Knowledge-Based Systems 240 (2022) 108022

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Efficient semantic place categorization by a robot through active line-of-sight selection

Jose Luis Matez-Bandera, Javier Monroy *, Javier Gonzalez-Jimenez

Machine Perception and Intelligent Robotics group (MAPIR), Dept. of System Engineering and Automation, University of Malaga, Spain
Instituto de Investigación Biomédica de Málaga–IBIMA, Spain

ARTICLE INFO

Article history:
Received 29 July 2021
Received in revised form 26 October 2021
Accepted 18 December 2021
Available online 24 December 2021

Keywords:
Semantic knowledge
Mobile robots
Attention mechanism
Place categorization
Markov decision processes

ABSTRACT

In this paper, we present an attention mechanism for mobile robots to face the problem of place categorization. Our approach, which is based on active perception, aims to capture images with characteristic or distinctive details of the environment that can be exploited to improve the efficiency (quickness and accuracy) of the place categorization. To do so, at each time moment, our proposal selects the most informative view by controlling the line-of-sight of the robot's camera through a pan-only unit. We root our proposal on an information maximization scheme, formalized as a next-best-view problem through a Markov Decision Process (MDP) model. The latter exploits the short-time estimated navigation path of the robot to anticipate the next robot's movements and make consistent decisions. We demonstrate over two datasets, with simulated and real data, that our proposal generalizes well for the two main paradigms of place categorization (object-based and image-based), outperforming typical camera-configurations (fixed and continuously-rotating) and a pure-exploratory approach, both in quickness and accuracy.

© 2021 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Mobile robots are increasingly gaining presence in human-centered environments, like houses [1,2] or convention centers [3,4]. For a mobile robot to autonomously and safely navigate on such challenging scenarios, it needs a geometric representation or map from which: plan paths, support map-based localization, avoid obstacles, etc. Acquiring such geometric representation is then a first step when deploying a mobile robot. However, for intelligent operation, human interaction and assistance, geometric information must be complemented with semantic knowledge (SK) such as the recognition of the objects and places within the environment, or their contextual relations [5,6]. SK is required to improve the integration and usefulness of mobile robots, improving object classification [7,8], enabling high-level decision processes [1,9] or modulating the robot behavior (e.g. navigation speeds, path planning, voice level, HRI tuning, etc.) [10,11].

In this work we focus on the problem of semantic place categorization, which refers to the problem of assigning a semantic label to places or parts of the environment once their geometry is already known [12]. The usual approach involves a mobile robot inspecting the environment autonomously while capturing images that are used to infer the semantic label of each place (also referred in the literature as visual place categorization). For example, a robot in an unfamiliar home environment should be able to recognize the nature of the rooms it visits, such as kitchen, bedroom, etc. Yet, a major concern when this place categorization is carried out by a mobile robot is the need of a powerful attention mechanism to automatically identify characteristic objects or distinctive views of a given place [13]. For instance, a person taking a picture of a kitchen will naturally frame the image to include representative details such as the stove, the sink, etc. In contrast, the video acquired by a mobile robot with a fixed on-board camera will probably include many non-informative images, since the line-of-sight is fixed, typically along the tangent of the robot path. This has an important impact on the categorization efficiency by constraining the observed areas of the environment, being necessary to either adapt the robot's path during the inspection [14], or to employ advanced image representations and temporal fusion methods to overcome it.

In this work, we propose an attention mechanism based on active perception to the place categorization problem that relies on a pan-only unit to control the line-of-sight of the camera. Our approach does not alter the navigation path of the robot but exploits the ability of recent robots to dynamically change the head orientation during navigation to continuously select the line-of-sight of the camera that maximizes the information

* Corresponding author at: Machine Perception and Intelligent Robotics group (MAPIR), Dept. of System Engineering and Automation, University of Malaga, Spain.
E-mail addresses: josematez@uma.es (J.L. Matez-Bandera), jgmonroy@uma.es (J. Monroy), javiergonzalez@uma.es (J. Gonzalez-Jimenez).

https://doi.org/10.1016/j.knosys.2021.108022
0950-7051/© 2021 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

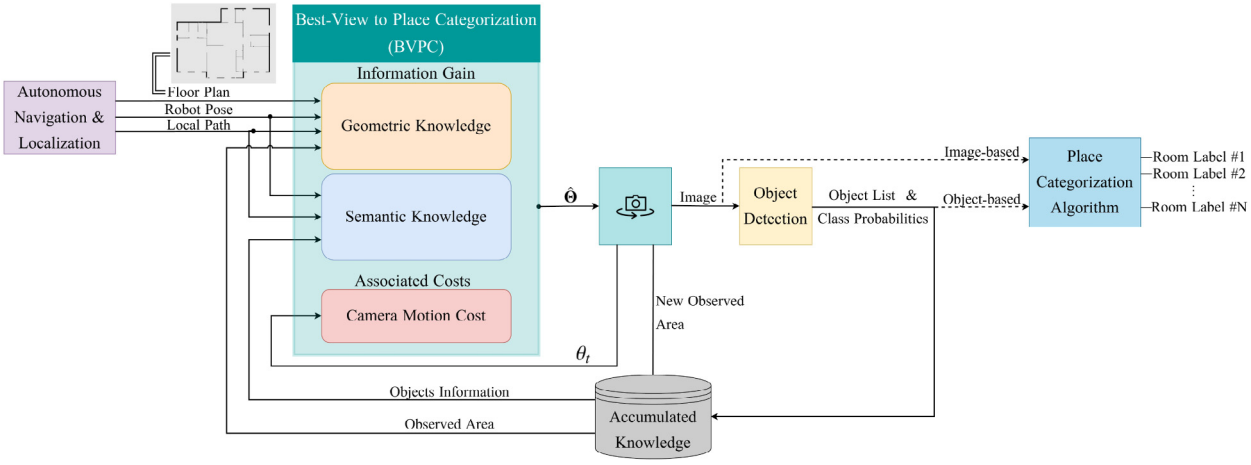


Fig. 1. System-diagram of the proposed line-of-sight selection mechanism for semantic place categorization. By exploiting knowledge of the environment (geometry and semantics), the robot pose, the expected robot path and the camera motion-cost, we select the most informative camera line-of-sight. Square blocks are the different subsystems related to our proposal. Cylinder block represent the whole accumulated knowledge of the environment. Black slashed lines refers to possibles data type inputs of place categorizations paradigms.

gain. The robot is assumed to be navigating the environment, either autonomously, by doing use of the previously gathered geometric map, or teleoperated by a human operator. In any case, an external agent controls the navigation path, which can be consulted but not altered by our method. Therefore, to select the line-of-sight, we exploit knowledge about the robot pose, the short-time estimated navigation path, the camera parameters and the environment geometry, including the segmentation of the different rooms or spaces in it. The latter is a common assumption in place categorization problems where the goal is to determine the most probable label for each segmented space in the environment [15].

We propose a probabilistic framework built upon an information maximization scheme formalized as a next-best-view problem, as well as on Markov Decision Processes (MDP) to exploit the expected short-term robot path and the previously gathered semantic knowledge. The efficiency of the system is evaluated based on two criteria: (i) quickness (a fast place categorization is desired in order to free resources such as GPUs and cameras that can be used by other tasks) and (ii) accuracy (minimizing errors in place categorization). A wide range of experiments employing two state-of-the-art datasets with real and simulated images demonstrate that our approach outperforms traditional configurations with fixed or continuously-rotating cameras and a pure-exploratory approach, while adding a small computational overhead.

Moreover, we test our approach under the two main paradigms of place categorization: image-based and object-based. The former are the most straightforward methods, inferring directly the place category from RGB images [16,17], generally employing state-of-the-art neuronal networks. On the contrary, object-based approaches first recognize the objects in the scene from a set of RGB images, and then infer the place category from the list of detected objects [18,19]. In this case, the list of detected objects together with their uncertainties are also taken into account to select the camera line-of-sight. An overview of the system-diagram of our proposal, outlined for both place categorization paradigms, is shown in Fig. 1.

2. Related work

In the last years, given the increasing interest for service robots being able to coexist with humans and to perform high-level automated tasks, multiple contributions have been presented to complement geometrical information of the environments with semantic knowledge. These contributions covers different problems such as semantic mapping [7,8] or semantic place categorization (see Section 2.1 for a review of relevant works). For the latter, which is in the scope of this work, most contributions perform place categorization by controlling the robot motion and not considering the camera rotation (i.e. assuming a fixed camera orientation w.r.t. the robot) [7,17,20]. Yet, taking into account the camera rotation is an interesting fact which enables to perform place categorization and maintain the semantic knowledge updated in parallel to the normal activity of the robot, without requiring to alter the robot navigation path. The latter is addressed in this work in which we propose to control the camera line-of-sight by an attention mechanism based on active perception. Throughout this section, we discuss related works on place categorization algorithms (see Section 2.1) and attention mechanisms based on active perception applied to other problems in mobile robotics (see Section 2.2).

2.1. Place categorization

Place categorization contributions can be divided according to the type of input data employed: object-based and image-based. On the one hand, object-based contributions rely on the semantic relationships between the category of the detected objects and their location in the environment (e.g. a bed is usually located in the bedroom, while a fridge is usually found in the kitchen). For example, Ruiz-Sarmiento et al. [21] presented a Conditional Random Field (CRF) model to categorize objects and rooms jointly from RGB-D images by exploiting contextual relations (object–object and object–room), while a prior of these relations was encoded in the form of Human Knowledge (HK) in an ontology [22]. Luo et al. [23] proposed a semantic mapping framework based on a hybrid metric-topological map. A Convolutional Neural Network (CNN) recognizes the objects that are stored as nodes in the topological map. The map is segmented by rooms and the topological nodes are clustered based on the map segmentation. Applying a Multivariate Bernoulli Naïve Bayesian (MBNB) model to semantic information from topological clusters, room labels

are inferred. Brucker et al. [14] applied a CRF to infer scene labels from recognized objects and prior knowledge (statistics correlating object presence to scene types). Ahmed et al. [18] train a Multi-class Logistic Regression (McLR) for scene classification with a set of object classes and their features. Object class and its uncertainty are obtained by applying a kernel function to object's features such as signatures and local descriptors and using the Estimated Intersection over Union (EIoU). Fernandez-Chaves et al. [19] and Oyebode et al. [24] face room categorization by using a Bayesian probabilistic framework which combines recognized objects by a CNN and its semantics encoded in an ontology.

