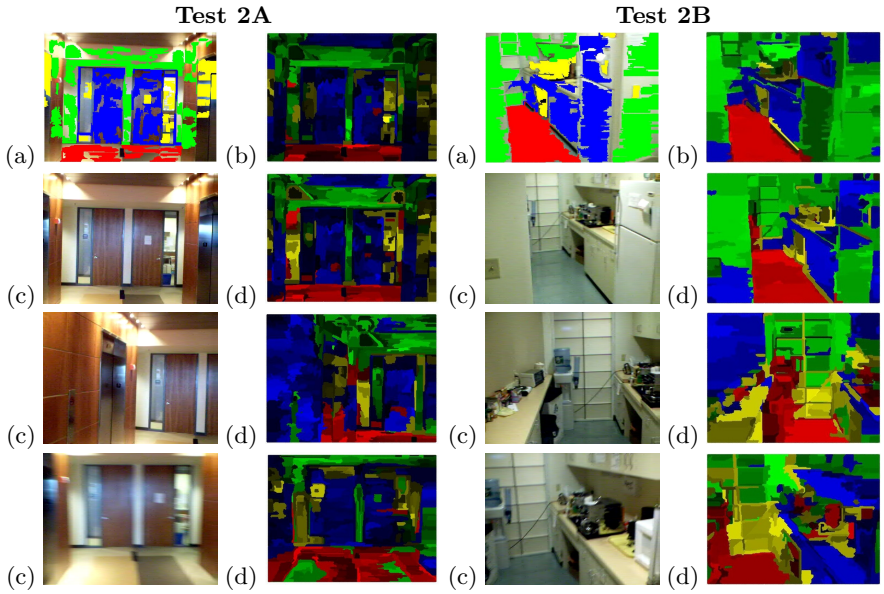


Fig. 5. Test 2. sequences labeling: walls (green), floor (red), doors in 2A or kitchen furniture in 2B (blue) and other (yellow). Reference labels (a), first frame classification (b) and other sequence frames (c) with classification results (d).

4 Conclusions

We have demonstrated an approach for semantic labeling of large regions of uniform appearance in indoor environments. The main goal is to segment the whole sequence, from minimal supervision in the initial frame, into dominant regions and smaller areas containing unknown objects. We introduced a two stage classification strategy, where we first learn the probabilities of superpixels belonging to individual regions regardless their spatial relationships, followed by solving a MRF classification. The main novelty of the approach was the idea of updating the appearance model using the high confidence regions and region correspondences. The updating yielded notably better classification, reduced dramatically the number of required labeled examples and enables us to use rather simple generative models. The currently encountered errors are due to the fact that the features used are rather weak and do not capture any geometrical relationships

[1] `http://labelme.csail.mit.edu/`

of individual labels. Furthermore in many cases the confusions occur in case the concepts cannot be disambiguated based on visual information only. In the future work we plan to extend the work by incorporating additional features and explore alternative means for modeling temporal relationships in the sequence.

References

1. Stachnis, C., Martinez-Mozos, O., Rottman, A., Burgard, W.: Semantic labeling of places. In: ISRR. (2005)
2. Vasuvedan, S., Gachter, S., Nguyen, V., Siegwart, R.: Cognitive maps for mobile robots - an object based approach. Robotics & Autonomous Systems **55**(5) (2007)
3. Weijer, J., Schmid, C., Verbeek, J.: Using high-level visual information for color constancy. In: ICCV. (2007)
4. Pronobis, A., Caputo, B., Jensfelt, P., Christensen, H.I.: A discriminative approach to robust visual place recognition. In: IROS. (2006)
5. Anguelov, D., Koller, D., Parker, E., Thrun, S.: Detecting and modelling doors with mobile robots. In: ICRA. (2004)
6. Limketkai, B., Liao, L., Fox, D.: Relational object maps for mobile robots. In: IJCAI. (2005)
7. Posner, I., Cummings, M., Newman, P.: Fast probabilistic labeling of city maps. In: RSS. (2008)
8. Douillard, B., Fox, D., Ramos, F.: A spatio-temporal probabilistic model for multi-sensor multi-class objet recognition. In: RSS. (2007)
9. Dahlkamp, H., Kaehler, A., Stavens, D., Thrun, S., Bradski, G.R.: Self-supervised monocular road detection in desert terrain. In: RSS. (2006)
10. Luo, J., Pronobis, A., Caputo, B., Jensfelt, P.: Incremental learning for place recognition in dynamic environments. In: IROS. (2007)
11. Angelova, A., Matthies, L., Helmick, D., Perona, P.: Learning and prediction of slip using visual information. In: Journal of Field Robotics. (2007)
12. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost: Joint appearance, shape and context modeling for multi-class ob ject recognition and segmentation. In: ECCV. (2006)
13. Yang, L., Meer, P., Foran, D.J.: Multi class segmentation using unified framework over mean-shift patches. In: CVPR. (2007)
14. Schroff, F., Criminisi, A., Zisserman, A.: Single-histogram class models for image segmentation. In: ICVGIP. (2006)
15. Gould, S., Rodgers, J., Cohen, D., Elidan, G., Koller, D.: Multi-class segmentation with relative location prior. Int. J. Comput. Vision (2008)
16. Opelt, A., Pinz, A., Zisserman, A.: Incremental learning of ob ject detectors using a visual shape alphabet. In: CVPR. (2006)
17. Felzenszwalb, P.F., Huttenlocher, D.: Efficient graph-based image segmentation. Int. J. Comput. Vision **59**(2) (2004) 167–181
18. Torralba, A., Murphy, K., Freeman, W.: Sharing features: efficient boosting procedures for multiclass object detection. In: CVPR. (2004)
19. Werner, T.: A linear programming approach to max-sum problem: A review. IEEE Trans. Pattern Analysis and Machine Intelligence **29**(7) (July 2007) 1165–1179
20. Kolmogorov, V.: Convergent tree-reweighted message passing for energy minimization. IEEE Trans. Pattern Analysis and Machine Intelligence **28**(10) (October 2006) 1568–1583
21. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: ECCV. (2006)