Engineering Applications of Artificial Intelligence 48 (2016) 173–187

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Robot navigation via spatial and temporal coherent semantic maps

Ioannis Kostavelis [a,*], Konstantinos Charalampous [a], Antonios Gasteratos [a], John K. Tsotsos [b]

[a] Laboratory of Robotics and Automation, Production and Management Engineering Department, Democritus University of Thrace, Vas. Sophias 12, GR-671 00 Xanthi, Greece

[b] Electrical Engineering and Computer Science Department, York University, 4700 Keele Street, Toronto, Ontario, Canada M3J 1P3

ARTICLE INFO

Article history:
Received 23 January 2015
Received in revised form
7 October 2015
Accepted 12 November 2015
Available online 2 December 2015

Keywords:
Semantic mapping
Robot navigation
Augmented navigation graph
Place recognition
Time proximity

ABSTRACT

The ability of mobile robots to sense, interpret and map their environments in human terms is decisive for their applicability to everyday activities hereafter. Bearing this view in mind, we present here, for the first time, an integrated framework that aims: (i) to introduce a semantic mapping method and (ii) to use this semantic map, as a means to provide a hierarchical navigation solution. The semantic map is formed in a bottom-up fashion, along the robot's course, relying on the conceptual space quantization, the time proximity and the spatial coherence integrated into the *labeled sparse topological map*. A novel time-evolving *augmented navigation graph* determines the semantic topology of the explored environment and the connectivity among the recognized places expressed by the inter-place transition probability. The robot navigation part is addressed through an interface that facilitates human robot interaction. High level orders are passed to the robots and unfolded recursively, in a top-down fashion, into local navigation data. The performance of the proposed framework was evaluated on long range real world data and exhibited remarkable results.

1. Introduction

For mobile robots to be able to enter the market in the near future, they should include the capacity to produce meaningful internal perceptual representations of their own environment, making them able to cope with a range of real-life situations and tasks. In response to this challenge, coordinated research efforts to build cognitive robots to competently perceive and understand their surroundings along with cooperation with humans are underway. In particular, as far as mapping and navigation are concerned, robots should comprehend human concepts about places, to skillfully operate in human environments. With this point of view, *semantic mapping* comprises an active research topic, in which substantial progress has been reported in recent years (Kostavelis and Gasteratos, 2015, 2013; Girdhar et al., 2013). A semantic map is an augmented representation of the robot's environment that – supplementary to the geometrical knowledge – encapsulates characteristics compatible with human conception. Such characteristics might be an abstraction of the explored space, recognition of diverse places, objects or shapes and even a sense of the vicinity of the examined spots (Kostavelis and Gasteratos, 2015). In fact, a semantic map can be more efficient when it is built upon a consistent geometrical one. Therefore, the evolvement of semantic mapping, one might say, is grounded on the prior advancements in *simultaneously localization and mapping* (SLAM), as described in Filliat and Meyer (2003), Meyer and Filliat (2003), and Thrun et al. (2005).

Although SLAM methods are capable of modeling the explored environment with great accuracy, precisely driving robots into specific target positions, they lack the higher level comprehension that would allow them to push the human–robot interaction forward as dictated in Galindo et al. (2008) and Zender et al. (2008).

Towards this end, the proposed work is oriented towards forming consistent bottom-up multilayered semantic maps and exploiting them in hierarchical top-down robot navigation. Therefore, the main contributions of this work are expanded both in the semantic mapping and robot navigation axes, capitalizing on attributes which are described in the following section:

The semantic mapping embraces:

- A topometric mapping framework is introduced, where the explored environment is expressed as a geometrical representation, and abstracted as a topological map.
- A place classification method, relying on histogram based representations of the places to be memorized firstly presented in Kostavelis and Gasteratos (2013) but also described here shortly for the sake of completeness.

* Corresponding author.
E-mail addresses: gkostave@pme.duth.gr (I. Kostavelis), kchara@pme.duth.gr (K. Charalampous), agaster@pme.duth.gr (A. Gasteratos), tsotsos@cse.yorku.ca (J.K. Tsotsos).

http://dx.doi.org/10.1016/j.engappai.2015.11.004
0952-1976/

- An unsupervised place partitioning, which firstly involves a segmentation of the environment into distinguishable annotated places, forming a *labeled sparse topological map* (LSTM) and, secondly, models the transitions among the places on the map using the first order Markov Model.
- A conceptual representation of the detected places is introduced herein for the first time and expressed as an *augmented navigation graph* (ANG).

The robot navigation takes advantage of:

- A graph traversing strategy of the ANG, enabling transmission of high level commands to the robot, so as to move from one memorized place to another, such as "*go to the living room*".
- A global navigation methodology, utilizing the formed LSTM as global path planner, the nodes of which act as landmarks the robot should closely follow.
- Local navigation routines that employ the nodes in the topometric map and recall 3D point clouds to construct 2D occupancy grids for the robot to reach the target location, discussed in Dayoub et al. (2013) but integrated here with ANG and LSTM.

It should be noted that apart from the place classification, which has been proposed in Kostavelis and Gasteratos (2013), the rest of the functional modules mentioned above are novel or utilized in an integrated framework, for the first time, to work as part of a system. An overview of this work is illustrated in Fig. 1. We propose an integrated system, which provides effective solutions for semantic mapping and knowledge-based robot navigation. During the formation of the semantic map, the sole prior knowledge this system employs is a learned visual vocabulary of abstract representations of the possible place categories the robot may visit while wandering around. Along the robot's itinerary, a 3D map of the explored environment is constructed, which is abstracted as a topological map. Simultaneously, the LSTM is built by partitioning the robot's surroundings into places according to their spatiotemporal relations and labels. Considering the temporal proximity a first order Markov model is progressively unfolded and grouped, in proportion to places already learned. The Markov model conducts the formation of the ANG, which expresses the abstraction of the places learned, as well as the transition probability among them. In a nutshell, the proposed multilayered semantic map encompasses spatiotemporal and semantic attributes by seamlessly integrating the low level topometric maps into the LSTM and then into the high level ANG. The constructed semantic map (represented as LSTM and ANG) enables the user to give high level commands, in order to redirect the robot from one mapped place to another. The execution of this action is treated hierarchically firstly by employing a graph traversing algorithm operating on the ANG and detecting the sequence of the places the robot should progressively follow. Subsequently, the respective vertices on the LSTM are triggered to provoke the recall mechanism to actuate the nodes on the topometric map. These nodes evoke the respective 3D point clouds to form a 2D occupancy grid, thus facilitating the robot's local navigation. Notwithstanding that *human machine interaction* (HMI) is out of the scope of this paper, with a view to establish a basic liaison between humans and robots, a dynamically evolved *graphical user interphase* (GUI) has also been developed.