On the other hand, image-based contributions follow a straightforward approach for place categorization by obtaining the place category directly from a given image. For instance, Sünderhauf et al. [7] train a CNN that given an RGB image as input, it is able to categorize the place between 205 different categories. For the CNN training, the authors used Places205 [16], a state-of-the-art indoor/outdoor environments' dataset. To overcome closed-set limitations, the authors integrate the CNN with one-vs-all classifiers that allow to learn new places categories online. Temporal coherence is ensured using a Bayesian filter framework. Uršič et al. [25] propose a part-based model for room categorization. Proposal regions are obtained by applying object-type-agnostic part generation. Each extracted region is codified as a descriptor by using a state-of-the-art image descriptor extractor. A mixture model of proposed parts is used to infer room category. Mancini et al. [20] implement a CNN-NBNN for semantic place categorization, an integration of a CNN with a Naïve Bayes Nearest Neighbor (NBNN). This model unifies feature extraction and classifier learning steps. In the first stage, a CNN with fully-connected layers replaced by standard convolutional layers maps an input image of arbitrary size into a set of regions. Then, a NBNN categorizes the place based on previous extracted regions. Mancini et al. [26] propose a deep learning framework for semantic place categorization facing Domain Generalization (DG) which aims to work properly under any environmental condition (occlusions, lightning changes, etc.). Pal [17] propose a combination of the object recognition CNN YOLOv3 [27] and the place categorization CNN Places365 [16] to boost the performance of visual scene recognition systems. Othman et al. [28] introduce a model that integrates different CNN architectures by using a multi-binary classifier referred as Error-Correcting Output Codes (ECOC).

However, they all share a common drawback in terms of efficiency due to a non-optimal observation of the scene, meaning that most methods will struggle when observing walls or empty areas of the environment which provide few or none information about the place category. The latter is the focus of this work, improving the categorization efficiency by selecting the most informative camera line-of-sight at each time moment.

2.2. Attention mechanisms in robotics

Generally speaking, and in the context of mobile robotics, attention mechanisms based on active perception aim to improve the efficiency of relevant tasks such as navigation [29,30], object modeling [31,32] or object manipulation [33,34], among others. Without loss of generality, most contributions formalizes active perception through a *next-best-view* problem. Given a subset of potential viewpoints, the goal is to estimate the expected information gain at each potential point and perform the selection in terms of information maximization. For example, in [31] a Hidden Markov Model (HMM) was used to optimize the viewpoint of a robot for object modeling, while in [29], authors introduced an uncertainty reduction strategy for robotic navigation through a reward function based on Partially Observable Markov Decision Processes (POMDPs). Also, noticeable are those contributions that handled information gain in terms of entropy minimization through a cost function [35,36].

3. Problem formulation

Given a mobile robot equipped with an RGB-D camera mounted on a pan unit, we seek to improve the efficiency of place categorization systems by selecting the camera's line-of-sight that maximizes the expected information gain at each time moment.

Defining $V'(\mathbf{x_t}, \theta_t)$ as the expected information gain given a robot pose ($\mathbf{x}$) and the angle of the pan unit (θ) at the time moment t, the problem can be expressed as:

$$\hat{\theta}_t = \underset{\theta}{\operatorname{argmax}}\{V'(\mathbf{x_t}, \theta_t)\}. \tag{1}$$

Our approach, from now on BVPC (Best-View to Place Categorization), considers the following assumptions:

1. Availability of a pre-segmented geometric map of the environment ($\mathcal{M}$), common in place categorization problems [15].
2. A robot equipped with an RGB-D camera mounted on a pan unit to control the observation angle (θ). The camera should be placed at a minimum height from the floor ($\geq$ 1 m) in order to have a meaningful view of the environment.
3. Control over the camera observation angle (θ), but zero control over the robot motion. We assume that the robot is moving within the environment, but we cannot alter its path, we just can consult the short-term planned navigation path computed by the navigation planner. This allows the method to anticipate the next robot poses and make more optimal and consistent decisions ahead of time than just acting reactively.
4. Only for object-based place categorization, an object detector that given an input image yields a list of detected objects and their class probabilities.

Given that the categorization of a place considers a sequence of images taken while the robot navigates it, we present in Section 3.1 our proposal for time-optimization. Then, we derive the expected information gain attending to the categorization paradigm, covering object-based categorization in Section 3.2, and image-based categorization in Section 3.3.

3.1. Time-optimization through Markov decision processes

We pursue to maximize the information gain along a specified time-horizon γ given a short-time prediction of the next robot movements. Then, the objective is to find the set of angles $\hat{\boldsymbol{\Theta}} = \{\hat{\theta}_t, \hat{\theta}_{t+1}, \ldots, \hat{\theta}_{t+\gamma}\}$ that maximizes the information gain. We formulate the problem in terms of information maximization through a Markov Decision Process (MDP). A MDP is represented by the 5-tuple $\langle S, A, T, R, \gamma \rangle$:

- $\boldsymbol{S}$ represents the finite set of states. A state s_t is defined by the observation angle θ_t and the expected robot pose $\mathbf{x_t}$ at time t, and a flag that indicates if the current state is goal (i.e. the space has already been categorized).
- $\boldsymbol{A}$ is the set of possible actions, being an action $\mathcal{R}$ composed of γ successive rotations of the pan unit over time: $\mathcal{R} = \{r_0, r_1, \ldots, r_i, \ldots, r_{\gamma-1}\}$, and r_i the pan unit rotation from θ_{t+i} to θ_{t+i+1}. For each rotation, we consider three options: maintain the current orientation, turn left or turn right $\Delta\theta$ degrees.

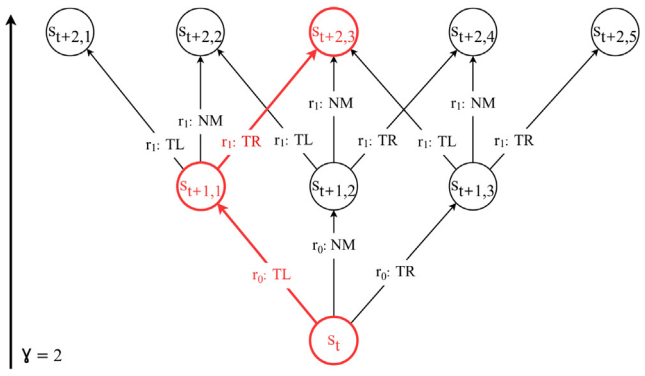


Fig. 2. An example of a tree-based representation of our MDP model for a value of $\gamma = 2$. An instance of action $\mathcal{R} = \{\text{Turn left, Turn right}\}$ is highlighted in red. TL: Turn Left, TR: Turn Right and NM: No Move.

- $\boldsymbol{T} : S \times A \rightarrow S$ stands for the transition function and denotes the probability of reaching a state s_{t+i+1} after executing a rotation $r_i \in \mathcal{R}$ from state s_{t+i}.
- $\boldsymbol{R} : S \times A \rightarrow \mathbb{R}$ is the reward that the decision framework expects to receive when executing action $\mathcal{R}$.
- $\boldsymbol{\gamma}$ is the time-horizon and it specifies the number of rotations r_i that compose an action $\mathcal{R}$, limiting the evaluation horizon. We define a tree-based representation $S(\gamma)$ of the state space given by all the possible actions $\mathcal{R} \in A$, as shown in Fig. 2.

Once the MDP model is defined, we apply the action–value paradigm, i.e. we estimate the expected gain of information for each action $\mathcal{R} \in A$. Defining the action–value function $V(s_t, \mathcal{R})$ as the expected accumulated gain of information when moving from state s_t to $s_{t+\gamma}$ through an action $\mathcal{R}$, it can be expressed according to Bellman expectation as:

$$V(s_t, \mathcal{R}) = R(s_t, \mathcal{R}) + \sum_{i=0}^{\gamma-1} T(s_{t+i+1}|s_{t+i}, a_i) V'(s_{t+i+1}), \tag{2}$$

where $T(s_{t+i+1}|s_{t+i}, a_i)$ is the probability that rotation r_i updates the current state from s_{t+i} to s_{t+i+1} (in our case, $T(\cdot) = 1$ since the states transitions are deterministic) and $V'(s_{t+i+1})$ is the expected gain of information after reaching state s_{t+i+1}, as proposed in Eq. (1). $R(s_t, \mathcal{R})$ is the reward of executing the action $\mathcal{R}$, and serves to disambiguate between actions entailing equal accumulated information gain $V(s_t, \mathcal{R})$, promoting those with few pan unit rotations.

Applying Eq. (2) to the set of possible actions A, we seek to select the action $\hat{\mathcal{R}}$ that maximizes the expected rate of information gain as:

$$\hat{\mathcal{R}} = \underset{\mathcal{R}}{\text{argmax}}\{V(s_t, \mathcal{R})\}, \text{ with } \mathcal{R} \in A. \tag{3}$$

For solving the Bellman equation, we employ *Dynamic Programming* (DP), concretely the value-iteration method, known as Bellman's update [37]. The value-iteration method works as follows: (i) from the current state s_t, we obtain the tree-based representation $S(\gamma)$ that defines the state-space A; (ii) we estimate the expected information gain for each possible action $V(s_t, \mathcal{R})$; (iii) the action $\hat{\mathcal{R}}$ with maximum expected rate of information is selected and executed; (iv) after reaching state s_{t+1} we check whether the space is categorized. If the space is categorized, the task is completed. Otherwise, a new iteration is carried out.

