2010 IEEE International Conference on Robotics and Automation
Anchorage Convention District
May 3-8, 2010, Anchorage, Alaska, USA

Indoor Scene Recognition Through Object Detection

P. Espinace, T. Kollar, A. Soto, and N. Roy

Abstract— **Scene recognition is a highly valuable perceptual ability for an indoor mobile robot, however, current approaches for scene recognition present a significant drop in performance for the case of indoor scenes. We believe that this can be explained by the high appearance variability of indoor environments. This stresses the need to include high-level semantic information in the recognition process. In this work we propose a new approach for indoor scene recognition based on a generative probabilistic hierarchical model that uses common objects as an intermediate semantic representation. Under this model, we use object classifiers to associate low-level visual features to objects, and at the same time, we use contextual relations to associate objects to scenes. As a further contribution, we improve the performance of current state-of-the-art category-level object classifiers by including geometrical information obtained from a 3D range sensor that facilitates the implementation of a focus of attention mechanism within a Monte Carlo sampling scheme. We test our approach using real data, showing significant advantages with respect to previous state-of-the-art methods.**

I. INTRODUCTION

Mobile robotics has made great advances, however, current mobile robots have very limited capabilities to understand their surrounding. As an example, most mobile robots still represent the environment as a map with information about obstacles and free space. In some cases, this representation is enhanced with information about relevant visual landmarks, but the semantic content is still highly limited. Clearly, to increase the complexity of the tasks that mobile robots can perform in natural environments, we must provide them with a higher semantic understanding of their surrounding. Scene recognition appears as a fundamental part of this understanding. In particular, the ability to identify indoor scenes, such as an office or a kitchen, is a highly valuable perceptual ability to execute high-level tasks using mobile robots.

Scene recognition, also known as scene classification or scene categorization, has been extensively studied in areas such as Cognitive Psychology and Computer Vision [1][2]. Historically, the main source of controversy has been between achieving scene recognition using low-level features to directly capture the gist of a scene versus using intermediate semantic representations. Typically, these intermediate representations can be obtained by processes such as region segmentation or object recognition.

In terms of cognitive psychology previous studies have shown that humans are extremely efficient in capturing the overall gist of natural images, suggesting that intermediate representations are not needed [1]. Following this idea, early work in computer vision attempted to achieve scene recognition using supervised classifiers that directly operate over low-level image features such as color, texture, and shape [3] [4] [5]. The main problem with these approaches has been their inability to generalize from the training data to new scenes [2]. As discussed in [6], this problem has been particularly relevant for the case of indoor scenes.

In an attempt to overcome the previous limitation, recent work has started to include intermediate representations to bridge the gap between low-level image properties and the semantic content of a scene. The typical approach is based on image segmentation, where the input image is segmented into local regions that are later tagged with a semantic label (e.g. sky, mountain, grass, etc.) [7] [8]. Unfortunately, this approach inherits the usual poor performance of segmentation algorithms. This is particularly relevant in the case of indoor scenes, where the presence of a large number of objects usually produces scenes with significant clutter that are difficult to segment. As an alternative, some work avoids the problems of image segmentation by introducing more elaborated manual strategies to identify relevant intermediate properties [9] [10], however, the significant extra work to obtain representative training data usually precludes the proper scaling of such techniques.

Borrowing ideas from text mining, recent work on scene recognition has focused on hierarchical probabilistic methods that use unsupervised techniques in conjunction with bag-of-words schemes to obtain relevant intermediate representations [11][12]. Currently, these approaches represent the state-of-the-art for scene recognition, however, they do not perform well in the type of scenes usually visited by an indoor mobile robot. As we demonstrate in this paper, and has also been recently demonstrated in [6], these techniques show a significant drop in performance for the case of indoor scenes. This can be explained by the fact that, as opposed to outdoor scenes, indoor scenes usually lack distinctive local or global visual textural patterns.

In a related research track, recently there has been significant progress in the area of object recognition. In particular, it has been shown that it is possible to achieve real time category-level object recognition without relying on image segmentation, but instead using a sliding window approach in conjunction with a focus of attention mechanism [13]. Furthermore, several results have shown the advantage of using massive online data sources to automatically obtain relevant training data to feed the object recognition models [14]. In particular, correlations between object categories, and between objects and abstract labels (semantic labels such as kitchen), can be learned from online databases such as Flickr [15].

From the previous analysis, a key insight is the relevance of including semantic information in the scene recognition process. Furthemore, new advances in object recognition and convenient new sources of training data suggest a direct use

P. Espinace and A. Soto, Department of Computer Science, Pontificia Universidad Catolica de Chile, (pespinac, asoto@ing.puc.cl)
T. Kollar and N. Roy, Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, (tkollar, nickroy@csail.mit.edu)

of common objects as a key intermediate representation to achieve robust scene recognition. We believe that such an approach is particularly relevant for indoor environments, where current techniques do not provide satisfactory results.

In this paper we propose a new approach for indoor scene recognition based on a probabilistic hierarchical representation that uses common objects as an intermediate semantic representation. Our main intuition is that we can associate low-level features to objects through object classifiers, and we can also associate objects to scenes using contextual relations. In this respect, the natural semantic meaning of common objects facilitates the acquisition of traininig data from public web sources. We base our category-level object detectors on Adaboost classifiers operating on Gabor, HOG, and grayscale features. Additionally, we enhance our pure visual based classifiers using geometrical information obtained from a 3D range sensor that facilitates the implementation of a focus of attention mechanism within a Monte Carlo sampling scheme.