2. Related work

A consistent geometrical map constitutes the cornerstone for an efficient semantic map. However, it should be mentioned that the semantic learning of visual concepts has also been widely used in the image retrieval field providing (Li et al., 2013) a significant assistance in the area of semantic mapping for robots. Progress would not be possible in the area of semantic mapping, without prior advancement in SLAM. Essentially, SLAM provides solutions to the problem according to which, a mobile robot located at any unknown spot in an unexplored environment incrementally builds a consistent metric map, while simultaneously determines its location within this environment map (Thrun et al., 2005). Various research attempts successfully provided remarkable solutions to this problem and an analytical summary is presented in a two-part review paper (Durrant-Whyte and Bailey, 2006; Bailey and Durrant-Whyte, 2006). Several semantic mapping approaches find their roots in SLAM as, except for the knowledge-based characteristics for mapping and navigation, they also embrace geometrical ones, both for indoors and outdoors scenarios. Speaking of indoors scenarios, an attempt to construct a taxonomy of the existing semantic mapping methods results in two categories, according to the scale the geometrical map is expanded to, namely

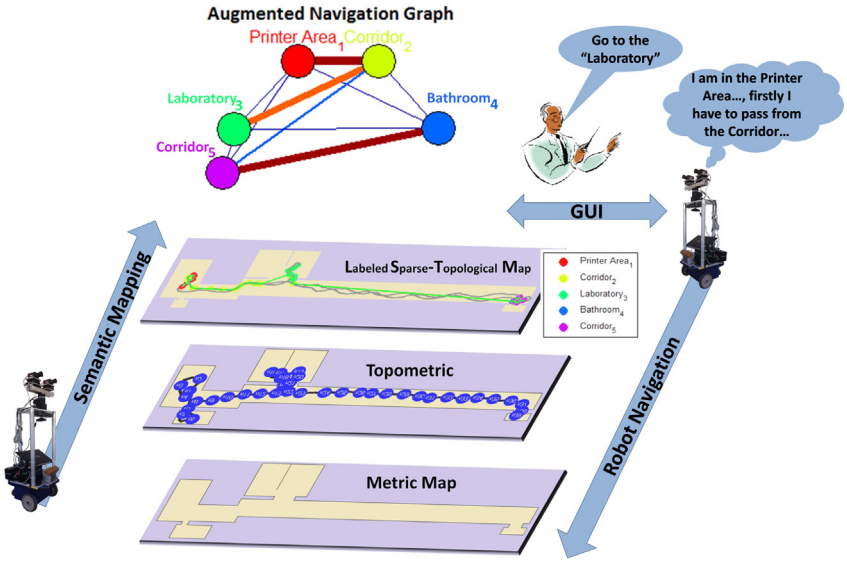

Fig. 1. Overview of the proposed semantic mapping and navigation method.

the single scene maps and the large-scale ones (Kostavelis and Gasteratos, 2015). The first class gathers those methods where reasoning about an instance frame is done with respect to a local coordinate system and, moreover, conceptual attributes about the observed objects in the scene are also provided. An early work is described in Nielsen et al. (2004), where the authors approximate the semantic mapping as an interface between robots and humans. They introduce a single-frame snapshot application as a way to grab real-world pictures and store them with the aim to augment a metric map. In particular the map advancement is accomplished by means of icons or symbols, thus adding significance to places and objects of interest. In a more recent work (Trevor et al., 2015) a single scene point cloud segmentation utilizing connected components through RGB-D[1] data has been introduced. Firstly a planar segmentation step is performed on the point cloud data to distinguish the dominant planes in the scene. Then an L_2[2] norm based clustering and a connected component labeling mask are applied on the color image, in order to detect objects on a tabletop. On the other hand, large-scale approaches progressively construct a metric map, with respect to a global coordinate system, which is simultaneously annotated with high-level features, such as object types, place labels and shape interpretation. A generic pipeline for indoor data processing and semantic information extraction based on 3D point clouds is described in Tamas and Goron (2014). The authors in Blodow et al. (2011) exploited progressively acquired laser scans merged with a 2D-3D registration routine to form the metric map. In further detail, 3D laser data have been exploited, firstly to extract geometrical information about the planes in kitchen environments and secondly to infer the object labels within the organized point clouds. Segmentation techniques are applied to generate initial hypotheses about the significance of objects, such as furniture drawers and doors. The authors in Rusu et al. (2009) and Rusu et al. (2008) augmented the geometrical map by processing large input datasets and extracting relevant objects. The objects modeled are the ones found in a kitchen environment, for instance kitchen appliances, cupboards, tables, and drawers, all having a specific significance for a domestic robot. In Kostavelis and Gasteratos (2013) an novel appearance-based approach was introduced, with the aim to address the place classification issue. Another RGB-D based method (Günther et al., 2015) adopted the SLAM6D toolkit (Nüchter et al., 2007) to register the subordinate point clouds into a consistent full-scene map. The authors in that work introduced an RGB-D based system capable of reconstructing surfaces in the point clouds, detecting different types of furniture and then estimating their poses. The result is a consistent mesh representation of the environment enriched by CAD[3] models corresponding to the detected pieces of furniture.

A primitive intention of the semantic mapping methods is to express the surroundings in terms of significance to humans. Indeed, such methods take advantage of the research efforts made in the area of place and object recognition and categorization (Linde and Lindeberg, 2004; Ullah et al., 2008). The authors in Pronobis et al. (2006) proposed a discriminative approach for place recognition based on visual information. A global descriptor operating directly on the robot acquired images assisted by a *support vector machine* (SVM) to produce appearance based interpretations. This method was embodied on a robot platform and tested for indoors place recognition scenarios, yielding notable results. Towards the same end, the authors in Pronobis et al. (2010a) combined an incremental extension of SVMs with a method that reduces the number of support vectors needed to build the decision function and without any loss in performance. Moreover, in Cadena et al. (2012) a stereo vision place recognition algorithm is presented which considers both appearance and geometrical information of points of interest in the images. Emphasis is given in the loop closure verification, which is addressed by comparing *conditional random fields*. Additionally, in Fazl-Ersi and Tsotsos (2012) *histograms of oriented uniform patterns* were utilized, providing strong discriminative capacities facilitating the solution of the place recognition problem. Moreover, the authors in Milford (2013) proved that the place recognition capabilities of a robot increase when long matching sequences are utilized instead of short ones. Additionally, it is explained how the accuracy of place recognition is affected by the quality of the acquired images.

Object recognition is another very common attribute of the semantic mapping methods. One straightforward technique implemented in Ekvall et al. (2006), was targeted at building a robot that localizes its position in a metric map, recognizes objects on its travel and assigns them accordingly on the map. In a more sophisticated approach (Meger et al., 2008), the authors developed an object recognition algorithm supported by a saliency attentional model. More precisely, in order to perform successful recognition in a real world scenario, they combined peripheral-foveal vision system, a bottom-up visual saliency with structure from stereo and metric mapping. A SIFT-based object recognition system was adopted, while a stereo camera was used to recognize the object and to obtain its coordinates on the metric map. In Ranganathan and Dellaert (2007) a model for place recognition using objects as the basic unit of representation was adopted. Stereo range data were used to compute the 3D locations of the objects, while the *Swendsen–Wang algorithm*, a *Markov chain Monte Carlo* cluster method, was adopted to solve the correspondence problem between image features and objects. In Nüchter and Hertzberg (2008) a similar algorithm based on the place geometry and the object information was presented. In particular, a laser scanner mounted on a mobile robot enables the acquisition of dense point clouds, which are then partitioned and annotated with semantic labels using object recognition techniques. Additionally, the authors in Anand et al. (2013) introduced a method that detects and recognizes objects in 3D point clouds of indoor scenes. This method utilizes various features and contextual relations, obtained from dense point clouds to train a maximum-margin algorithm capable of distinguishing multiple classes of objects with remarkable accuracy.

There is also a sufficient number of applications that exploit not only one, but multiple cues to draw semantic conclusions. These methods might either combine different modalities to apprehend the robot's surroundings or exploit multiple perception facets deriving from the same sensory input to deduce observed scenes. The occurrence frequencies for visual place and object recognition are utilized in Viswanathan et al. (2009) and Aydemirn et al. (2011), in order to construct a spatial-semantic model. The authors in Ko et al. (2015) constructed consistent semantic maps exploiting place categorization accompanied by object matching techniques. Geometrical information is also employed to augment the spatial relationships among the detected objects. In more sophisticated works such as those reported in Pronobis and Jensfelt (2011, 2012) the authors utilize multiple sensors to deduce about different features of the scene. For example place and object recognition is performed through vision, while the shapes of the rooms are extracted by utilizing laser scanners. All these retained concepts are fused under generalized SVM models to produce probabilistic inferences concerning the explored areas.