An MDP model becomes more complex as the value of the time-horizon γ increases, i.e. the number of possible actions $\mathcal{R}$ and the length of each action (number of pan unit rotations) increases exponentially. In brief, for a fixed value of γ, we have 3^γ possible actions, each one composed by γ rotations.

3.2. Expected information gain for object-based categorizers

Object-based place categorization relies on inferring the place category from a set of recognized objects. Its performance is, therefore, largely dependent on the set of recognized objects and their associated uncertainty [19,24]. In this sense, we seek to maximize the number of detected objects by forcing exploration of unobserved areas in the environment, while also trying to provide robustness to the recognition of previously detected ones. For the latter, multiple observations of the same object from different points of view has been demonstrated an effective approach [38].

Formally speaking, the expected gain of information after reaching state s_{t+i+1} from state s_{t+i} can be defined in terms of these two factors as:

$$V'(s_{t+i+1}) = \lambda\, \Psi_{exp}(s_{t+i+1}) + (1 - \lambda)\, \Psi_{obj}(s_{t+i+1}), \tag{4}$$

where $\Psi_{exp}(\cdot)$ and $\Psi_{obj}(\cdot)$ represent the expected information gain by exploring unobserved areas and by providing robustness to previous detected objects, respectively, and λ is a configurable weight that trades off the importance of each term ($0 \leq \lambda \leq 1$). Next, we describe in detail both terms.

3.2.1. Exploring unobserved space

The expected information gain when exploring unobserved areas of the environment stems from the possibility to discover new objects that may contribute to the inference of the place label. To numerically assess this term we make use of the probabilistic occupancy grid map of the environment $\mathcal{M}_1$, and the camera Field-of-View (FOV) parameters (angle and depth range) to apply a 2D ray-tracer over $\mathcal{M}_1$ in order to generate an observability binary grid map $\mathcal{M}_2$ (with same dimensions as $\mathcal{M}_1$), as shown in Fig. 3. Each cell in $\mathcal{M}_2$ takes the value $b_j = 1$ if the cell has already been observed, otherwise $b_j = 0$.

We evaluate the expected information gain by the ratio between the number of cells that will be observed for the first time, and the total number of cells within the FOV of the camera:

$$\Psi_{exp}(s_{t+i+1}) = \frac{\mathcal{C}(b_j = 0 \mid \theta_{t+i+1}, \mathbf{x}_{t+i+1}, \mathcal{M}_1, \mathcal{M}_2)}{\mathcal{C}(\mathcal{M}_1)}, \tag{5}$$

where $\mathcal{C}(b_j = 0 \mid \theta_{t+i+1}, \mathbf{x}_{t+i+1}, \mathcal{M}_1, \mathcal{M}_2)$ represents the number of cells observed for the first time, given the point of view $(\mathbf{x}_{t+i+1}, \theta_{t+i+1})$, and $\mathcal{C}(\mathcal{M}_1)$ is a constant defined by the number of observable cells given the FOV and the depth range of the RGB-D camera.

It must be stressed that although more elaborated approaches to quantify the information gain can be considered (e.g. relying on the concept of entropy [31]), in this work we rely on a simple, yet efficient, formulation to keep the computational overhead low. The latter is particularly important when optimizing through a time-horizon while not stopping neither slowing down the robot navigation, being fundamental to take quick decisions.

3.2.2. Temporal coherence in object classification

Without loss of generality, we can consider o^t as the observation of an unknown object o at time moment t, which is defined by the object detector probabilities $p(O_m|o^t), \forall O_m \in \boldsymbol{O}$, being $\boldsymbol{O}$ the set of M object classes. This classification, provided by an off-the-shelf object detector (a CNN in our case), is based only on a single observation and, therefore, is prone to failures. To improve it, we incorporate temporal coherence in this classification by applying a recursive Bayes filter that accounts for all the previous

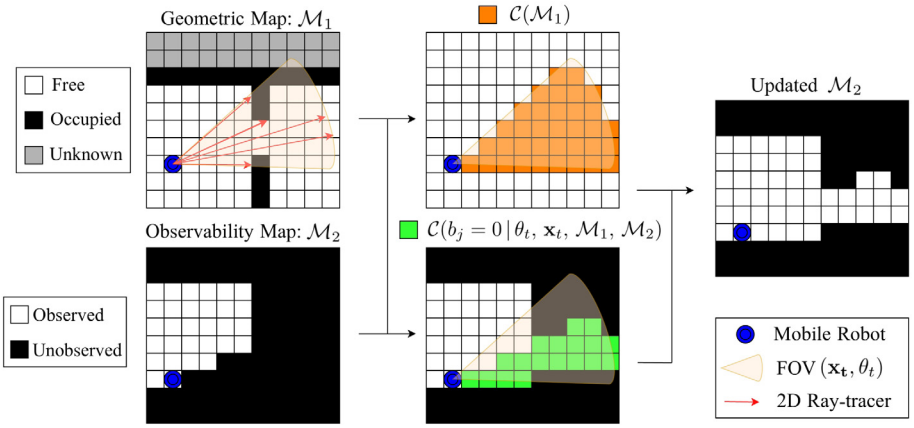


Fig. 3. Expected gain information of exploring unobserved area from a state s_t. The area to discover is estimated by running a 2D ray-tracer over $\mathcal{M}_1$.

observations concerning object o. Concretely, we express our belief $Bel(O_m^t) = p(O_m|o^{1:t})$ (i.e. the overall posterior probability over object classes) recursively over time, as:

$$Bel(O_m^t) = \frac{\mathcal{L}(o^t|O_m)\, Bel(O_m^{t-1})}{p(o^t)} \propto \mathcal{L}(o^t|O_m)\, Bel(O_m^{t-1}), \tag{6}$$

where $Bel(O_m^t)$ is our belief at time t that a specific object belongs to each possible object class O_m, $p(o^t)$ is assumed to be a constant scale factor for all $O_m \in \boldsymbol{O}$ and $\mathcal{L}(o^t|O_m)$ represents the likelihood function. Assuming first order Markov properties, this probability can be related, through Bayes theorem, to the posterior probabilities that the object detector yields:

$$\mathcal{L}(o^t|O_m) = \frac{p(O_m|o^t)p(o^t)}{p(O_m)} \propto \frac{p(O_m|o^t)}{p(O_m)}, \tag{7}$$

where $p(o^t)$ is a scale factor for all $O_m \in \boldsymbol{O}$, and $p(O_m)$ is a prior that can be learned from experimental data. For example, the probability of any object to be *chair* is, a priori, higher than that of being *toilet*, as typically there are more chairs than toilets in the environment (see Table 1).

To have an estimation of the classification uncertainty of each detected object o, we employ the well known Shannon Entropy:

$$\mathcal{H}(o) = \sum_{O_m} -Bel(O_m^t)\log(Bel(O_m^t)) \tag{8}$$

Yet, this standard entropy does not account for the relevance of the different object classes in the place categorization problem. That is, a chair, which can be found on almost any place, does not help to discern the true category of the place, while a bed, for example, provides an unequivocal contribution for the place to be a bedroom. Therefore, we define a weighted entropy $\mathcal{H}'$ where we introduce the parameter ω_m that allows us to select which objects deserve further attention (i.e. more observations):

$$\mathcal{H}'(o) = \sum_{O_m} -\omega_m Bel(O_m^t)\log(Bel(O_m^t)). \tag{9}$$

This weighting is application-dependent and user-defined (an example of this weighting can be seen in Section 4.4).

At this point, we define the expected information gain Ψ_{obj} to promote the re-observation of previously detected objects with a low number of observations, as well as those with a high entropy (seeking to reduce it with further observations):

$$\Psi_{obj}(s_{t+i+1}) = \tanh\left(\sum_{n=1}^{N} \frac{\mathcal{H}'(o_n)}{Z_n}\right), \tag{10}$$

where $\tanh(\cdot)$ is the hyperbolic tangent function, which is applied to normalize the contributions, N is the number of previously recognized objects visible from state s_{t+i+1} (see an instance in Fig. 4), and Z_n is the number of observations of the object o_n.

Table 1
Objects–room relations and appearing frequency.

N# rooms	Kitchen 31	Bedroom 70	Living room 39	Bathroom 53	$p(O_m)$ –	ω_m
Bed	0	74	0	0	0.12	3.12
Toilet	0	0	0	53	0.08	3.08
Couch	0	0	37	0	0.06	3.06
Microwave	30	0	0	0	0.06	3.05
Oven	21	0	0	0	0.03	3.03
Sink	32	0	0	53	0.14	2.14
Dining table	18	0	25	0	0.07	2.07
Chair	51	36	137	0	0.35	1.35
TV	7	15	32	0	0.09	1.09

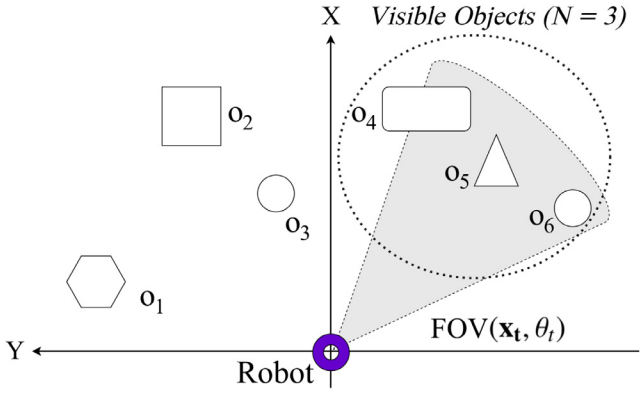
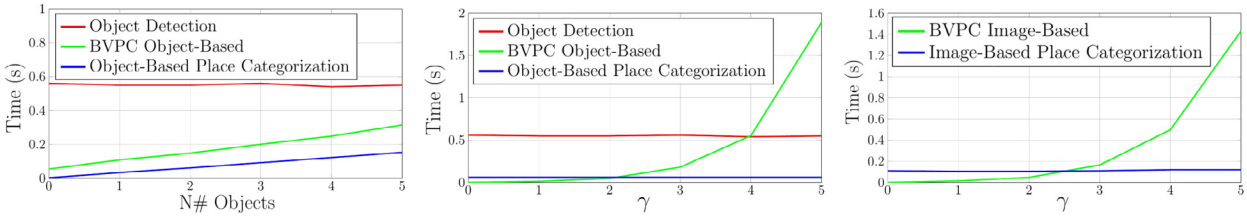


Fig. 4. An example where 3 recognized objects are visible from state s_t, i.e. $N = 3$. Note that an object o is considered as visible when at least half of the object is expected to be seen by the camera, i.e. the object center is inside of the camera FOV.