Accordingly, the main contributions of this work are: i) A new probabilistic generative model for scene recognition based on the detection of relevant common objects, ii) A new focus of attention mechanism based on a 3D range sensor that fully exploits the embedded nature of a mobile robot by directly measuring physical properties of objects such as size, height, and range disparity. iii) An empirical evaluation of the proposed method showing significant advantages with respect to previous state-of-the-art methods.

The rest of this paper is organized as follows. Section II discusses relevant previous work on visual scene recognition. Section III presents the mathematical framework behind our model to achieve scene recognition. Section IV provides details about the probabilistic models used in this work. Section V presents an evaluation of the proposed method and a comparison with state-of-the-art approaches. Finally, Section VI presents the main conclusions of this work and future avenues of research.

II. RELATED WORK

Early methods for scene recognition are based on global image features. These approaches extract low-level features, such as color or texture, and use those features to classify different scene categories. Vailaya et al. [3] use this approach for classifying city vs. landscape images. Later, they extend the method to the case of a hierarchical classification scheme [16], where images are first classified as indoor or outdoor. Chang et al. [4] estimate a belief or confidence function among the available scene labels, also using low-level global features for scene classification. During training, one classifier is built for each available scene category, then, all classifiers are applied to each test image, computing a confidence value for that image belonging to each of the categories. An important disadvantage of methods based on global image features is a poor generalization capability beyond training sets.

More reliable global approaches use low-level signatures to summarize global image statistics or semantics. Ulrich and Nourbakhsh [5] use color histograms as the image signature and a k-nearest neighbors scheme for classification. They apply their method to topological localization of an indoor mobile robot, but re-training is needed for each specific indoor environment. Oliva and Torralba [9] use an image representation based on features such as naturalness or openness, each of which corresponds to one dimension in a space that they call spatial envelope. These features are computed using coarsely localized spectral information. Siagian and Itti [17] build image signatures by using orientation, color, and intensity low-level visual saliency maps that are also shared by a focus of attention mechanism [18]. They test their approach by recognizing scenes using an outdoor mobile robot.

In terms of methods based on local image features, early approaches use a straightforward extension of low-level global approaches, where the input image is broken into local blocks or patches. Features and classifiers are applied to each of the blocks and then combined through a voting strategy [19], or a mixture of probabilistic classifier outputs [20]. The problem with these techniques is that they share the same limitations of their predecessors.

A second group of methods based on local image features uses semantic image regions such as sky, grass, or mountains, in order to classify the underlying scene. To obtain the relevant regions, these methods use an image segmentation procedure and afterward apply a classifier to each segmented region [7] [8]. Limitations of these methods rely on obtaining a good automatic image segmentation, a problem that is still hard to solve in computer vision.

Recent approaches have achieved good results in scene classification by using bag-of-words schemes. Fei-Fei and Perona [11] recognize scenes using an automatically obtained intermediate representation that is provided by an adapted version of the Latent Dirichlet Allocation (LDA) model. Bosch et al. [12] achieve scene classification by combining probabilistic Latent Semantic Analysis (pLSA) with local invariant features. Lazebnik et al. [21] modify bag-of-words representations by using a spatial pyramid that partitions the image into increasingly fine sub-regions. The main idea is to capture spatial relations among different image parts.

Recently, Quattoni and Torralba [6] propose an indoor scene recognition algorithm based on combining local and global information. They test their approach using 67 indoor image categories with results that outperform current approaches for the case of indoor scenes. Interestingly, although they do not explicitly use objects in their approach, they remark that some indoor scenes are better characterized by the objects they contain, indicating that object detection might be highly relevant to improve scene recognition for the case of indoor environments. Unfortunately, given lack of 3D information, we could not test our approach over the indoor dataset used by this work.

In terms of robotics, besides the fact that some of the already mentioned methods are applied to this field, extensive work has been done in the case of topological localization using visual landmarks [22] [23]. The main limitation of these approaches is that landmarks are usually environment specific, thus, generalization to different places usually produces poor results.

Finally, it is worth mentioning that Bosch et al. [2] provide a full bibliographic review in the field of scene recogntion (up to 2007), including a deeper description of some of the methods mentioned above.

III. PROBLEM FORMULATION

Next, we present the mathematical formulation behind our method to use objects as an intermediate semantic representation between low-level features and high level scene concepts. First, we present the core of our method considering only visual features and leaving aside 3D properties. Then, we show how 3D geometrical properties can be incorporated to enhance our formulation. Finally, we provide a mathematical approximation that makes our method computationally feasible.

A. Scene recognition using visual features

In order to model our scene recognition approach, we include the following terms:

- Define ξ to be a scene type, $\xi \in \Xi$.
- Define $s \in \{1, \dots, S\}$ to be an object class.
- Let $o_s \in [0, 1]$ indicate the presence/absence of instances of objects of class s in a given scene.
- Let $p(\xi|o_s)$ be the probability that ξ is the underlying scene, given that an object of class o_s is present in the scene.
- Define I to be an image.
- Define w_i, $i \in \{1, \dots, L\}$ to be a rectangular window that covers a specific part of image I that defines an object location.
- Let $c_{w_i} \in \{0, \dots, S\}$ indicate the output of an object classifier c when applied to image location w_i. Output 0 indicates that no object is found.
- Let $c_{1:w_L}$ be a vector describing the outputs of L classifiers calculated over a set of L windows.
- Define $f^j_{w_i}$ to be the output of feature j on window w_i.
- Let $\vec{f}_{w_i}$ be a vector describing the output of all the image features calculated over w_i.
- Let $\vec{f}_{1:w_L}$ be the complete set of features calculated over the set of L windows.