[1] RGB-D stands for *red green blue and depth*.
[2] L_2 stands for the *Euclidean distance*.
[3] CAD stands for *computer aided design*.

3. Building the semantic map

3.1. Place memorization and recall

The place memorization method refers to the one developed and thoroughly discussed in Kostavelis and Gasteratos (2013), which provides an effective solution to the place classification (recognition and categorization) problem. The place recognition method is deployed in two phases: the first one being executed off-line, whereas the second one on-line.

During the off-line phase, a visual vocabulary is constructed and the SVM models have their parameters optimized via training. In particular, a labeled sequence of images is considered, describing specific type of facilities relative to the ones the robot will come across when it is in normal operation. The *scale-invariant feature transform* (SIFT) (Lowe, 2004) is applied on each single image of the sequence and the detected feature points are concatenated in a framewise manner. This comprises a *bag-of-features* (BoF) model also denoted as bag of visual words (Csurka et al., 2004), which is motivated by the bag of words model assessed in information retrieval to describe a document by its written content. Accordingly, the BoF describes a scene by its pictorial content, thus it retains a substantial description of the entire feature space to be memorized. The feature space $\mathbf{S}$ is then clustered by the Neural Gas algorithm (Martinetz et al., 1991) and the collection of the quantization centers constitute the visual vocabulary, which provides a satisfactory representation of the BoF. The visual words are then utilized to create an appearance based histogram for each image of the sequence. Specifically, given the detected features we form a representative consistency histogram $\mathbf{h}_{S_k} \in \mathbb{R}^{N_Q}$ for each image $k = 1, 2, \ldots, M$ over the N_Q visual words. The L_2 norm between the detected features and the visual words is calculated and the representative histogram is formed by increasing the corresponding bin by one, according to the smaller distance. All the computed histograms, labeled accordingly, are utilized as training samples for the SVM classifiers, trained following the one-versus-all strategy. Consequently, the number of SVM classifiers needed in order to employ the aforementioned strategy becomes N_C, which equals the number of classes. Each SVM classifier derives a separation hyperplane between a class C_j and the remaining ones, labeled as "1" and "−1", respectively. This scheme derives the following N_C separating functions:

$$g^j(\mathbf{h}_{s_t}) = \sum_{k=1}^{N_{sv}^j} \lambda_k^j y_k^j \mathbf{K}(\mathbf{h}_{s_k}, \mathbf{h}_{s_t}) + b^j, \quad j = 1, 2, \ldots, N_C \tag{1}$$

where $\lambda_k{}^j$ are the Lagrange multipliers, $y_k{}^j$ the labels of the training samples, $\mathbf{K}(\cdot,\cdot)$ the kernel product of the training samples and the testing one, b^j the bias term and N_{sv}^j the number of support vectors of the corresponding class j. The main advantage of Neural Gas is that its update scheme is a global one and, consequently, local minimum solutions are avoided. Although this solution increases the computational burden of our system, this algorithm is utilized only once during the off-line phase to produce the representative visual vocabulary from the BoF.

During the on-line phase which is the construction of the semantic map, the system makes inferences about the places lately visited, by retaining the visual vocabulary and the trained SVM models, as they produced during the off-line memorization phase. Particularly, as the robot explores an unknown environment, each frame s_t at time t is converted into an appearance based histogram $\mathbf{h}_{s_t}$ utilizing the pre-computed visual vocabulary. Then, the SVM models $g^j(\cdot)$ are queried with each histogram and the selection of the current class is performed relying on the maximum marginal distance of the hyperplane (Chang and Lin, 2011), i.e. $\max\{g^1(\mathbf{h}_{s_t}), g^2(\mathbf{h}_{s_t}), \ldots, g^{N_C}(\mathbf{h}_{s_t})\}$, $j = 1, 2, \ldots, N_C$. The utilized methodology is illustrated in Fig. 2, where the off-line memorization and the on-line inference mode are highlighted.

Since the developed methodology is applicable to indoors, especially university environments, the utilized database concerns corresponding types of localities only, such as "corridors", "laboratories", "meeting-rooms", "offices", etc. This might be a limitation for the proposed method, considering that the system will not be able to distinguish a new location, unless it is trained for it. However, this disadvantage can turn to an advantage elsewhere, since the proposed system is capable of generalizing and inferring on similar facilities for which it possesses no prior knowledge. This is due to the fact that each frame is transformed into an appearance based representation that holds great separability capacities, which assists the unsupervised place clustering procedure.

3.2. Metric and topological mapping

During the robot's roaming in an unexplored environment a SLAM algorithm is executed, employing an RGB-D sensor only. The SLAM procedure employs the robot's incremental motion estimations and the registration of the computed transformations on the 3D point clouds with reference to a global coordinate system. The motion estimation among the successive frames is performed by 3D feature tracking, with matches being filtered by means of an outlier detection procedure (Kostavelis et al., 2013). The rigid body transformation between the consecutive time instances is solved by *singular value decomposition* (SVD). This transformation is applied on the respective 3D point clouds, resulting in a well-shaped 3D map. With the view to revise the 3D map to a more consistent representation, a *random sample consensus* (RANSAC) plane detection step is applied, which detects the dominant planes among the consecutive point clouds. The points that belong to detected planes are matched using the *iterative closest point* (ICP) algorithm providing a more solid artwork, similar to the method

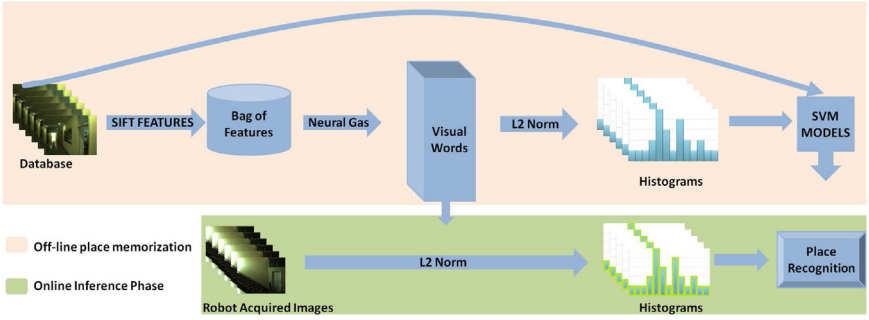

Fig. 2. The upper figure describes the off-line memorization procedure, while the lower figure depicts the on-line procedure that operates in each frame acquired by the robot along the normal operation. The inference mode involves the formation of the appearance based histograms using the pre-computed vocabulary and the L_2 norm for the SVM classification.

described in Kostavelis and Gasteratos (2013). The output of the 3D SLAM algorithm is illustrated in Fig. 3.

The aforementioned metric map can be abstracted by a topological map, i.e. a graph $G^{TM}=(V^{TM},E^{TM})$, where V^{TM},E^{TM} is the set of vertices and edges, respectively. The vertices of this graph bear both semantic and geometrical attributes. Each vertex is added on the topological map on the condition that a minimum distance T_δ has been covered. With the view to accomplish this constraint, the robot's location (x_r,y_r) is estimated in each step and a new vertex is added to the topological graph *iff* $\forall v_i \in V^{TM}$: $\sqrt{(x_r-x_{v_i})^2+(y_r-y_{v_i})^2} > T_\delta$, such as $V^{TM'}=V^{TM}\cup\{v_{|V^{TM}|+1}\}$ and $E^{TM'}=E^{TM}\cup\{e_{|V^{TM}|,|V^{TM}|+1}\}$. That is, the current position of the robot and the last created node should not exceed a specific distance $T_\delta=1$ m, which is empirically estimated in this work and measured on the metric map. Moreover, each node of the topological map is registered with a specific point cloud which is centered around this vertex with respect to the global pose of the robot. This point cloud is utilized to facilitate the robot's local navigation (see Section 4.2). Additionally, each vertex on the topological map bears a specific label tagging the category the place currently being visited belongs to. This is achieved by setting forth queries to the SVM-based place recognition algorithm about the area observed by the robot. The place label is also considered during the global navigation of the robot. Therefore, for each node on the topological map, the respective point cloud and the place category are stored. An example of this topometric (topological and metric) map with the respective semantic information is illustrated in Fig. 4.