3.3. Expected information gain for image-based categorizers

Image-based place categorization just require an RGB image to infer the place category, being unnecessary to maximize the number of recognized objects or to provide robustness to the object classification, as done for object-based approaches. Since image-based approaches only use the information included in the image, these images should be as informative as possible [25]. Namely, images framing visually-coherent composition of elements and scene-related characteristics are more appropriate for



(a) Execution time for object-based place categorization with different number of detected objects and $\gamma = 2$.

(b) Execution time for object-based place categorization with different γ values and one detected object ($K = 1$).

(c) Execution time for image-based place categorization with different γ values.

Fig. 5. Execution times of each component involved in the place categorization (data obtained with a nVidia GeForce GTX 1050 and an i7-8750H @2.20 GHz).

the place categorization than low-textured images such as an image of a wall. In terms of information gain, camera viewpoints that maximize the observed area of the place should cover a more general view of the space and hence, images will be more informative for the categorization.

Based on the knowledge about the occupancy grid map $\mathcal{M}_1$, the camera parameters, and taking into account the computational limitations previously mentioned, we evaluate the expected information gain as the ratio between the number of observable cells and, the total number of cells within the camera FOV:

$$V'(s_{t+i+1}) = \frac{\mathcal{C}(\theta_{t+i+1},\, \mathbf{x}_{t+i+1},\, \mathcal{M}_1)}{\mathcal{C}(\mathcal{M}_1)}, \tag{11}$$

where $\mathcal{C}(\theta_{t+i+1},\, \mathbf{x}_{t+i+1},\, \mathcal{M}_1)$ is the number of observable cells from state s_{t+i+1}, computed by a 2D ray-tracer over the area of the map $\mathcal{M}_1$ that is covered by the camera FOV (see Fig. 3).

3.4. Time complexity

The time complexity of place a categorization algorithm mainly depends on the paradigm employed. For object-based, it is determined by the object detection stage and the categorization process. In our case, the detection of objects is carried out through a CNN $\mathcal{O}(w\,h)$, where w and h are the width and height of the input image, respectively, while the categorization process is based on a Bayesian network $\mathcal{O}(K)$ [39], being K the number of detected objects. In contrast, for image-based paradigm, the complexity corresponds solely to the CNN used for inference, which is $\mathcal{O}(w\,h)$ as previously mentioned. Moreover, since most of CNNs tend to require a fixed input image size, their time complexity can be simplified to $\mathcal{O}(1)$.

Regarding the complexity overhead added by our BVPC method, it mainly depends on the time-horizon or number of possible actions (3^γ). However, while for image-based the time complexity is just $\mathcal{O}(3^\gamma)$, for object-based we also need to consider the semantic information gain estimation, resulting in $\mathcal{O}(K\,3^\gamma)$, which is proportional to the number of recognized objects.

Fig. 5 shows a comparative analysis of such time complexities decoupling each contribution. As can be seen, despite the fact that an MDP model scales exponentially with the time-horizon (γ), as long as the γ-value is set within the range [1–4], the time overhead is not excessive.

4. Experimental setup

This section covers the setup of the multiple experiments carried out to evaluate the proposed algorithm, including an outline of the datasets and robots employed, an overview of the place categorization method implemented, the description of the different camera-configurations used for comparison, as well as some comments on parameter selection.

4.1. Datasets and robotic platforms

Two state-of-the-art robotic datasets have been selected for the evaluation: Robot@Home [40] and Robot@VirtualHome [38]. The usage of datasets is motivated by their inherent repeatability, enabling to reproduce different experiments under identical conditions and to make a fair comparison. Both datasets include geometric pre-segmented maps by spaces and the localization of the robot during exploration.

Robot@Home [40] is a collection of raw data recorded in five real households by the mobile robot Giraff, which is equipped with a rig of 4 overlapping RGB-D cameras with an overall field-of-view of 180 degrees, and a 2D laser scanner (see Fig. 6a). The cameras are placed at a height of 1.05 m from the floor, which enables a meaningful view over the environment (see Fig. 6b–c for an instance of RGB and depth images). Yet, Robot@Home does not explicitly offer a controllable pan unit. Hence, in this work we consider a virtual one[1] with 135 degrees of pan motion and a maximum rotation speed of 20 degrees per second, generated by interpolating the view from the four available fixed cameras. From Robot@Home, we selected the households *Anto*, *Alma* and *Rx2* as representative environments of large, medium and small size, respectively. Their respective maps with ground-truth room labels and the path followed by the robot are shown in Fig. 7a–c.

Robot@VirtualHome [41] is a set of 30 synthetic realistic-looking houses, recreated from real environments. The robot employed for the environment categorization is a virtualization of Giraff robot provided of a laser scanner and an RGB-D camera with 45 degrees of FOV. The latter is placed at a height of 1.05 m from the floor, mounted on a controllable pan-only unit with 135 degrees of pan motion and maximum rotation speed of 20 degrees per second. For evaluation, we selected three virtual environments: *House12* and *House18* because both contains a high number of rooms to categorize, and *House30* as a challenging environment composed by small rooms with objects in close proximity to each other. Both maps with ground-truth room labels and the path followed by the robot are depicted in Fig. 7d–f.

Since both datasets focus on households, we consider the set of place categories: *living room, bedroom, bathroom and kitchen*.

[1] In order to provide realistic movements to the virtual pan unit, we coded a script to control the camera including technical parameters such as maximum rotation speed or maximum rotation range.

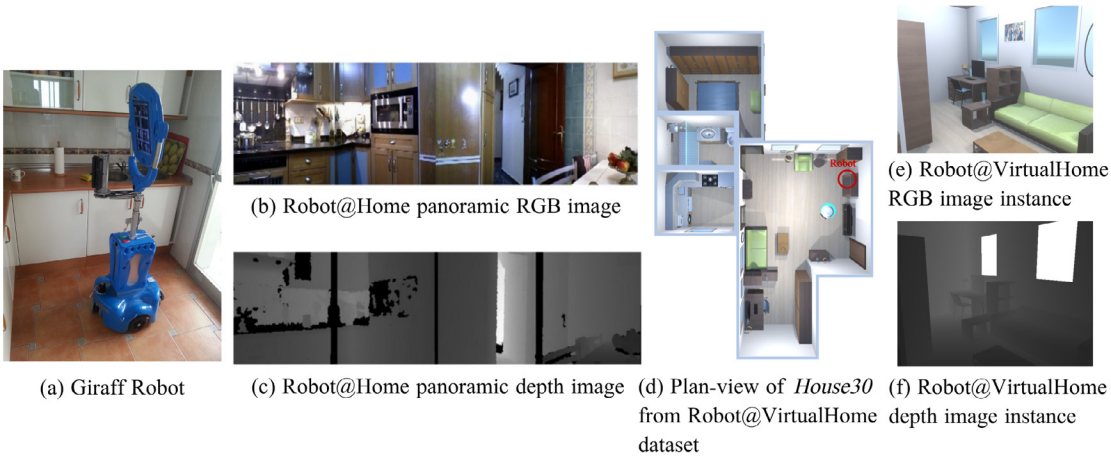


Fig. 6. (a) Giraff mobile robot used to record the Robot@Home dataset. The images (b) and (c) are an example of the resulting panoramic RGB and depth, respectively.