Given these terms, the probability of a place ξ given a set of features $\vec{f}_{1:w_L}$ is:

$$\begin{aligned} p(\xi|\vec{f}_{1:w_L}) &= \sum_{o_{1:S}} \sum_{c_{1:w_L}} p(\xi|o_{1:S}, c_{1:w_L}, \vec{f}_{1:w_L}) \dots \\ &\quad \dots p(o_{1:S}, c_{1:w_L}|\vec{f}_{1:w_L}) \\ &= \sum_{o_{1:S}} \sum_{c_{1:w_L}} p(\xi|o_{1:S}) p(o_{1:S}|c_{1:w_L}) p(c_{1:w_L}|\vec{f}_{1:w_L}) \end{aligned} \tag{1}$$

Let's now consider $p(o_{1:S}|c_{1:w_L})$ in Equation (1), using the Naive Bayes approximation that objects are independent given the classifier outputs, we have:

$$p(o_{1:S}|c_{1:w_L}) = \prod_s p(o_s|c_{1:w_L}) \tag{2}$$

Also, let's assume that we have detector models relating the presence of an object of class s to the output of a classifier c in any possible window, such that:

$$p(o_s = 1|c_{w_{(\cdot)}} = o_k) = p_{o_s, c_{o_k}} = 1 - p_{\bar{o_s}, c_{o_k}} \tag{3}$$

Then, considering that $p(o_s|c_{1:w_L}) = p(o_{s,w_1} \cup \dots \cup o_{s,w_L}|c_{1:w_L})$ and assuming that windows are independent, we have:

$$p(o_{1:S}|c_{1:w_L}) = \prod_s [1 - \prod_k (p_{\bar{o_s}, c_{o_k}})^{n_k}]^{o_s} [\prod_k (p_{\bar{o_s}, c_{o_k}})^{n_k}]^{1-o_s} \tag{4}$$

where $k \in \{0, \dots, S\}$ ranges over the possible classifier outputs and n_k is the number of classifications in $c_{1:w_L}$ with an output value o_k. $k = 0$ represents the case of no-object in the respective image window. The assumption of independent windows is very strong and leads to overconfident posteriors, however, in practice we have not observed significant failures due to this approximation.

As an alternative to Equation (4), when particular error models are not available for each possible classifier output, one can establish general error terms, such as:

$$\begin{aligned} p(o_s = 1|c_{(\cdot)} = o_s) &= p_{o_s, c_{o_s}} \\ p(o_s = 1|c_{(\cdot)} \neq o_s) &= p_{o_s, c_{\bar{o_s}}} \end{aligned} \tag{5}$$

In this case, Equation (4) is given by:

$$\begin{aligned} p(o_{1:S}|c_{1:w_L}) = \prod_s [1 - (p_{\bar{o_s}, c_{o_s}})^{n_s} (p_{\bar{o_s}, c_{\bar{o_s}}})^{(L-n_s)}]^{o_s} \dots \\ \dots [(p_{\bar{o_s}, c_{o_s}})^{n_s} (p_{\bar{o_s}, c_{\bar{o_s}}})^{(L-n_s)}]^{1-o_s} \end{aligned} \tag{6}$$

Let us now consider $p(c_{1:w_L}|\vec{f}_{1:w_L})$ in Equation (1), assuming independence among the visual information provided by each window, we have:

$$p(c_{1:w_L}|\vec{f}_{1:w_L}) = \prod_i p(c_{w_i}|\vec{f}_{w_i}) \tag{7}$$

Therefore, using Equation (4), we can finally express Equation (1) as:

$$\begin{aligned} p(\xi|\vec{f}_{1:w_L}) = \sum_{o_{1:S}} \sum_{c_{1:w_L}} p(\xi|o_{1:S}) \prod_s [1 - \prod_k (p_{\bar{o_s}, c_{o_k}})^{n_k}]^{o_s} \dots \\ \dots [\prod_k (p_{\bar{o_s}, c_{o_k}})^{n_k}]^{1-o_s} \prod_i p(c_{w_i}|\vec{f}_{w_i}) \end{aligned} \tag{8}$$

Note that this formulation can operate with any object detector able to classify objects from low-level visual features.

B. Adding 3D geometric information

In order to include 3D geometric information, we add the following terms to our model:

- Let D be a set of routines that calculate 3D geometric properties of an image.
- Define $d^j_{w_i}$ be the output of property j on window w_i.
- Let $\vec{d}_{w_i}$ be a vector describing the output of all the 3D geometric properties calculated over w_i.
- Let $\vec{d}_{1:w_L}$ be the complete set of geometric properties calculated over a set of L windows.