3.3. Unsupervised place partitioning

On top of the topological map we have introduced the LSTM, which models both the *spatial* and *temporal* information, so as to pave the way for the semantic interpretation of the explored area. This layer utilizes the memorized visual vocabulary and the robot's pose to express the environment as an abstract representation, supporting two functionalities: (i) to partition the places visited according to their category label and (ii) to exploit their temporal proximity, resulting from the robot's motion, to conceptualize the physical constraints of the environment. Particularly, during the robot's motion, the places are notionally divided into classes without strict spatial boundaries and the transitions among the characterized classes are modeled via their temporal proximity. This layer operates as an intermediate communication channel between humans and robots, so that robots can eventually comprehend high level commands from humans. Following the architecture described in Fig. 1, this module evolves simultaneously with the metric and topological layers and is described in more detail in the rest of this section.

3.3.1. Labeled sparse topological map

The LSTM is exposed to the successive frames and memorizes a subset of the appearance based histograms up to the end of the robot's itinerary. The subset of such histograms, eventually added to the LSTM, corresponds to the vertices of this map retaining both geometrical and semantic information. Consequently, a weighted, undirected and symmetric graph $G^{LSTM}=(V^{LSTM},E^{LSTM})$ is formed, where an edge $e_{ij}\in E^{LSTM}$ connecting two nodes $v_i,v_j\in V^{LSTM}$ is computed via $e_{ij}=\sqrt{(x_{v_i}-x_{v_j})^2+(y_{v_i}-y_{v_j})^2}$. The procedure of

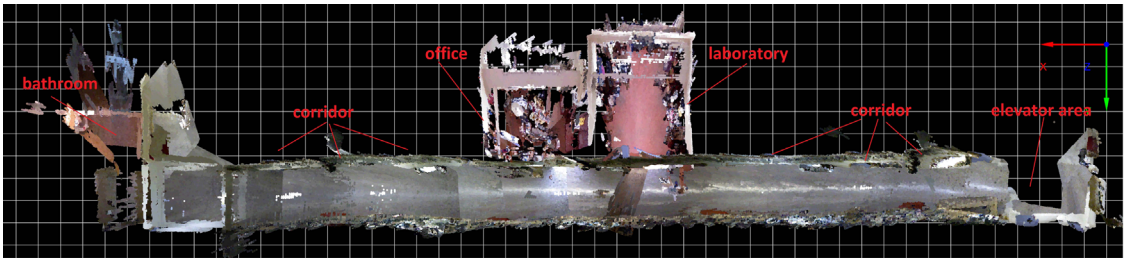

Fig. 3. The 3D metric map of the explored environment, which corresponds to the 2nd floor of the Democritus University of Thrace, Department of Production and Management Engineering. Note that a voxel grid filtering has been applied on the merged point clouds in order to discard duplicated points. It is worth noting that the resolution of the 3D map of the explored environment is less than 1 cm.

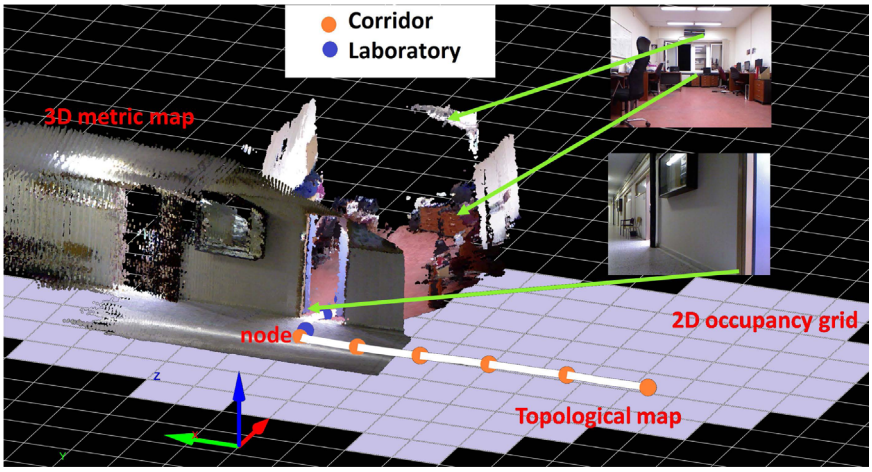

Fig. 4. The topological and metric (topometric) map of the perceived environment. The topological map represents the world as a globally consistent graph. Each graph vertex bears a specific place label according to the inference of the SVM models. Additionally, each node is centered on the occupancy grid constructed using 3D point clouds. These are recalled during the local robot navigation.

adding a new node, i.e. $V^{LSTM'} = V^{LSTM} \cup \{v_{|V^{LSTM}|+1}\}$ is governed by two criteria, both of which have to be met, as follows:

- The extracted histogram $\mathbf{h}_{S_t} \in \mathbb{R}^{N_Q}$, corresponding to the frame at time t, is checked against $\mathbf{h}_{S_{v_i}}$ of existing nodes, i.e. $L_2(\mathbf{h}_{S_t}, \mathbf{h}_{S_{v_i}})\ \forall v_i \in V^{LSTM}$ and taken as different in case the respective L_2 norm is found greater than a predefined threshold T_{LSTM}. In this case $\mathbf{h}_{S_t}$ is marked as a candidate to be retained by LSTM as a new vertex $v_{|V|+1}$ otherwise, it is ignored.
- A voting procedure of the SVM models $g^j(\cdot), j = 1, 2, \ldots, N_C$ (described in Section 3.1) endorsing the place label. The w neighbors of the $\mathbf{h}_{S_t}$ participate in a majority vote process with aim to assist the inference about the place label of the new candidate. This constraint exploits the time proximity of the successive frames during the robot's exploration and boosts confidence in the place categorization, as follows:

$$\max_{i=t-w/2}^{t+w/2} \left\{ \max_{j=1}^{N_C} \left\{ g^j(\mathbf{h}_{S_i}) \right\} \right\} \tag{2}$$

Each vertex $v_i \in V^{LSTM}$ in the G^{LSTM} is registered with the current location estimation of the robot. A unique attribute of this graph is that a new vertex is added each time the appearance based information, i.e. the set of nodes in the LSTM, is not sufficient to provide the semantics of the explored environment. Particularly, a robot operating in a large room might need to create more than one vertex on the LSTM to outline the semantic attributes of this place, due to the unevenness of the visual information within it. Let us consider the case of a laboratory, with a lab bench and a whiteboard in different spots. Although the two entities exhibit great appearance dissimilarities, they both signify the same type of place. Therefore, this knowledge should be stored within the existing appearance based histograms of the LSTM. The main difference from the standard topological map described in Section 3.2 is that the LSTM is a sparse one, the sparsity of which refers to the geometrical non-uniformity it retains. More precisely, whereas in the topological map each node is added whenever the robot has traveled a certain distance, in the LSTM a new vertex is added only when the existing vertexes are insufficient to comprehend the explored area. Another point is that the lower the edge weights $e_{ij} \in E^{LSTM}$ the richer the content of the area in terms of appearance and visual variation. Moreover, the nodes classified in the same class according to Eq. (2), describe a specific area providing semantic attributes in the LSTM. Fig. 5 illustrates a conceptual example where the resulting LSTM is depicted against the topological map, during the robot's pass from a corridor to an elevator area. Note that new vertices are added in the LSTM only when the surroundings exhibit a significant visual variation as compared to ones in previous instances. Later, in Section 3.4 we will explain how the LSTM is decomposed into multiple interconnected subgraphs, each of which describes a distinguishable place in the explored environment and exhibiting excessive spatial coherence.