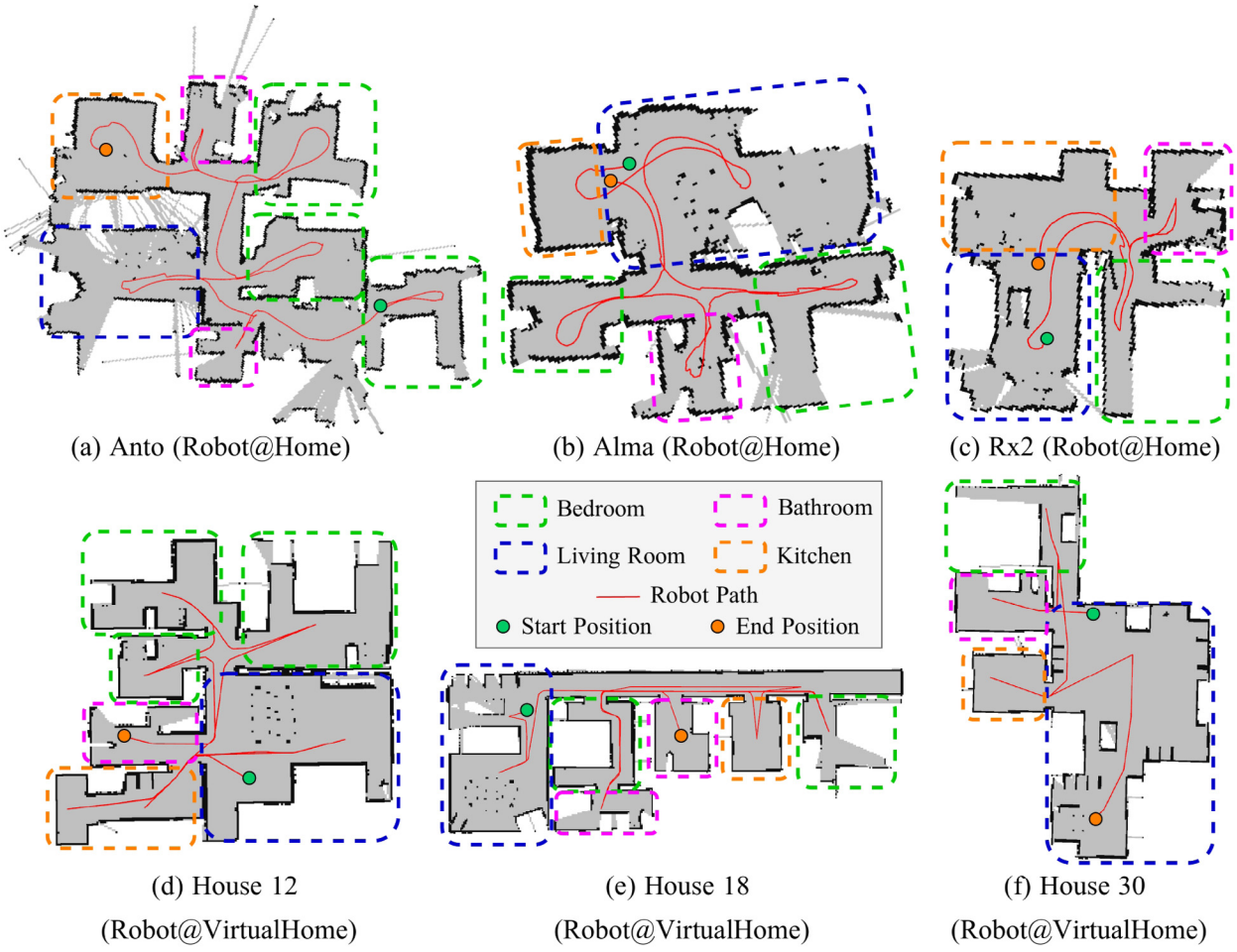


Fig. 7. Geometric maps with room category ground-truth and the path followed by the robot in each household.

4.2. Place categorization methods

To analyze the impact of our proposal when applied to object-based and image-based categorizers, we have relied on state-of-the-art methods. On the one hand, for object-based systems, and taking into consideration that objects and their properties constitute the semantic knowledge that is exploited by our proposed attention mechanism, we rely on the state-of-the-art object detection Mask R-CNN [42]. Concretely, we employ the implementation Detectron2 [43] pre-trained on MS COCO dataset [44]. This choice is motivated by the outstanding runtime performance demonstrated in recent works [45,46]. Detectron2 yields the input image with the object masks and a list with the class probabilities. Next, knowing the intrinsic parameters of the camera and the depth of each pixel of the mask obtained from the depth channel, we locate in the 3D world each detected object. The

set of detected objects in each room is then fed to a state-of-the-art Bayesian object-based place categorizer [19], that uses an ontology as the internal representation of the semantic knowledge. The latter is used in this work to accumulate the semantic information obtained from the environment, while exploiting it to select the optimal line-of-sight of the camera. The Bayesian probabilistic categorizer requires as input a list of recognized objects together with their class probabilities, and returns a probability distribution over the room categories.

On the other hand, for image-based systems, we leverage the excellent performance of the scene categorization CNN Places365 [16] shown in recent works [47,48], to classify an RGB image between 365 different categories of indoor and outdoor environments. For each individual RGB image $\mathcal{I}_t$, Places365 yields a combined probability $p(\mathbf{C}|\mathcal{I}_t)$ over the set of 365 possible place labels $\mathbf{C} = \{c_0, \ldots, c_i, \ldots, c_{364}\}$. Yet, we know in advance that just four types of indoor places can be found in the evaluation environments. Hence, as done in [7], treating the problem as Bayesian, we can incorporate prior knowledge $p(\mathbf{C})$ about the places that are unlikely to be encountered as follows:

$$p(\mathbf{C}|\mathcal{I}_t) \propto p(\mathbf{C}) \cdot p(\mathbf{C}|\mathcal{I}_t), \tag{12}$$

where $p(\mathbf{C})$ is the prior term that encodes whether each place category c_i is unlikely to be observed ($p(c_i) = 0$) or not ($p(c_i) = 1$). Note that in this expression, the prior term works as a scale factor, hence the scaled result must be normalized in order to meet the definition of a probability distribution (i.e. the sum of all the probabilities must equal 1).

Moreover, although images are classified individually, they are acquired consecutively and thus, a temporal dimension can be exploited between their classifications. Considering the place categorization as a Bayesian probabilistic estimation problem and assuming first order Markov properties, we can integrate over time the classification by employing a Bayesian filter [7]:

$$Bel(\mathbf{C}^t) = p(\mathbf{C}|\mathcal{I}_{0:t}) \propto p(\mathbf{C}) \cdot p(\mathbf{C}|\mathcal{I}_t) \cdot Bel(\mathbf{C}^{t-1}), \tag{13}$$

where $Bel(\mathbf{C}^t)$ is the belief or overall posterior probability given the set of all images taken until time moment t. Note that this belief is estimated recursively over time along the sequence of images $\mathcal{I}_{0:t}$.

4.3. Camera-configurations for comparison

Most state-of-the-art works that seek to improve the efficiency of place categorization present algorithms that control the robot motion during the inspection [7,17,20,49], not taking into account the camera rotation (i.e. assuming the camera is fixed w.r.t. the robot). These methods are not directly comparable with our approach, as controlling the robot path allows observation of any area of interest with no time restriction, as opposed to when only controlling the camera line-of-sight. Hence, for comparison, we include in this work comparison with seven different configurations. The first four ones correspond to different fixed camera-configurations typically employed in place categorization, which are referred as *RGBD1*, *RGBD2*, *RGBD3* and *RGBD4*, each one with a specific angle of observation w.r.t. robot (see Fig. 8). Moreover, to exploit the rotation capability of the camera, a fifth configuration referred as *Continuous Exploration* (CE) [50,51] is also considered. This configuration attempts to mimic the behavior of a human seeking to maximize the information acquired from an environment without prior knowledge, by continuously moving the camera from left to right and vice versa.

Finally, the last two configurations are rooted on the concept of exploration [49], where the goal is to maximize the explored area of the environment by moving the robot to their so-called frontier points. Frontiers are boundaries that separate known space from unknown space. Thus, our adaptation is to select the camera line-of-sights that maximize the observation of unknown space. As this exploratory approach allows for time optimization, we also compare with the frontier-based method working on our proposed MDP model in order to exploit the short-time estimated navigation path.

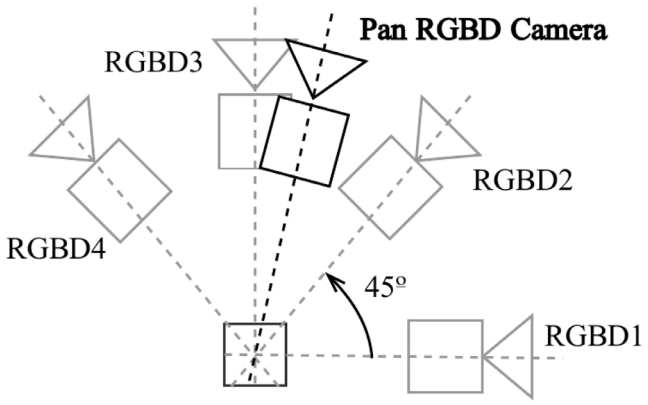


Fig. 8. Proposed camera-configurations for comparison. Fixed camera-configurations are obtained from Robot@Home dataset and replicated to Robot@VirtualHome. RGBD3 represents the typical camera-configuration employed in most state-of-the-art contributions, were the camera points in the direction of the robot path.

4.4. Parameter selection

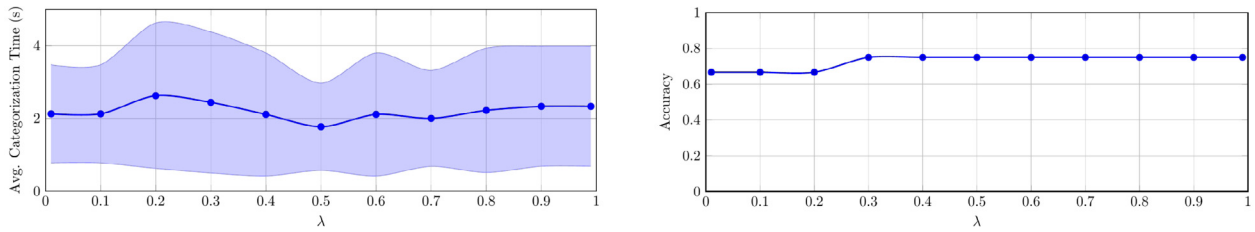
To obtain representative results, while avoiding a high computational overhead, we set the time-horizon of the MDP model as $\gamma = 3$ and the temporal step between successive MDP states to $\Delta t = 1$ s. The latter means that a new camera line-of-sight is selected at a frequency of 1 Hz. Additionally, knowing that our transition function is deterministic (i.e. we assume that the camera always reach the desired orientation since only reachable states are proposed) and the MDP temporal step is 1 s, we set the angle step of the pan unit to $\Delta\theta = 15°$, a value below the rotation speed of the camera (20 degrees per second for our employed cameras) to guarantee that the desired position is always reached while keeping a minimum overlapping between images to avoid information loss.

Related to the weighted entropy in Eq. (9), we set the weights ω_m that assess the relevance of each object class in the inference process attending to their occurrence frequency, as:

$$\omega_m = \bar{N}^m_{places} + p(O_m), \tag{14}$$

where $\bar{N}^m_{places}$ is the number of place categories where an object class O_m is not expected to be found and is computed based on human knowledge. For example, from the set of rooms considered in this work (*kitchen, bedroom, bathroom* and *living room*), a bed is expected to be found only in bedrooms, hence $\bar{N}^{bed}_{places} = 3$. Additionally, $p(O_m)$ is the prior probability which encodes the relative frequency of appearing a certain object class in a household. In practice, this probability is difficult to estimate and hence, typically it is learned from the prior knowledge retrieved from the existing data. Particularly, in this work we extract the prior probability by observing the objects distribution and occurrence frequency in both employed datasets (see Table 1).