Given this information, our original problem in Equation (1) becomes

$$p(\xi|\vec{f}_{1:w_L},\vec{d}_{1:w_L}) = \sum_{o_{1:S}}\sum_{c_{1:w_L}} p(\xi|o_{1:S},c_{1:w_L},\vec{f}_{1:w_L},\vec{d}_{1:w_L})... \\ ...p(o_{1:S},c_{1:w_L}|\vec{f}_{1:w_L},\vec{d}_{1:w_L}) \\ = \sum_{o_{1:S}}\sum_{c_{1:w_L}} p(\xi|o_{1:S})p(o_{1:S}|c_{1:w_L})p(c_{1:w_L}|\vec{f}_{1:w_L},\vec{d}_{1:w_L}) \tag{9}$$

In this case, $p(\xi|o_{1:S})$ and $p(o_{1:S}|c_{1:w_L})$ are as before. In terms of $p(c_{1:w_L}|\vec{f}_{1:w_L},\vec{d}_{1:w_L})$ we have:

$$p(c_{1:w_L}|\vec{f}_{1:w_L},\vec{d}_{1:w_L}) = \prod_i p(c_{w_i}|\vec{f}_{w_i},\vec{d}_{w_i}) \tag{10}$$

Using Bayes Rule and a conditional independence assumption, we can transform Equation (10) into

$$p(c_{1:w_L}|\vec{f}_{1:w_L},\vec{d}_{1:w_L}) = \prod_i \alpha p(\vec{d}_{w_i}|c_{w_i})p(c_{w_i}|\vec{f}_{w_i}) \tag{11}$$

In our case, we use depth information to calculate three geometric properties: object size, object height, and object depth dispersion. We respectively denote these properties as: ds_{w_i}, dh_{w_i}, and dd_{w_i}. Then, $\vec{d}_{w_i} = \{ds_{w_i}, dh_{w_i}, dd_{w_i}\}$, so Equation (11) becomes:

$$p(c_{1:w_L}|\vec{f}_{1:w_L},\vec{d}_{1:w_L}) = \prod_i \alpha p(ds_{w_i},dh_{w_i},dd_{w_i}|c_{w_i})... \\ ...p(c_{w_i}|\vec{f}_{w_i}) \tag{12}$$

Assuming conditional independence among the different geometric priors,

$$p(c_{1:w_L}|\vec{f}_{1:w_L},\vec{d}_{1:w_L}) = \prod_i \alpha p(ds_{w_i}|c_{w_i})p(dh_{w_i}|c_{w_i})... \\ ...p(dd_{w_i}|c_{w_i})p(c_{w_i}|\vec{f}_{w_i}) \tag{13}$$

Finally, Equation (8) becomes

$$p(\xi|\vec{f}_{1:w_L},\vec{d}_{1:w_L}) = \sum_{o_{1:S}}\sum_{c_{1:w_L}} p(\xi|o_{1:S}) \prod_s [1 - ... \\ ... \prod_k (p_{\bar{o_s},c_{o_k}})^{n_k}]^{o_s} [\prod_k (p_{\bar{o_s},c_{o_k}})^{n_k}]^{1-o_s} \prod_i \alpha p(ds_{w_i}|c_{w_i})... \\ ...p(dh_{w_i}|c_{w_i})p(dd_{w_i}|c_{w_i})p(c_{w_i}|\vec{f}_{w_i}) \tag{14}$$

The geometric properties are independent from visual information, thus, they can be used in combination with any chosen object classifier to enhance detection performance.

C. Reducing dimensionality

As can be seen, our mathematical formulation depends on two nested summations over combinations of objects and windows. In computational terms, we can estimate the complexity of our method as follows:

- The inner summation considers the presence of all possible objects in all possible windows, thus, its complexity is ${N_{obj}}^{N_{win}}$, where N_{obj} is the number of objects being used, and N_{win} is the number of windows.
- The outer summation considers the presence of all possible objects in the scene, thus, its complexity is $2^{N_{obj}}$.
- Considering both summations, the complexity of the method is $2^{N_{obj}} * {N_{obj}}^{N_{win}}$.

A complexity of $2^{N_{obj}} * {N_{obj}}^{N_{win}}$ is intractable, particularly when N_{obj} may grow to the order of tens and N_{win} is in the order of thousands. Fortunately, many of the cases considered in these summations are highly unlikely. For example, some of the cases may include non-realistic object combinations, or may consider objects that according to the classifiers are not present in the current image. Furthermore, we can use the 3D information to discard unlikely object locations and sizes. Considering this, we can effectively reduce the computational complexity by focusing processing in likely cases. To achieve this goal, we use Monte Carlo techniques to approximate the relevant summations in Equation (14) using a sampling scheme based on a focus of attention principle.

For the outer summation we have

$$p(\xi|\vec{f}_{1:w_L},\vec{d}_{1:w_L}) = \sum_{o_{1:S}}\sum_{c_{1:L}} p(\xi|o_{1:S})p(o_{1:S}|c_{1:w_L})... \\ ...p(c_{1:w_L}|\vec{f}_{1:w_L},\vec{d}_{1:w_L}) \tag{15}$$

We can take the first term out of the inner summation and using Bayes Rule we obtain:

$$p(\xi|\vec{f}_{1:w_L},\vec{d}_{1:w_L}) = \sum_{o_{1:S}} \frac{p(o_{1:S}|\xi)p(\xi)}{p(o_{1:S})} \sum_{c_{1:w_L}} p(o_{1:S}|c_{1:w_L})... \\ ...p(c_{1:w_L}|\vec{f}_{1:w_L},\vec{d}_{1:w_L}) \tag{16}$$

This is equivalent to:

$$\sum_{o_{1:S}} p(o_{1:S}|\xi)F(o_{1:S}) \tag{17}$$

where

$$F(o_{1:S}) = \frac{p(\xi)}{p(o_{1:S})} \sum_{c_{1:w_L}} p(o_{1:S}|c_{1:w_L})p(c_{1:w_L}|\vec{f}_{1:w_L},\vec{d}_{1:w_L}) \tag{18}$$

We solve the summation by sampling from $p(o_{1:S}|\xi)$ and evaluating the samples in $F(o_{1:S})$. In the evaluation, we need to solve the inner summation.