3.3.2. Markov model based place transitioning

The propagation of the temporal information presented in this section aims to retrieve the existing relationships among the V^{LSTM} nodes of the G^{LSTM} graph. It requires the assumption that given a frame at time t, having a histogram $\mathbf{h}_{S_t}$ from a class C_i, the following one $\mathbf{h}_{S_{t+1}}$ can be equiprobably assorted to any class $C_j, j = 1, 2, \ldots, N_C$. It therefore forms the stepping stone in finding and quantifying the temporal proximity among places. A well-known approach to describe the underlying class dependence among the V^{LSTM} nodes is the Markov chain rule. For a given sequence of states (in this case the set of states being $\Omega = \{v_1^{LSTM}, v_2^{LSTM}, \ldots, v_{|V^{LSTM}|}^{LSTM}\}$) the first order Markov model presumes that:

$$P\left(v_{i_t}^{LSTM} \mid v_{i_{t-1}}^{LSTM}, v_{i_{t-2}}^{LSTM}, \ldots, v_{i_1}^{LSTM}\right) = P\left(v_{i_t}^{LSTM} \mid v_{i_{t-1}}^{LSTM}\right) \tag{3}$$

In more detail, at each time t the histogram $\mathbf{h}_{S_t}$ is associated with a vertex $v_{i_t}^{LSTM}$ with the respective proximity measures, as described in Section 3.3.1. Thus, given that the histograms $\mathbf{h}_{S_{t-1}}, \mathbf{h}_{S_{t-2}}, \ldots, \mathbf{h}_{S_1}$ are in accordance with the vertices $v_{i_t}^{LSTM}, v_{i_{t-1}}^{LSTM}, \ldots, v_{i_1}^{LSTM}$, the transition probability at $v_{i_t}^{LSTM}$, at time t solely depends upon the state histogram $\mathbf{h}_{S_{t-1}}$ is derived.

The transition probability matrix, in this paper called the *temporal adjacency matrix* (TAM), is $T \in \mathbb{R}^{|V^{LSTM}| \times |V^{LSTM}|}$. This is a dynamically expanded matrix that follows the Markov chain rule. The latter is justified by the fact that (i) it continuously updates the transition probabilities between the states over the robot's perambulation and (ii) it increases its dimensionality when a new vertex is appended in the graph G^{LSTM}. The aforementioned case implies that if the examination of two successive histograms $\mathbf{h}_{S_{t-1}}, \mathbf{h}_{S_t}$ triggers the nodes v_i^{LSTM} and v_j^{LSTM} then, the respective transition probability $T(i,j)$ is updated. In the case where a new node $v_{|V^{LSTM}|+1}^{LSTM}$ is appended to the graph G^{LSTM}, matrix T expands its dimensionality such as $T \in \mathbb{R}^{|V^{LSTM}|+1 \times |V^{LSTM}|+1}$ and initializes its new elements. The semantic interpretation of this component can be found in the state set Ω and the continuous update of the transition probabilities. The Ω set consists of nodes, essential to represent the traveled space of the robot in terms of visual representation and semantic variability, as defined in previous sections. While the LSTM graph appends nodes based on the appearance histograms, which directly affect Ω, the first order Markov matrix imprints the successive physical transitions between places and quantifies as probabilities the subsequent position when the present one is given.

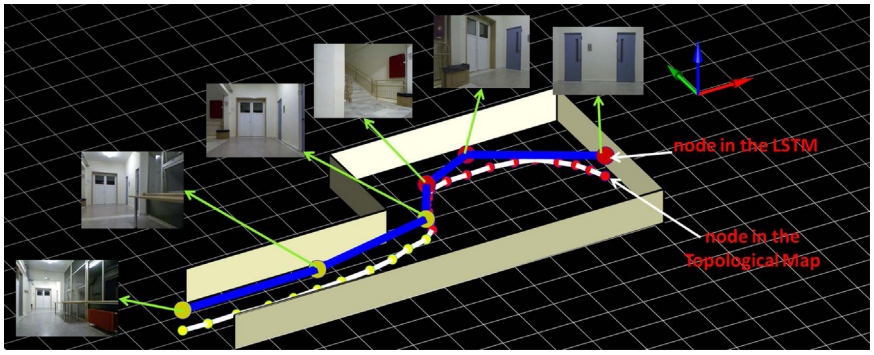

Fig. 5. An example of a LSTM against a topological map during the robot's transition from a corridor to an elevator area. Note that the LSTM is sparser than the topological map, adding new vertices only when the new frame significantly differs (in terms of the L_2 norm of the stored appearance based histograms) from the existing ones. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

We need to note here, that contrary to the early work in Rottmann et al. (2005), where transition among places are modeled by means of hidden Markov models, in the proposed work, we introduced Markov chains into the topological map to further exploit the resulting geometrical constrains of the environment. It is to be shown in the next section that we take advantage of properties of the graphs formed to distinguish among multiple places with the same label, modeling thus the physical constraints of the environment during the robot's travel.

3.4. Augmented navigation graph and place connectivity

In this section we demonstrate the capacity of the LSTM and the first order Markov model to conceptualize the perceived environment as a novel abstract connectivity graph. This is ANG, a weighted graph $G^{ANG}=(V^{ANG},E^{ANG})$ enclosing spatial, temporal and semantic attributes and being defined here for the first time. The term *augmented* is justified by the nature of this graph, due to the fact that vertices correspond to detected places enhanced with labels that comply with the human understanding of the environment, while the edges model the transition probabilities – calculated during the robot's perambulation – between detected places. This conceptualization facilitates high level navigation due to the fact that the structural function of the explored environment is imprinted on the graph and can be made use of in a direct way. For example, in Fig. 6(a), the robot explores an area which consists of a corridor, a laboratory and a meeting room. In order to go from the laboratory to the meeting room (either a robot or a human) should pass through the corridor first. This is a physical constraint concerning connectivity between the two places and is expressed as the transition probability in the ANG (Fig. 6(b)).

The ANG derives from the segmentation of the TAM, which at a given time t encapsulates the respective number of the LSTM nodes, i.e. $|V^{LSTM}|$. However, these nodes are further accompanied with a class label $l_{C_i}, i=1,2,\ldots,N_C$, where their number reveals the least number of nodes in the ANG, that is, $|V^{ANG}| \geq$ *number of classes*. The aim of this procedure is to group together the labeled places exhibiting access-time proximity, while at the same time it estimates the transition probability from one node $v_i^{ANG}\in V^{ANG}$ to the other according to the physical arrangement of the nodes on the LSTM. Particularly, the most recently formed TAM is partitioned into groups with the goal being to divide the set of vertices on the LSTM into spatial and temporal coherent subgroups that correspond to different places. On the one hand, the temporal vicinity according to which the places were accessed is obtained by utilizing the sequential temporal transitions of TAM. On the other hand, the spatial locality of the places is ensured due to the fact that TAM is partitioned into specific groups by taking into consideration the place tagging of the nodes. Each of the groups formed contains LSTM vertices with identical labels, thus belonging to the same place. The nodes that belong to the same group are most likely to exhibit both spatial and temporal adjacency, expressed by both their geometrical arrangement on the LSTM and their aggregated transitions on the TAM. The transition probability among different places (groups of nodes) is estimated by computing the intersection of the transitions between the vertices of different groups in the TAM, that is:

$$e_{ij}^{ANG}=\frac{\sum_{i\in l_{C_i}}\sum_{j\in l_{C_i}}T(i,j)}{\sum\sum T(i,j)} \tag{4}$$

where $e_{ij}^{ANG}\in E^{ANG}$ is the edge connecting the vertex $v_i^{ANG}\in V^{ANG}: v_i^{ANG}\widehat{=}\,l_{C_i}$ to another one $v_j^{ANG}\in V^{ANG}: v_j^{ANG}\widehat{=}\,l_{C_j}$. It is worth noting that each vertex in the graph consists solely of appearance based histograms similarly classified by the SVM models and the weights in the graph denote the transition probability between different groups. The formation of the ANG is also performed upon user's request indicating that even if the robot has not concluded the semantic mapping of the entire environment, the metric, the topological, the LSTM and the TAM are formed progressively over time and, hence, the ANG can be recalled at any time abstracting sufficiently the environment explored so far.