Finally, to select the value of the parameter λ controlling the influence of the different contributions to the expected information gain (see Eq. (4)), we carried out an empirically evaluation shown in Fig. 9. In terms of time performance, it can be seen that this parameter has no relevant impact, being the optimal



(a) Average time for successful categorization. (b) Place categorization accuracy.

Fig. 9. Average time for successful categorization and accuracy of object-based place categorization working under BVPC configuration with different values of the information gain parameter λ.

Table 2
Accuracy performance for each experiment and camera-configuration. Best results are marked in bold.

	Camera configuration	Robot@Home				Robot@VirtualHome				Global avg.
		Anto	Alma	Rx2	Avg.	House 12	House 18	House 30	Avg.	
Object-based	BVPC	**87.50%**	**100.00%**	**75.00%**	**88.24%**	**100.00%**	**100.00%**	75.00%	**93.75%**	**90.91%**
	Frontiers [49]	75.00%	**100.00%**	50.00%	76.47%	50.00%	83.33%	75.00%	68.75%	72.73%
	Frontiers [49] + MDP	75.00%	**100.00%**	**75.00%**	82.35%	83.33%	66.67%	75.00%	75.00%	78.80%
	CE	75.00%	85.00%	**75.00%**	77.94%	83.33%	83.33%	**100.00%**	87.50%	82.58%
	RGBD1	62.50%	80.00%	**75.00%**	70.59%	33.33%	66.67%	50.00%	50.00%	60.61%
	RGBD2	68.75%	90.00%	68.75%	75.00%	33.33%	66.67%	25.00%	43.75%	59.85%
	RGBD3	75.00%	75.00%	68.75%	73.53%	50.00%	83.33%	75.00%	68.75%	71.21%
	RGBD4	65.63%	60.00%	**75.00%**	66.18%	66.67%	66.67%	**100.00%**	75.00%	70.46%
Image-based	BVPC	**75.00%**	**80.00%**	**100.00%**	**82.35%**	**83.33%**	**66.67%**	**75.00%**	**75.00%**	**78.79%**
	Frontiers [49]	62.50%	60.00%	50.00%	58.82%	66.67%	**66.67%**	50.00%	62.50%	60.61%
	Frontiers [49] + MDP	62.50%	**80.00%**	50.00%	64.71%	66.67%	**66.67%**	50.00%	62.50%	63.64%
	CE	50.00%	60.00%	75.00%	58.82%	66.67%	**66.67%**	50.00%	62.50%	60.61%
	RGBD1	62.50%	60.00%	75.00%	64.71%	50.00%	50.00%	25.00%	43.75%	54.55%
	RGBD2	62.50%	40.00%	75.00%	58.82%	66.67%	16.67%	25.00%	37.50%	48.49%
	RGBD3	62.50%	**80.00%**	75.00%	70.59%	66.67%	**66.67%**	50.00%	62.50%	66.67%
	RGBD4	62.50%	40.00%	50.00%	52.94%	66.67%	**66.67%**	50.00%	62.50%	57.58%

values in the range [0.4–0.7], approximately. However, looking to Fig. 9(b), it can be observed that low λ values are detrimental to the categorization accuracy. The latter indicates that the exploratory contribution of the expected information gain should not be underrated, avoiding situations where the algorithm focuses too much on previously detected objects, not exploring the whole environment and thus recognizing a smaller number of them. Therefore, although the exact value λ is not particularly important, it is recommended to be within the range [0.4–0.7]. In this work we select the value $\lambda = 0.5$. It must be stressed that this range only applies to house environments with no significant differences in room dimensions, being necessary to repeat a similar analysis for environments of different nature.

5. Experimental results

5.1. Evaluation of place categorization accuracy

Table 2 summarizes the accuracy results for the proposed camera configurations and testing environments under both place categorization paradigms. As can be seen, our approach outperforms other camera-configurations in almost any environment, improving a ∼8.33% the average accuracy for object-based paradigm w.r.t. the second-best configuration (CE), while for image-based, BVPC exceeds the top-2 configuration (RGBD3) by a ∼12.12%. These results demonstrate that the orientation of the camera plays an important role during place categorization. Furthermore, results obtained from Frontiers compared to Frontiers + MDP shows that exploiting the short-time estimated navigation path of the robot through the MDP contributes to obtain more representative views for both place categorization paradigms.

Moreover, Table 2 illustrates that moving the camera actively (BVPC, Frontiers [49], Frontiers [49] + MDP and CE) increases the overall performance of object-based methods, as a larger area of the environment is observed (see Fig. 10), allowing the recognition of a greater number of objects. However, this fact is not reflected in image-based methods, in which the focus is on maximizing the area observed per frame, seeking to avoid non-informative frames. In this sense, RGBD3 shows better results than moving actively the camera without appropriate constraints (Frontiers) or even without constraints as does CE. The latter is because both Frontiers and CE maximizes the global observed area of the environment (see Fig. 11), but not the observed area per frame, leading sometimes to non-informative frames such as walls, while RGBD3 is oriented in the direction of the robot movement, so frames are generally more appropriate.

Finally, in terms of accuracy, we can see that under identical conditions of camera-configuration and environment, object-based paradigm shows better results than image-based. This fact is explained because object-based exploits contextual relations between objects and rooms and hence, more information is taken into account during the inference process, while image-based just consider the information included in the processed frame.

5.2. Analysis of time performance

For a comprehensive efficiency evaluation, we measure the required time for a successful categorization of a place (see Table 3). Notice that wrong or inconclusive categorizations are not taken into account, so we also indicate in brackets the average accuracy per room category. Results for both place categorization paradigms illustrates that our approach generalizes well for environments of different characteristics (e.g. real/synthetic, small/large, etc.), obtaining, on average, the shortest time for a

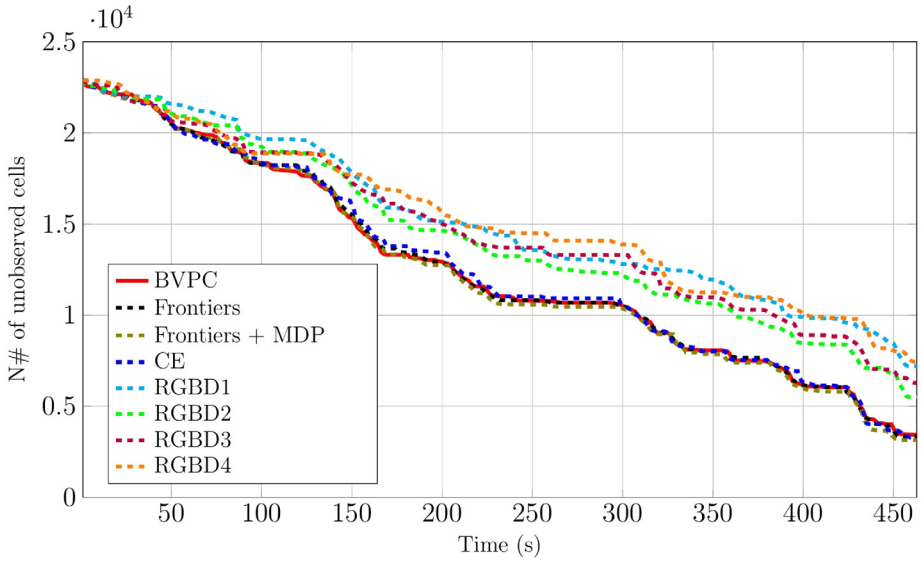


Fig. 10. Temporal evolution of the remain unobserved cells for each camera-configuration in Anto (Robot@Home) under object-based paradigm. Results shows that moving the camera actively (BVPC, Frontiers [49], Frontiers [49] + MDP and CE) allows to maximize the observed area of the environment, which increases the possibility of observing a greater number of objects.

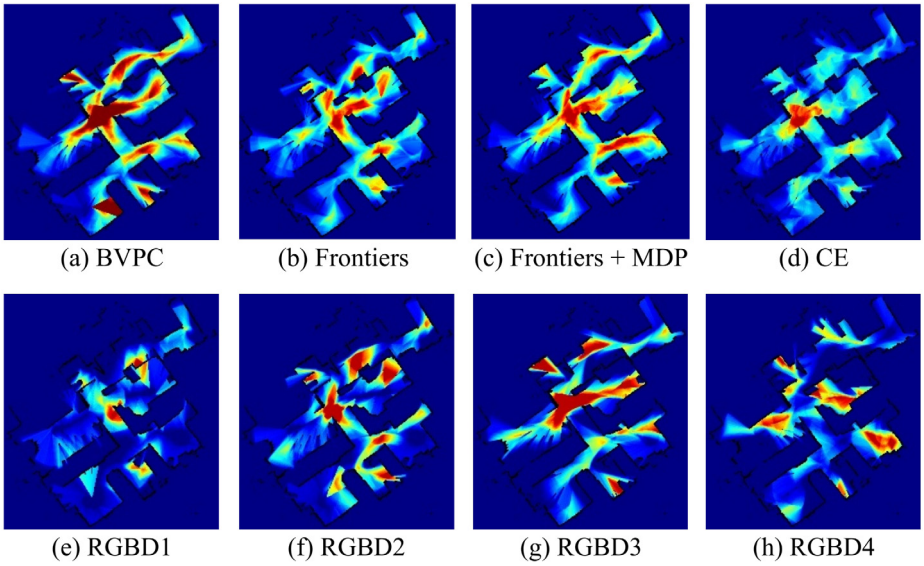


Fig. 11. Representation of the most observed areas for each camera configuration working under the image-based paradigm in the household Anto (Robot@Home) through heatmaps. Blue color in a cell indicates low number of observations and red color indicates high number of observations. BVPC tend to cover general views of the environment while Frontiers [49] and Frontiers [49] + MDP seek to discover the maximum space of the environment. Concerning typical camera-configurations, CE shows a similar number of observations for the whole environment while the fixed camera-configurations tend to concentrate their views in certain spaces of the environment.

successful categorization and the highest accuracy. However, it can be noticed that sometimes a fixed camera-configuration can be faster than BVPC. This fact is more common under object-based paradigms and it is usually associated to a low categorization accuracy, meaning that only a few samples have been taken into account to compute the categorization time, while most of the times the categorization fails. In these rooms, as the robot step in, a relevant object happens to be inside the FOV of the camera, so the object is continuously observed. Naturally, this fact does not always occurs, because objects are not always placed in the same location for a given place category (e.g. a toilet is not always in the left side of the bathroom). Thus, analyzing a greater number of rooms, this fact is compensated.