For the inner summation we have

$$\sum_{c_{1:w_L}} p(o_{1:S}|c_{1:w_L})p(c_{1:w_L}|\vec{f}_{1:w_L},\vec{d}_{1:w_L}) \tag{19}$$

Again, we approximate the summation using a Monte Carlo scheme by sampling from $p(c_{1:w_L}|\vec{f}_{1:w_L},\vec{d}_{1:w_L})$ and evaluating the samples in $p(o_{1:S}|c_{1:w_L})$. Here, we use

the combination $o_{1:S}$ that comes from the current sample of the outer summation. In order to sample from $p(c_{1:w_L}|\vec{f}_{1:w_L}, \vec{d}_{1:w_L})$, we use our assumption of independence among windows:

- A combination $x \in c_{1:w_L}$ can be seen as a binary array of length L, where each element in the array represents the object that is present in one particular window (zero if nothing is present).
- A sample x_k can be obtained by getting a sample for each of the windows, $x_k = \{x_k^1, x_k^2, ..., x_k^L\}$, where each element x_k^i is obtained according to the probability distribution of the presence of objects in the corresponding window.
- For each window w_i, we build a multi-class probability distribution for the presence of objects in the window by joining a set of two-class object classifiers and normalizing afterwards.

IV. BUILDING THE SCENE DETECTOR

Next, we show how we compute each of the terms in the previous probabilistic model.

A. Category-level object detection

In this sub-section, we present our approach to category-level object detection and show how we compute $p(c_{1:w_L}|\vec{f}_{1:w_L}, \vec{d}_{1:w_L})$. As shown before, this term can be expressed as $\alpha p(\vec{d}_{w_i}|c_{w_i})p(c_{w_i}|\vec{f}_{w_i})$, therefore, we focus on these two sub-terms.

1) Computing $p(c_{w_i}|\vec{f}_{w_i})$: First, we apply an offline training procedure to obtain classifiers for each object class. We collect a representative dataset using selected images from 3 main sources: Label Me [24], Caltech 101, and Google images. Then, we extract a group of features for each training instance. Following [25], we explore an extremely large set of potentially relevant features to increase the hypothesis space, and rely on learning to select features relevant to each object model. Specifically, we use a pyramidal decomposition similar to the approach in [26], computing the same features at different image patches within a single image. This allows us to extract global and local information from each object instance. In our approach we use a 3-level pyramid, obtaining a total of 21 image patches per object instance. For each of these 21 patches, we extract 3 types of features:

1) Grayscale features given by the mean and standard deviation of the intensity value within each patch (2 features).
2) Gabor features given by 2-D Gaussian-shaped bandpass filters with dyadic treatment of the radial spatial frequency range and multiple orientations. We use 8 different scales and 8 different orientations and calculate the mean and standard deviation of the convolved region (128 features total).
3) Histogram of oriented gradients (HOG) [27] given by the magnitude of the gradients of a patch in different orientations. We use histograms with 4 different number of bins (36, 18, 8 and 4 bins), and consider each bin as one feature (66 features total).

Using these features, we learn models for each object class using AdaBoost, with weak classifiers that use linear separation of a single feature. We use the feature selection properties of AdaBoost, so from the original set of 4116 available features, each final classifier uses fewer than 100.

At execution time, we apply the classifiers using a sliding window procedure that allows us to compute $p(c_{w_i}|\vec{f}_{w_i})$. For efficiency, similarly to previous approaches [13], we arrange the AdaBoost voting scheme in a cascade that only uses each further weak classifier if the performance of the previous classifier is above a suitable threshold. For each window, we approximate a probability distribution that considers the aggregated votes of the ensemble of weak classifiers that have operated so far over the window. At each stage of the cascade, any window with classifier response below the corresponding threshold receives a probability value of zero for the presence of the corresponding object, allowing to discard unlikely image places quickly. Windows that successfully reach the end of the cascade receive an estimation of $p(c_{w_i}|\vec{f}_{w_i})$

2) Computing $p(\vec{d}_{w_i}|c_{w_i})$: To obtain this term we use a 3D swiss ranger that provides a pixel level estimate of the distance from the camera to the objects in the environment (depth map). Given an image and its corresponding depth map, we use the camera parameters and standard projective geometry to calculate features $\vec{d} = \{ds, dh, dd\}$ for each candidate window containing a potential object, where ds is the object size given by its width and height, dh is the object altitude given by its distance from the floor plane, and dd is the object internal disparity given by the standard deviation of the distances inside the object. Each of these individual properties has its associated term in our equations and their probabilities take the form of a Gaussian distribution with mean and covariance that is learned from data,