An additional attribute of the ANG is that, during its formation, it simultaneously allows handling more than one place of the same type within the explored environment, as for instance the multiple classrooms in a school building. The latter is achieved by constraining the formation of the ANG to the currently estimated location of the robot. For every node in the ANG graph, the ones corresponding in the LSTM graph are located. A new geometrical graph ${}^{ed}G({}^{ed}V,{}^{ed}E)$ is then formed, with its weights being the respective Euclidean distances (denoted by *ed* throughout this paper):

$$\forall v_i^{ANG} \quad \in V^{ANG}\Longrightarrow \tag{5}$$

$${}^{ed}V: \quad v_j^{LSTM}\subseteq V^{LSTM} \tag{6}$$

$${}^{ed}E:{}_{ed}e_{ij}=L_2\left(v_i^{LSTM},v_j^{LSTM}\right) \tag{7}$$

Thus, the detection of multiple places e.g. "office$_1$" and "office$_2$", that bear the same label is accomplished by applying the *minimal spanning tree* (MST) algorithm (Theodoridis and Koutroumbas, 2008) to the nodes of LSTM. In particular, the MST is applied to the coordinates of the nodes being grouped together during the partitioning of the TAM. The resulting edges connect all vertices according to their minimum Euclidean distance. For cases where an edge holds a value greater than an adaptive threshold T_{ed}, the corresponding group is partitioned and this procedure repeats until no in-between edge is greater than the threshold. The distribution of the intra-node distances is computed and the outliers determine the value of the threshold. The latter implies that if an edge of the MST corresponding to a ${}^{ed}G$ graph results with a value greater that T_{ed}, then a graph cut takes place by

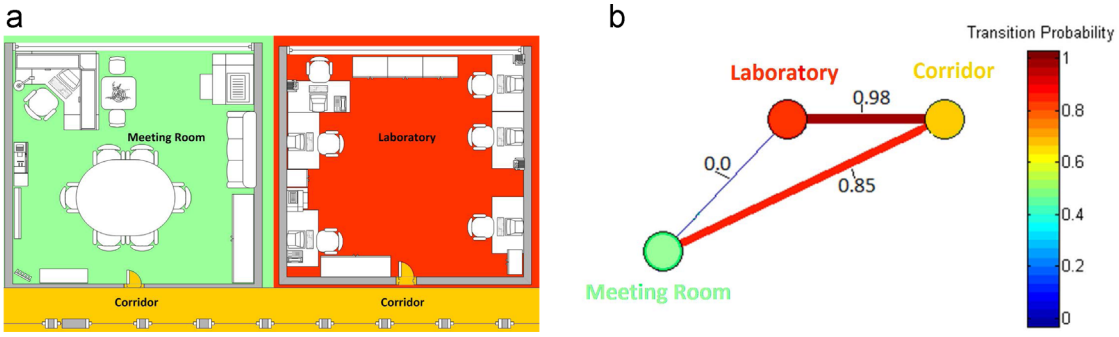

Fig. 6. (a) The top-down projection of a structured environment that comprises a laboratory, a meeting room and a corridor. (b) The ANG of the perceived environment. Note that in the ANG the transition probability among laboratory–corridor and meeting room–corridor is increased revealing the physical connectivity that exists between these two places. However, the direct connection between the laboratory and the meeting room has zero transition probability due to the interfering wall.

removing the corresponding edge. The latter leads to two subgraphs, namely ${}^{ed}G_1({}^{ed}V_1, {}^{ed}E_1)$ and ${}^{ed}G_2{}^{ed}V_2, {}^{ed}E_2)$, where ${}^{ed}V_2 = \overline{{}^{ed}V - {}^{ed}V_1}$. The utilization of the MST ensures the detection of similar places according to the low level geometrical adjacency, while it simultaneously retains the formation of the ANG in case the robot visits the same place more than once. Note that in the example presented in Fig. 5, the blue line indicates the MST among the retained vertices in the LSTM.

4. Robot navigation based on semantic maps

The proposed navigation framework aspires to consolidate the low level local navigation strategies into a high level navigation framework perceivable by humans. This is considered the major novelty of this work, i.e. the construction of a semantic map that preserves both geometrical and semantic attributes, understandable by robots, thus enabling them to perform navigation tasks in a human-like manner. The ANG provides the means to the user to directly pass high level orders to the robot, in order to initiate locomotion. Since the proposed method is an on-line one, the user is able to intervene at any given moment and issue a *go-to* command to the robot, regarding places included already on the semantic map. Following the architecture described in Fig. 1 the robot makes use of the semantic map to navigate in the explored environment. Thus, during the navigation, the semantic map is accessed hierarchically, in a top-down fashion, directing the robot from a global to a local goal.

4.1. Probabilistic place transitioning

The user can only interact with the ANG, since the topological map and the LSTM retain low level information, that is comprehensible only to the robot. The user is constantly aware about the current location of the robot, and can redirect it to any other mapped place by passing specific *go-to* commands. After receiving a specific command, the robot checks with the ANG to find the possible pathways between its current target places. The fact that the ANG retains information on place connectivities enables the robot to plan and execute a conceptual route. Considering the example presented in Fig. 6(b) the robot is unable to move from the "laboratory" to the "meeting room", unless it firstly passes through the "corridor". This high level graph traversing is accomplished using the Dijkstra algorithm (Theodoridis and Koutroumbas, 2008), where the corresponding weights in the ANG are expressed as the inverse transition probabilities. More precisely, the graph on which the lightest path is calculated, is the ${}'G^{ANG} = ({}'V^{ANG}, {}'E^{ANG})$, where ${}'V^{ANG} = V^{ANG}$ and $\forall\, {}'e_{ij}^{ANG} \in {}'E^{ANG} = \frac{1}{e_{ij}^{ANG}}, e_{ij}^{ANG} \in E^{ANG}$. This phase imposes the sequence of places the robot has to pass in order to reach the target goal. Given the detected nodes that were triggered on the ANG during this phase, only the respective vertices in the LSTM are retained after the selection of the target goal. Let $v_i^{ANG}, v_{i+1}^{ANG}, v_{i+2}^{ANG}$ be the triggered nodes of the G^{ANG}, where each one corresponds to a subset of nodes in the G^{LSTM}, i.e. $V_i^{LSTM}, V_{i+1}^{LSTM}, V_{i+2}^{LSTM} \subset V^{LSTM}$. These sets are utilized to form a temporary graph ${}^{ed}G^{temp} = ({}_{ed}V^{temp}, {}_{ed}E^{temp})$, ${}^{ed}V = V_i^{lstm} \cup V_{i+1}^{lstm} \cup V_{i+2}^{lstm}$ and ${}_{ed}E^{temp} : {}_{ed}e_{ij} = L_2(v_i^{temp}, v_j^{temp})$. Therefore, for the concatenated nodes of ${}^{ed}G^{temp}$ corresponding to the sequences of places that the robot has to pass through, the MST is computed. Having in mind that the MST is typically applied on nodes registered with the robot's localization, the minimum cost path that – in geometrical terms – connects the corresponding places is calculated. The formation of such a constrained path is physically explained taking into consideration that the corresponding sequence of places is indeed a feasible route connecting the current and the selected place.