Moreover, from the results for object-based in Table 3 we can observe how inspecting a larger area of the environment, and so recognizing a greater number of objects, is not sufficient for an efficient place categorization. In the case of CE and Frontiers, accuracy increases but also does the average required time. Due to the maximization of the observed space from CE and Frontiers, it increases the possibility to recognize more objects, yet it does not implement mechanisms to provide robustness to their classification. The latter makes difficult to discern between a false and a correct detection, being both equally considered in the inference process of the place category. This fact, also common for fixed camera-configurations, can be seen in Fig. 12, where a false detection cause fluctuations in the inference process of the room category. This leads to longer times for a successful place categorization, or, in the worst case, to failure.

Looking to Fig. 12, we demonstrate that the Bayesian filtering in BVPC provides robustness to object classes, which contributes

Table 3
Temporal evaluation of the average required time for successful place categorization of each experiment (wrong or inconclusive categorizations are not taken into account). All time measures are in seconds. In brackets, we indicate the average accuracy for each place category. Best results based on time are marked in bold.

		Bedroom	Bathroom	Living Room	Kitchen	Avg.
Object-based	BVPC	**6.04 s (92.31%)**	3.75 s (100.00%)	7.05 s (83.33%)	3.24 s (100.00%)	**5.02 s (90.91%)**
	Frontiers [49]	6.18 s (84.62%)	3.29 s (100.00%)	26.50 s (33.33%)	3.50 s (66.67%)	6.75 s (72.73%)
	Frontiers [49] + MDP	6.09 s (91.67%)	1.71 s (87.50%)	24.50 s (33.33%)	5.67 s (100.00%)	6.23 s (78.80%)
	CE	7.38 s (91.67%)	2.53 s (100.00%)	16.56 s (37.50%)	9.94 s (100.00%)	9.10 s (82.58%)
	RGBD1	12.46 s (75.00%)	3.20 s (62.50%)	**1.00 s (16.67%)**	10.99 s (83.33%)	6.91 s (60.61%)
	RGBD2	9.21 s (70.83%)	**1.48 s (62.50%)**	9.35 s (41.67%)	**1.92 s (62.50%)**	5.49 s (59.85%)
	RGBD3	9.31 s (66.67%)	4.66 s (100.00%)	1.91 s (45.83%)	6.89 s (79.17%)	5.69 s (71.21%)
	RGBD4	6.72 s (75.00%)	5.66 s (100.00%)	34.60 s (20.83%)	6.86 s (83.33%)	13.46 s (70.46%)
Image-based	BVPC	**1.50 s (83.33%)**	**1.00 s (50.00%)**	1.33 s (100.00%)	**1.17 s (100.00%)**	**1.25 s (78.79%)**
	Frontiers [49]	4.20 s (83.33%)	2.25 s (50.00%)	4.00 s (33.33%)	2.75 s (66.67%)	3.50 s (60.61%)
	Frontiers [49] + MDP	3.90 s (83.33%)	2.00 s (50.00%)	4.33 s (50.00%)	1.00 s (66.67%)	3.05 s (63.64%)
	CE	2.75 s (66.67%)	2.00 s (50.00%)	1.50 s (33.33%)	3.00 s (100.00%)	2.31 s (60.61%)
	RGBD1	4.67 s (75.00%)	2.75 s (50.00%)	**1.00 s (16.67%)**	3.50 s (66.67%)	2.98 s (54.55%)
	RGBD2	1.86 s (58.33%)	1.50 s (33.33%)	17.00 s (33.33%)	4.25 s (66.67%)	6.15 s (48.49%)
	RGBD3	1.67 s (75.00%)	1.33 s (50.00%)	2.25 s (66.67%)	1.80 s (83.33%)	1.76 s (66.67%)
	RGBD4	7.50 s (83.33%)	1.25 s (16.67%)	3.50 s (66.67%)	1.33 s (50.00%)	3.40 s (57.58%)

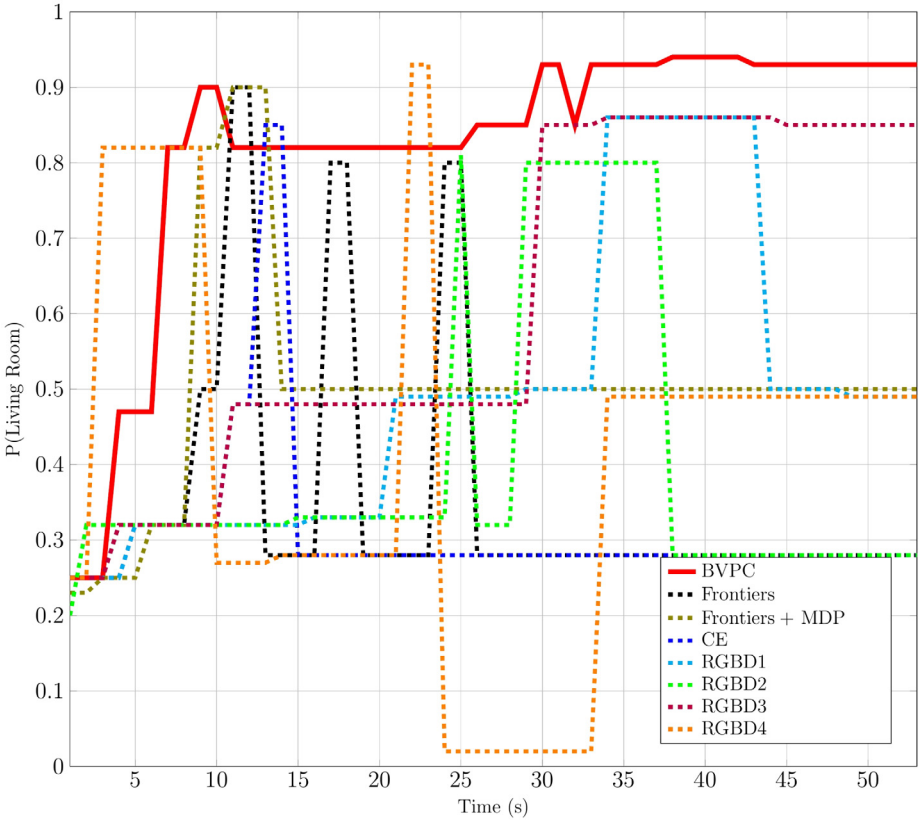


Fig. 12. Temporal overview of the inference process of the Living Room from Anto (Robot@Home) with object-based place categorization under the evaluated camera-configurations. BVPC is the faster configuration that categorizes well the environment, followed by RGBD3. The rest of configurations suffers from high fluctuations in the inference process, leading to inconclusive (Frontiers [49] + MDP, RGBD1 and RGBD4) and failure (Frontiers [49], CE and RGBD2) results.

to reduce fluctuations during the categorization, leading to an efficient place categorization. This fact is also observable between fixed cameras, as RGBD3 is oriented in the direction of the robot's movement, while the robot moves forward without rotating, the camera observes continuously the same objects, providing robustness to its object class. It implies a better performance (trade-off between quickness and accuracy) than others fixed camera-configurations.

Examining the results obtained for the image-based paradigm in Table 3, we can see that these methods are able to quickly categorize a room provided that the processed frames contain representative features of the scene. The latter is demonstrated in the results obtained for BVPC, which includes the proposed attention mechanism to maximize the area observed per frame, being 1.41× faster than the top-2 camera-configuration (RGBD3). Furthermore, comparing both place categorization paradigms, image-based tend to be faster than object-based for equivalent experiments. This fact is explained for the limited information employed for image-based while object-based manage more information and hence, require extra acquisition and computation time.

6. Conclusion and future work

In this work we presented an attention mechanism based on active perception to improve the efficiency (quickness and accuracy) of place categorization methods. To do so, our proposal selects the most informative line-of-sight of the robot's camera at each time moment by controlling a pan-only unit. The optimization is carried out in terms of information maximization, formalized as a next-best-view problem through a Markov Decision Process (MDP) model, which exploits the short-term robot navigation path planning in order to anticipate the next robot poses, being able to make consistent decisions.

We have discussed how the valuable information depends on the place categorization paradigm, proposing solutions for object-based and image-based methods. Results over multiple environments of heterogeneous characteristics (i.e. small/large, real/synthetic, etc.) demonstrate that the proposed attention mechanism improves the efficiency for both place categorization paradigms. In the case of object-based categorization, the accuracy increases when the observed area is maximized (i.e. when a higher number of objects is detected and taken into account to discern the place category), while the efficiency is also related to the robustness in the object detection. For image-based, maximizing the observed area per frame contributes to reduce the number of non-informative frames, improving the accuracy while reducing the required time for categorization.

As future work, we plan to extend the proposed algorithm from place categorization to semantic mapping, where prior knowledge is highly reduced. We will also explore the consideration of new parameters such as the camera's zoom or a pan–tilt unit.

CRediT authorship contribution statement

Jose Luis Matez-Bandera: Methodology, Software, Data curation, Formal analysis, Investigation, Writing – original draft. **Javier Monroy:** Methodology, Validation, Supervision, Writing – review & editing. **Javier Gonzalez-Jimenez:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the research projects WISER (DPI2017-84827-R) and ARPEGGIO (PID2020-117057), as well as by the Spanish grant program FPU19/00704. Funding for open access charge: Universidad de Málaga / CBUA.