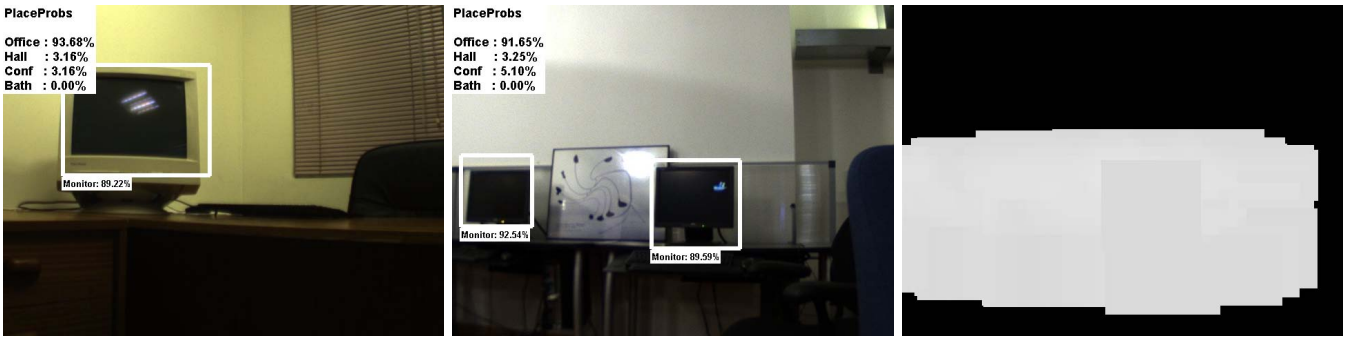
$$ds_i|c_{w_i} \sim N(\mu_{ds}, \Sigma_{ds})$$
$$dh_i|c_{w_i} \sim N(\mu_{dh}, \sigma_{dh}^2)$$
$$dd_i|c_{w_i} \sim N(\mu_{dd}, \sigma_{dd}^2)$$

Note that ds includes the height and width of the detection window, therefore is estimated using a 2-dimensional Gaussian.

In order to take full advantage of 3D information, we use the geometric properties described before as a focus of attention mechanism. As seen in Equation (12), the probability for the presence of an object in a window is a multiplication of a term that depends on 3D geometric features and a term that depends on visual features. We take advantage of this fact by using geometric properties at the initial steps of the cascade of classifier, quickly discarding windows that contain inconsistent 3D information, such as a door floating in the air. In our experiments, we found that by using geometric properties as an initial filtering step, we were able to reduce processing time by an average of 51.9% with respect to the case using just visual attributes.

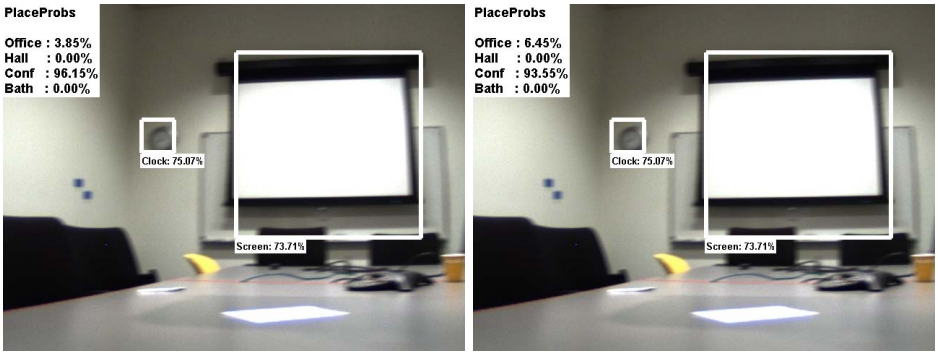
B. Classifiers confidence

Given that an object has been detected at a specific window, we require an estimate of the confidence of that detection. These confidence values correspond to the term $p(o_{1:S}|c_{1:w_L})$ in our model. We estimate this term by counting the number of true-positives and false-positives provided by our classifiers on test datasets. Actually, as stated in Equation (4), we estimate the probability that each classifier can confuse an object with each of the other objects.



(a) DCC-PUC (b) CSAIL-MIT (c) Focus of Attention for CSAIL-MIT

Fig. 1. a-b) Executions at two different office scenes. c) Focus of attention mechanism applied to image in b).



(a) Execution 1 (b) Execution 2

Fig. 2. Two different executions for the same image in a conference room scene. We can see that both executions are slightly different because of the sampling effect.

(a) Hall is the most likely place (b) Office is the most likely place

Fig. 3. Two different executions where doors are detected.

C. Prior of objects present in a scene

It is well known that some object configurations are more likely to appear in certain scene types than in others. As we show in [15], this contextual prior information can be inferred from huge datasets, such as Flickr. In our method, we follow this approach by using representative images from this dataset (in the order of hundreds for each scene type), computing the frequency of each object configuration in these images according to their tags, and normalizing to obtain the probability distributions included in the term $p(\xi|o_{1:S})$ of our model. See [15] for more details.

V. RESULTS

Our method was tested in two different indoor environments: i) Computer Science Department at Pontificia Universidad Católica de Chile (DCC-PUC), and ii) Computer Science and Artificial Intelligence Lab at Massachusetts Institute of Technology (CSAIL-MIT). In both environments, we defined four different scenes or places for which the method should compute a probability distribution given an input image: Office, Hall, Conference Room, and Bathroom. We use seven different objects to estimate place probabilities: PC-Monitor, Door, Railing, Clock, Screen, Soap dispenser,

and Urinal. Clearly, different objects are more or less related to different places. These relationships are reflected in the corresponding priors.