Even though an in-depth consideration of the HRI is out of the scope of this work, as mentioned before, for the sake of completeness a GUI has been developed, retaining information on detected places within the semantic map, as well as the ANG in its current form. As a result, the GUI can graphically illustrate conductivities of existing places and conceptually reports the current location of the robot, in order that the user can select a target place through the appropriate button. Since the GUI is directly updated from the robot's observations, it is a dynamically expanded interface, which automatically adds new nodes to the ANG, as well as the respective buttons when new places are detected. A depiction of the GUI is provided in Fig. 7, summarizing the available information during the robot's itinerary. More precisely, Fig. 7 (a) depicts the explored environment which consists of five detected places, viz. "printer $area_1$", "$corridor_2$", "$laboratory_3$", "$bathroom_4$" and "$corridor_5$", with subscripts indicating the order in which places have been detected. Note that, even if the robot operates in the same area, the system has detected two different corridors e.g. "$corridor_2$" and "$corridor_5$". This can be justified considering that, nodes already retained and describing different spots in the same corridor are separated by noticeable distance in terms of the robot's localization and, thus, the MST criterion has been triggered to differentiate these places as "$corridor_2$" and "$corridor_5$". However, the transitions between the two types of places are increased indicating physical connection between them. In Fig. 7(b) the next phase is presented, where the user has selected the "$laboratory_3$" to be the target place (given that the current one is the "printer $area_1$", see Fig. 7(a). The Dijkstra algorithm is applied on the ANG, where only the vertices linking the current and the target place are retained, i.e. "$corridor_2$" and "$laboratory_3$". That is, in order to move from the printer area to the laboratory, the robot has first to pass through the corridor. Moreover, in the LSTM only the respective nodes that bear the specific semantic attributes have been preserved, while the corresponding MST paths among those nodes have been concatenated, planning a physical route between the two places.

4.2. Global and local navigation through topometric maps

The sequence of nodes in the LSTM corresponding to an estimated physical route operates here as a global path planner. Besides, the local navigation strategy is carried out via the point clouds retained in the nodes of the topological map. For storage economy reasons, points beyond the robot's height have been truncated in the preserved point clouds. Subsequently, a recall mechanism that connects the LSTM and the topological map was developed. More specifically, the topological map and the LSTM are made to share the same semantic information, i.e. each node in the topological map is accompanied by a respective place label, identical to the label of the nodes in the LSTM (see Fig. 5). The latter is achieved owing to the fact that the topological map and the LSTM are evolved simultaneously over time, along the course of the robot. Accordingly, when the user enquires about the physical route between two places, the groups of vertices on the LSTM are concatenated with the respective vertices in the topological map also triggered. Furthermore, during this procedure, vertices that possess the same label and correspond to a similarly identified place such as "$corridor_2$" and "$corridor_5$" (example in Fig. 7(a)) are also triggered. However, these nodes of "$corridor_5$" are not included in the planned physical route (example in Fig. 7(b)) and, therefore, they should be omitted. In order to succeed with this, a simple geometrical thresholding step is applied between the concatenated MST routes and the nodes in the LSTM that describe the respective places "$corridor_2$" and "$corridor_5$".

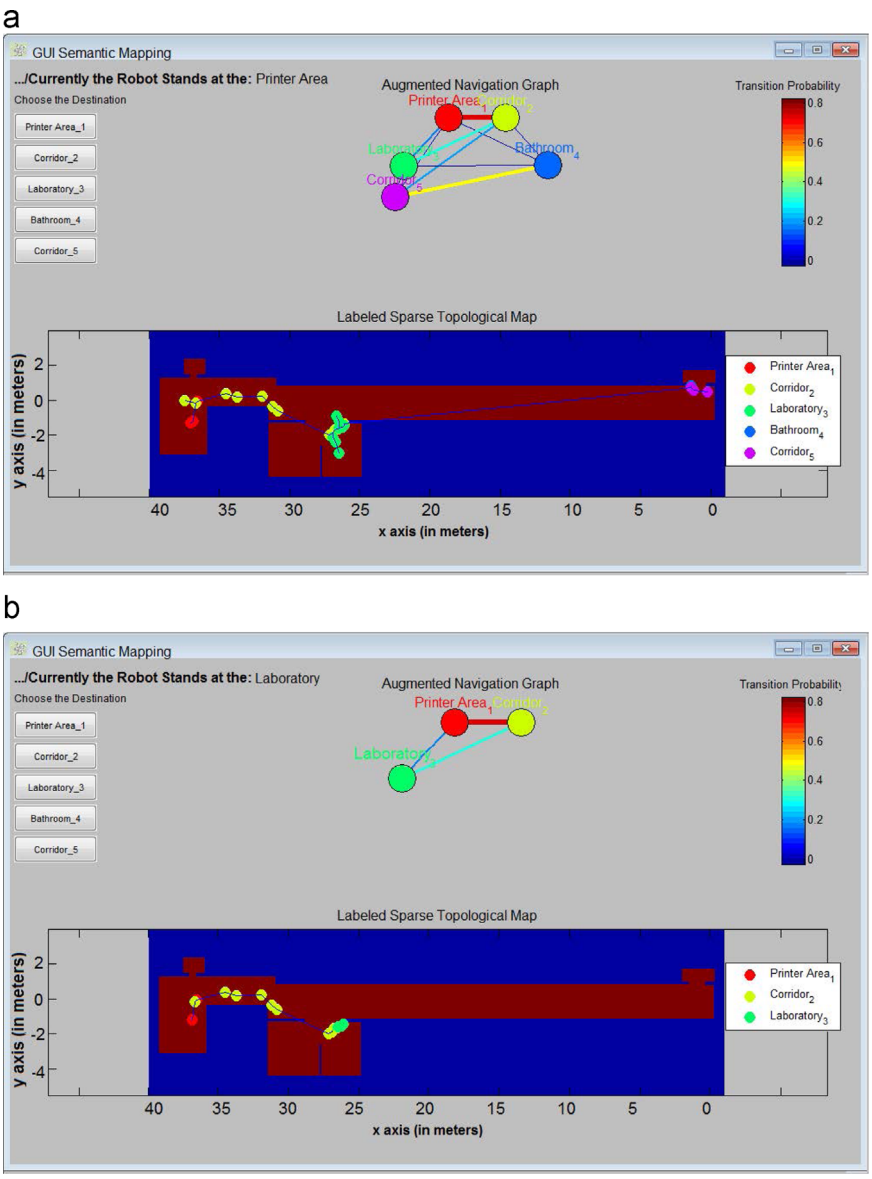

Fig. 7. (a) The GUI containing all the information the user needs to know in order to conceptually redirect to robot from the one place to the other. It comprises annotation about the current location of the robot (see upper left hand corner) in terms of human concepts, the formed ANG and the evolved LSTM superimposed on a metric map. (b) A state of the GUI after the user has selected to redirect the robot into a specific place. The ANG has changed according to the output of the Dijkstra algorithm and the LSTM has been reformed according the conceptual (place label) and the geometrical (MST) constraints.

During robot's local navigation planning, i.e. when a local map is being built about a particular node, a set of nodes is sought on the topological map close to the node the robot stands. A local map around the robot is then built and the truncated point clouds of these nodes are recalled. At this point it should be mentioned that the stored point clouds retain their transformation with respect to the global coordinate system and, therefore, no additional registration step is required upon their recall. In the next step, RANSAC is used to detect and remove the floor plane. All the remaining points are top-down projected and an aggregation step is performed considering a 2D occupation grid of 0.25 m resolution. According to the number scored in the aggregation step, the cells are characterized either "free" or "occupied" (Fig. 8). Mobile robot navigation based on occupancy grids is remains a widely used approach (Maohai et al., 2013). The adopted solution to track the position of the robot in the 2D occupancy grid is the one described in Dayoub et al. (2013), which is considered suitable for robot localization utilizing topometric maps.