References

[1] M. Zhang, G. Tian, Y. Zhang, P. Duan, Service skill improvement for home robots: Autonomous generation of action sequence based on reinforcement learning, Knowl.-Based Syst. 212 (2021) 106605.
[2] M. Luperto, J. Monroy, J.R. Ruiz-Sarmiento, F.-A. Moreno, N. Basilico, J. Gonzalez-Jimenez, N.A. Borghese, Towards long-term deployment of a mobile robot for at-home ambient assisted living of the elderly, in: 2019 European Conference on Mobile Robots, ECMR, IEEE, 2019, pp. 1–6.
[3] A. Ogle, D. Lamb, The role of robots, artificial intelligence, and service automation in events, in: Robots, Artificial Intelligence, and Service Automation in Travel, Tourism and Hospitality, Emerald Publishing Limited, 2019.
[4] L. Lu, R. Cai, D. Gursoy, Developing and validating a service robot integration willingness scale, Int. J. Hosp. Manag. 80 (2019) 36–51.
[5] J.-R. Ruiz-Sarmiento, C. Galindo, J. Gonzalez-Jimenez, Exploiting semantic knowledge for robot object recognition, Knowl.-Based Syst. 86 (2015) 131–142.
[6] J.-R. Ruiz-Sarmiento, C. Galindo, J. Monroy, F.-A. Moreno, J. Gonzalez-Jimenez, Ontology-based conditional random fields for object recognition, Knowl.-Based Syst. 168 (2019) 100–108.
[7] N. Sünderhauf, F. Dayoub, S. McMahon, B. Talbot, R. Schulz, P. Corke, G. Wyeth, B. Upcroft, M. Milford, Place categorization and semantic mapping on a mobile robot, in: Proc. IEEE Int. Conf. Robot. Autom., 2016, pp. 5729–5736.
[8] J.-R. Ruiz-Sarmiento, C. Galindo, J. Gonzalez-Jimenez, Building multiversal semantic maps for mobile robot operation, Knowl.-Based Syst. 119 (2017) 257–272.
[9] C. Galindo, J.-A. Fernández-Madrigal, J. González, Multihierarchical interactive task planning: Application to mobile robotics, IEEE Trans. Syst. Man Cybern. B 38 (3) (2008) 785–798.
[10] Z. Wang, G. Tian, X. Shao, Home service robot task planning using semantic knowledge and probabilistic inference, Knowl.-Based Syst. 204 (2020) 106174.
[11] L.V. Gómez, J. Miura, Ontology-based knowledge management with verbal interaction for command interpretation and execution by home service robots, Robot. Auton. Syst. 140 (2021) 103763.
[12] A. Pronobis, O. Martinez Mozos, B. Caputo, P. Jensfelt, Multi-modal semantic place classification, Int. J. Robot. Res. 29 (2–3) (2010) 298–320.
[13] J. Wu, H.I. Christensen, J.M. Rehg, Visual place categorization: Problem, dataset, and algorithm, in: 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2009, pp. 4763–4770.
[14] M. Brucker, M. Durner, R. Ambruş, Z.C. Márton, A. Wendt, P. Jensfelt, K.O. Arras, R. Triebel, Semantic labeling of indoor environments from 3D RGB maps, in: Proc. IEEE Int. Conf. Robot. Autom, 2018, pp. 1871–1878.
[15] R. Ambruş, S. Claici, A. Wendt, Automatic room segmentation from unstructured 3-d data of indoor environments, IEEE Robot. Autom. Lett. 2 (2) (2017) 749–756.
[16] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: A 10 million image database for scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2017) 1452–1464.
[17] A. Pal, C. Nieto-Granda, H.I. Christensen, DEDUCE: Diverse scene detection methods in unseen challenging environments, in: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, 2019, pp. 4198–4204.
[18] A. Ahmed, A. Jalal, K. Kim, A novel statistical method for scene classification based on multi-object categorization and logistic regression, Sensors 20 (14) (2020) 3871.
[19] D. Fernandez-Chaves, J.-R. Ruiz-Sarmiento, N. Petkov, J. Gonzalez-Jimenez, From object detection to room categorization in robotics, in: Proc. 3rd Int. Conf. Appl. Intell. Syst., 2020, pp. 1–6.
[20] M. Mancini, S.R. Bulò, E. Ricci, B. Caputo, Learning deep NBNN representations for robust place categorization, IEEE Robot. Autom. Lett. 2 (3) (2017) 1794–1801.
[21] J.-R. Ruiz-Sarmiento, C. Galindo, J. González-Jiménez, Joint categorization of objects and rooms for mobile robots, in: Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst., 2015, pp. 2523–2528.
[22] M. Uschold, M. Gruninger, Ontologies: Principles, methods and applications, Knowl. Eng. Rev. 11 (2) (1996).
[23] R.C. Luo, M. Chiou, Hierarchical semantic mapping using convolutional neural networks for intelligent service robotics, IEEE Access 6 (2018) 61287–61294.

[24] K. Oyebode, S. Du, B.J. Van Wyk, K. Djouani, A sample-free Bayesian-like model for indoor environment recognition, IEEE Access 7 (2019) 79783–79790.
[25] P. Uršič, A. Leonardis, M. Kristan, et al., Part-based room categorization for household service robots, in: Proc. IEEE Int. Conf. Robot. Autom., 2016, pp. 2287–2294.
[26] M. Mancini, S.R. Bulo, B. Caputo, E. Ricci, Robust place categorization with deep domain generalization, IEEE Robot. Autom. Lett. 3 (3) (2018) 2093–2100.
[27] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, 2018, arXiv preprint arXiv:1804.02767.
[28] K.M. Othman, A.B. Rad, An indoor room classification system for social robots via integration of CNN and ECOC, Appl. Sci. 9 (3) (2019) 470.
[29] M. Ghasemi, U. Topcu, Online active perception for partially observable Markov decision processes with limited budget, in: Proc. IEEE Conf. Decis, 2019, pp. 6169–6174.
[30] X. Qi, W. Wang, Z. Liao, X. Zhang, D. Yang, R. Wei, Object semantic grid mapping with 2D LiDAR and RGB-D camera for domestic robot navigation, Appl. Sci. 10 (17) (2020) 5782.
[31] C. Potthast, G.S. Sukhatme, A probabilistic framework for next best view estimation in a cluttered environment, J. Vis. Commun. Image Represent. 25 (1) (2014) 148–164.
[32] J. Delmerico, S. Isler, R. Sabzevari, D. Scaramuzza, A comparison of volumetric information gain metrics for active 3D object reconstruction, Auton. Robots 42 (2) (2018) 197–208.
[33] B. Calli, W. Caarls, M. Wisse, P.P. Jonker, Active vision via extremum seeking for robots in unstructured environments: Applications in object recognition and manipulation, IEEE Trans. Autom. Sci. Eng. 15 (4) (2018) 1810–1822.
[34] B. Rasolzadeh, M. Björkman, K. Huebner, D. Kragic, An active vision system for detecting, fixating and manipulating objects in the real world, Int. J. Robot. Res. 29 (2–3) (2010) 133–154.
[35] Y. Wang, S. James, E.K. Stathopoulou, C. Beltrán-González, Y. Konishi, A. Del Bue, Autonomous 3-D reconstruction, mapping, and exploration of indoor environments with a robotic arm, IEEE Robot. Autom. Lett. 4 (4) (2019) 3340–3347.
[36] J.J. Acevedo, J. Messias, J. Capitán, R. Ventura, L. Merino, P.U. Lima, A dynamic weighted area assignment based on a particle filter for active cooperative perception, IEEE Robot. Autom. Lett. 5 (2) (2020) 736–743.
[37] M.L. Puterman, Markov Decision Processes: discrete Stochastic Dynamic Programming, John Wiley & Sons, 2014.
[38] D. Fernandez-Chaves, J. Ruiz-Sarmiento, N. Petkov, J. Gonzalez-Jimenez, ViMantic, a distributed robotic architecture for semantic mapping in indoor environments, Knowl.-Based Syst. 232 (2021) 107440.
[39] J. Su, H. Zhang, Full Bayesian network classifiers, in: Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 897–904.
[40] J.R. Ruiz-Sarmiento, C. Galindo, J. González-Jiménez, Robot@Home, a robotic dataset for semantic mapping of home environments, Int. J. Robot. Res. (2017).
[41] D. Fernandez-Chaves, J.R. Ruiz-Sarmiento, N. Petkov, J. Gonzalez-Jimenez, Robot@VirtualHome, an ecosystem of virtual environments and tools for realistic indoor robotic simulation, 2022, under review.
[42] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2961–2969.
[43] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, R. Girshick, Detectron2, 2019, [Online]. Available: https://github.com/facebookresearch/detectron2.
[44] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755.
[45] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European Conference on Computer Vision, Springer, 2020, pp. 213–229.
[46] B.A. Griffin, J.J. Corso, Depth from camera motion and object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1397–1406.
[47] L. Wang, S. Guo, W. Huang, Y. Xiong, Y. Qiao, Knowledge guided disambiguation for large-scale scene classification with multi-resolution CNNs, IEEE Trans. Image Process. 26 (4) (2017) 2055–2068.
[48] S. Garg, N. Suenderhauf, M. Milford, Don't look back: Robustifying place categorization for viewpoint-and condition-invariant place recognition, in: 2018 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2018, pp. 3645–3652.
[49] H. Umari, S. Mukhopadhyay, Autonomous robotic exploration based on multiple rapidly-exploring randomized trees, in: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, 2017, pp. 1396–1402.
[50] C. Richardt, Omnidirectional stereo, in: Computer Vision: A Reference Guide, Springer, 2020, pp. 1–4.
[51] Y.-i. Kim, Y.G. Min, P.S. Hee, J. Wun-Cheol, S. Soonyong, H. Tae-Wook, The analysis of image acquisition method for anti-UAV surveillance using cameras image, in: 2020 International Conference on Information and Communication Technology Convergence, ICTC, IEEE, 2020, pp. 549–554.