In all tests, we used a sliding window procedure that considers five different window shapes, including square windows, two different tall rectangular windows (height bigger than width in two different proportions), and two different wide rectangular windows (width bigger than height in two different proportions). All windows were applied using seven different image scales that emulate different window sizes. The total number of windows per image, considering all shapes and scales, was ≈ 50000.

A. Scene recognition

Figure 1 shows two different cases where PC-Monitors are detected, at DCC-PUC (figure 1.a) and CSAIL-MIT (figure 1.b). As monitors are more related to offices than to other places, Office is the most likely label for the corresponding scenes. We can see that the method makes a good decision when it finds a single object instance (DCC-PUC case) as well as when it finds more than one instance (CSAIL-MIT case). Due to our sliding window procedure, some of the instances are found inside square windows, while others are found inside wide rectangular windows. Additionally, Figure 1.c provides a view of the focus of attention mechanism applied to the case of Figure 1.b. We can see that the method discards unlikely places using only geometric properties, focusing processing in areas that are highly likely to contain monitors.

Figure 2 shows an example image where different executions produce slightly different results. This is due to the sampling procedure. In order to estimate a suitable number of samples, we tested our approach using different numbers of samples and we evaluated the variance over identical executions. As expected, increasing the number of samples reduces the variance. In our tests, we found that good results can be achieved by using a number of samples in the order of hundreds for each summation. In our final implementation, we use ≈ 1000 samples for the external summation and ≈ 100 for the internal summation in Equation (14).

Figure 3 shows that some objects, such as doors, are not very good for deciding between different places. In this example, both images were taken in Hall scenes. Figure 3.a shows an image where a door is detected and Hall becomes the most likely place, while Figure 3.b shows a case where a door is detected and Office becomes the most likely place. In our experiments, we have found that when only doors are detected, Hall is slightly more likely than other places, which is consistent with our object-scene priors. Figure 4 shows a scenario where no objects are detected, thus, the resulting place probability distribution is almost flat depending only on the priors.

B. State-of-the-art comparison

Next, we provide an experimental comparison of our method with respect to two alternative state-of-the-art approaches: i) Oliva and Torralba Gist approach (OT-G) [9], which is the same approach used as baseline for comparisons in [6], and ii) Lazebnik et al. spatial pyramid approach (LA-SP) [21]. In both cases we use an SVM for classification. For our approach, we use the most likely place as the scene

Fig. 4. Example image where no objects are detected.

detected for each image. We train all the methods using similar data obtained from the web. For testing, we use a total number of ≈ 100 images per class where at least one object is detected, mixing examples from both of our available environments (DCC-PUC and CSAIL-MIT). Tables 1-3 show the detection rates (confusion matrices) for each of the methods in each of the available scenes. We can see that our method outperforms the alternative approaches. In particular, we can see that the alternative methods tend to confuse Office and Conference Room, as both places may look very alike. Our approach present good performance for these scenarios, as it can use highly distinguishing objects, such as a proyector screen. Figure 5 shows an example where our method makes a good decision by assigning Conference Room to the underlying scene, despite partial occlusion of the only detected object. In this case, both OT-G and LA-SP detect the place as Office.

VI. CONCLUSIONS

In this work, we present an indoor scene recognition approach based on a semantic intermediate representation given by the explicit detection of common objects. During our development, we noticed the convenience of using such a high level representation, that not only facilitates the acquisition of training data from public web sites, but also provides an easy interpretation of the results, in the sense that we can easily identify failure cases where some relevant objects are not detected. This is not the case for current state-of-the-art LDA type of models, where the intermediate representation does not provide an easy interpretation. Furthermore, we believe that our representation can also facilitate the implementation of high-level task planners on mobile robots.

In terms of object detection, we show the relevance of using reliable 3D information, such as the one provided by a swiss ranger. In our case, the focus of attention mechanisms provided by the 3D geometrical properties is a key element to achieve an efficient window sampling scheme.

In terms of testing using training and test data not coming from an specific indoor environment, our approach clearly outperforms the alternative methods. This demonstrates the limitation of current state-of-the-art approaches to achieve good performance for the case of indoor scenes. Furthermore, we also test the alternative methods in the case of using testing images from the same environments used for training. In this case, the alternative methods are more competitive,

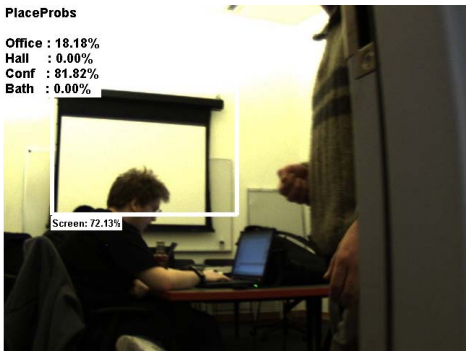


Fig. 5. Unlike alternative methods, our approach successfully detects a Conference Room scene.

although our method still presents the best results. This shows the limitations of current state-of-the-art methods to generalize their performance to new indoor environments.

Confusion matrix for the proposed method				
Scene	Office	Hall	Conference	Bathroom
Office	91%	7%	2%	0%
Hall	7%	89%	4%	0%
Conference	7%	7%	86%	0%
Bathroom	0%	6%	0%	94%

Confusion matrix for OT-G				
Scene	Office	Hall	Conference	Bathroom
Office	56%	12%	26%	6%
Hall	13%	52%	15%	20%
Conference	72%	7%	14%	7%
Bathroom	0%	9%	15%	76%

Confusion matrix for LA-SP				
Scene	Office	Hall	Conference	Bathroom
Office	44%	14%	31%	11%
Hall	19%	51%	17%	13%
Conference	38%	16%	41%	5%
Bathroom	2%	7%	13%	78%

One limitation of our approach is that images where no objects are detected cannot be identified. We claim that this is not a key problem for an indoor mobile robot because such images are usually the result of failed object detections due to artifacts such as viewpoint or illumination; the robot can move around to generate many images of a single scene with recognized objects. Additionally, a robot can use active perceptual behaviors that can guide its motions in order to find good views of key objects. This is an interesting research area for future work.

A second limitation of our method arises from the fact that running several object detectors, in addition to the scene recognition model, may result in a large execution time. Currently, depending of how many windows are discarded at early stages of the cascade of classifiers, our (non-optimized) implementation takes in the order of seconds to process each image in a regular laptop computer. Given that our method is highly parallelizable, we believe that is feasible to build a real time implementation, for example using GPU hardware.

VII. ACKNOWLEDGMENTS

This work was partially funded by FONDECYT grant 1095140.

REFERENCES

[1] S. Thorpe, C. Fize, and C. Marlot, "Speed of processing in the human visual system," *Nature*, vol. 381, pp. 520–522, 1996.
[2] A. Bosch, X. Muñoz, and R. Martí, "A review: Which is the best way to organize/classify images by content?" *Image and Vision Computing*, vol. 25, pp. 778–791, 2007.
[3] A. Vailaya, A. Jain, and H. Zhang, "On image classification: city vs. landscapes," *Pattern Recog.*, vol. 31, pp. 1921–1935, 1998.
[4] E. Chang, K. Goh, G. Sychay, and G. Wu, "Cbsa: Content-based soft annotation for multimodal image retrieval using bayes point machines," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 13, pp. 26–38, 2003.
[5] I. Ulrich and I. Nourbakhsh, "Appearance-based place recog. for topological localization," in *IEEE Int. Conf. on Rob. and Automation*, 2000.
[6] A. Quattoni and A.Torralba, "Recognizing indoor scenes," in *IEEE Conf. on Comp. Vision and Pattern Recog.*, 2009.
[7] A. Mojsilovic, J. Gomes, and B. Rogowitz, "Isee: Perceptual features for image library navigation," in *SPIE Human vision and electronic imaging Conf.*, 2002.
[8] C. Fredembach, M. Schroder, and S. Susstrunk, "Eigenregions for image classification," *IEEE Trans. on Pattern Analysis and Machine Intell.*, vol. 26, no. 12, pp. 1645–1649, 2004.
[9] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *Int. Journal of Comp. Vision*, vol. 42, pp. 145–175, 2001.
[10] J. Vogel and B. Schiele, "A semantic typicality measure for natural scene categorization," in *Pattern Recog. Symposium*, 2004.
[11] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *IEEE Int. Conf. on Comp. Vision and Pattern Recog.*, 2005.
[12] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification via plsa," in *European Conf. on Comp. Vision*, 2006.
[13] P. Viola and M. Jones, "Robust real-time face detection," *Int. Journal of Comp. Vision*, vol. 57, no. 2, pp. 137–154, 2004.
[14] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from google's image search," in *IEEE Int. Conf. on Comp. Vision*, 2005.
[15] T. Kollar and N. Roy, "Utilizing object-object and object-scene context when planning to find things," in *Int. Conf. on Rob. and Automation*, 2009.
[16] A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang, "Content-based hierarchical classification of vacation images," in *IEEE Int. Conf. on Multimedia Computing and Systems*, 1999.
[17] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Trans. on Pattern Analysis and Machine Intell.*, vol. 29, no. 2, pp. 300–312, 2007.
[18] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on Pattern Analysis and Machine Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.
[19] M. Szummer and R. Picard, "Indoor-outdoor image classification," in *IEEE Int. Conf. on Comp. Vision, Workshop on Content-based Access of Image and Video Databases*, 1998.
[20] S. Paek and S. Chang, "A knowledge engineering approach for image classification based on probabilistic reasoning systems," in *IEEE Int. Conf. on Multimedia and Expo*, 2000.
[21] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Int. Conf. on Comp. Vision and Pattern Recog.*, 2006.
[22] P. Espinace, D. Langdon, and A. Soto, "Unsupervised identification of useful visual landmarks using multiple segmentations and top-down feedback," *Rob. and Aut. Systems*, vol. 56, no. 6, pp. 538–548, 2008.
[23] M. Cummins and P. Newman, "FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance," *The Int. Journal of Rob. Research*, vol. 27, no. 6, pp. 647–665, 2008.
[24] B. Russell, A. Torralba, K. Murphy, and K. Freeman, "Labelme: a database and web-based tool for image annotation," *Int. Journal of Comp. Vision*, vol. 77, no. 1-3, pp. 157–173, 2008.
[25] D. Mery and A. Soto, "Features: The more the better," in *The 7th WSEAS Int. Conf. on Signal Processing, Computational Geometry and Artificial Vision*, 2008.
[26] A. Bosch, A. Zisserman, and X. Muñoz, "Image classification using random forests and ferns," in *IEEE Int. Conf. on Comp. Vision*, 2007.
[27] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *European Conf. on Comp. Vision*, 2005.