4.3. Implementation details

To achieve high level navigation through semantic mapping, described in this work, there are some implementation details that need to be addressed. More precisely, when the robot is prompted from one site to another, it firstly has to exactly locate itself within the metric map. Concerning the semantic part, a frame is acquired, converted into an appearance based histogram and the SVM infers about the label of the place. In case that the robot stands in a place with multiple equivalent tagging, e.g. "office$_1$", "office$_2$", then an additional check is carried out to determine the exact place of the robot. To achieve this, we compute the L_2 norms between the currently computed appearance based histogram and each of the embedded ones within the LSTM nodes. The calculated distances are accumulated and the final selection is performed according to the minimum aggregated distance of the respective group of nodes (note that each group of nodes is a labeled place in the semantic map as described in Section 4.1). Although the robot has semantically determined its location on the map, its absolute pose

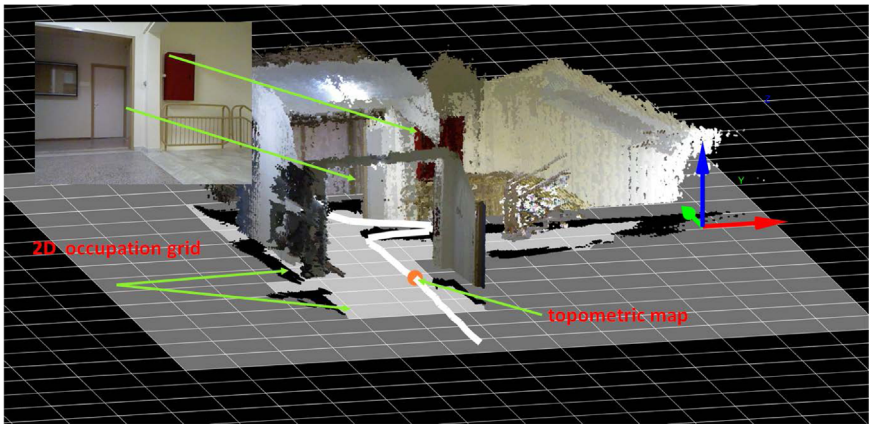

Fig. 8. A node on the topometric map surrounded by the 3D point clouds recalled, which are then converted into 2D occupancy grids – presented here in gray – for the local robot navigation. On top of the occupancy grid the respective 3D point cloud is superimposed. The texture on the point clouds serves for illustration purposes only.

still needs to be determined. Up to this point, only a rough estimation about the spot the robot might be located in has been carried out. In the next step, the respective nodes in the topological graph are triggered and the retained point clouds are recalled. By applying an ICP step between the currently acquired point cloud and the recalled ones, the robot's pose is accurately determined within the global coordinate system of the metric map. Hence, the starting location of the robot has been determined. In order to define the target location the robot should reach, the nodes of the LSTM are activated and the first node of the group selected. The ANG then operates as a conceptual global path planner, the starting and ending position of which are determined by utilizing the topometric map and the LSTM, as described in Section 3.4.

5. Experiments

Within this work, we evaluated the proposed algorithm both on pre-recorded data and on live trials with real robot. Given the necessity to prove the generalization capabilities of the place memorization algorithm, the selection of a dataset with similar place categories to those that the real robot would operate was mandatory. Therefore we have selected the COLD dataset which is a large collection of instances acquired in different universities, namely Freiburg, Ljubljana and Saarbrucken and is extensively described in Ullah et al. (2008). The next dataset on which our algorithm has been evaluated is the Cognitive Navigation one, a detailed description of which could be found in Kostavelis and Gasteratos (2013).

Place classification: In this section we evaluated the classification capabilities of the proposed off-line place memorization module, by comparing the algorithm with other state of the art solutions. For the comparison, we used the Composed Receptive Field Histograms (CRFHs) introduced in Linde and Lindeberg (2004) and utilized then by Pronobis et al. (2006) for place classification, and the local SIFT features combined with SVM employed in Pronobis and Caputo (2007) also for place recognition. Both methods have been implemented and tested on the Ljubljana section of the COLD dataset. The results are summarized and averaged over the various illumination conditions in Table 1, while the performance of the proposed place recognition algorithm used in this work is also shown. It is apparent that the utilized algorithm outperforms the rest of the methods proving that the appearance based histograms retain significant discriminative capabilities with great confidence interval under various illumination conditions.

Table 1
The averaged classification accuracy for the fused datasets including all the existed classes in the COLD Ljubljana sub-dataset for the various illumination conditions. The evaluation procedure comprises a 10-fold cross validation routine considering also the search for the optimal parameters in each case. The respective standard deviations for the 10-fold cross validation procedure are also presented.

Metric results	Method in Pronobis and Caputo (2007)	Method in Pronobis et al. (2006)	Proposed method
Averaged accuracy (%)	85.45	90.34	96.62
Standard deviation (%)	±4.24	±3.16	±3.02

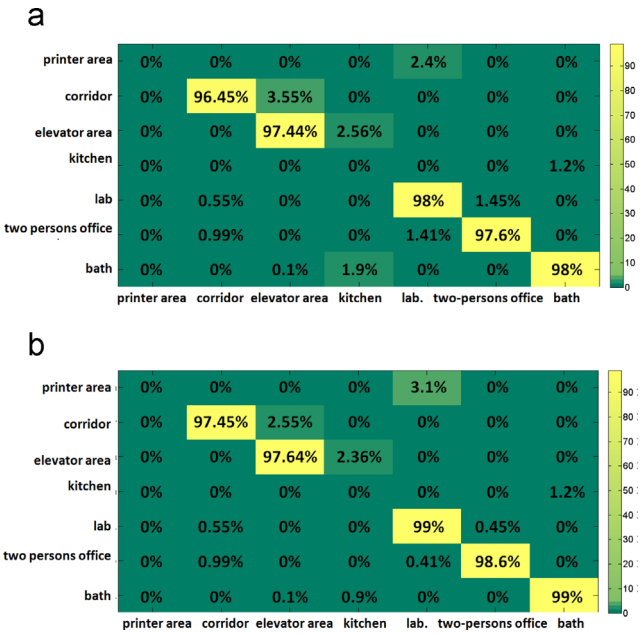

Fig. 9. The confusion matrix for the evaluation of (a) the place categorization algorithm on the Part A and (b) the Part B of the Cognitive Navigation dataset. The exhibited results are averaged via a 10-fold cross validation and the optimal parameter C that penalizes the large errors was found to be equal to 100.

It is essential to evaluate the place recognition algorithm on both known and unknown places. We retained the formulated visual vocabulary and the SVM models formulated on the Ljubljana dataset and we evaluated the categorization accuracy on both Part A (natural illumination conditions) and Part B (artificial illumination condition) of the Cognitive Navigation data, the

Table 2
Conceptual summarization of various semantic mapping approaches.

Examined method	Metric map	Topological map	Place recognition	Temporal coherency	Space segmentation	Navigation facilitation	HRI
Pronobis et al. (2006)	✓	✓					
Martinez Mozos et al. (2007)	✓	✓	✓				
Pronobis et al. (2010b)			✓		✓		
Krishnan and Krishna (2010)	✓				✓		
Blodow et al. (2011)	✓		✓		✓		
Case et al. (2011)	✓					✓	
Pronobis and Jensfelt (2012)	✓	✓	✓	✓	✓		
Kostavelis and Gasteratos (2013)	✓	✓	✓				
This work	✓	✓	✓	✓	✓	✓	✓

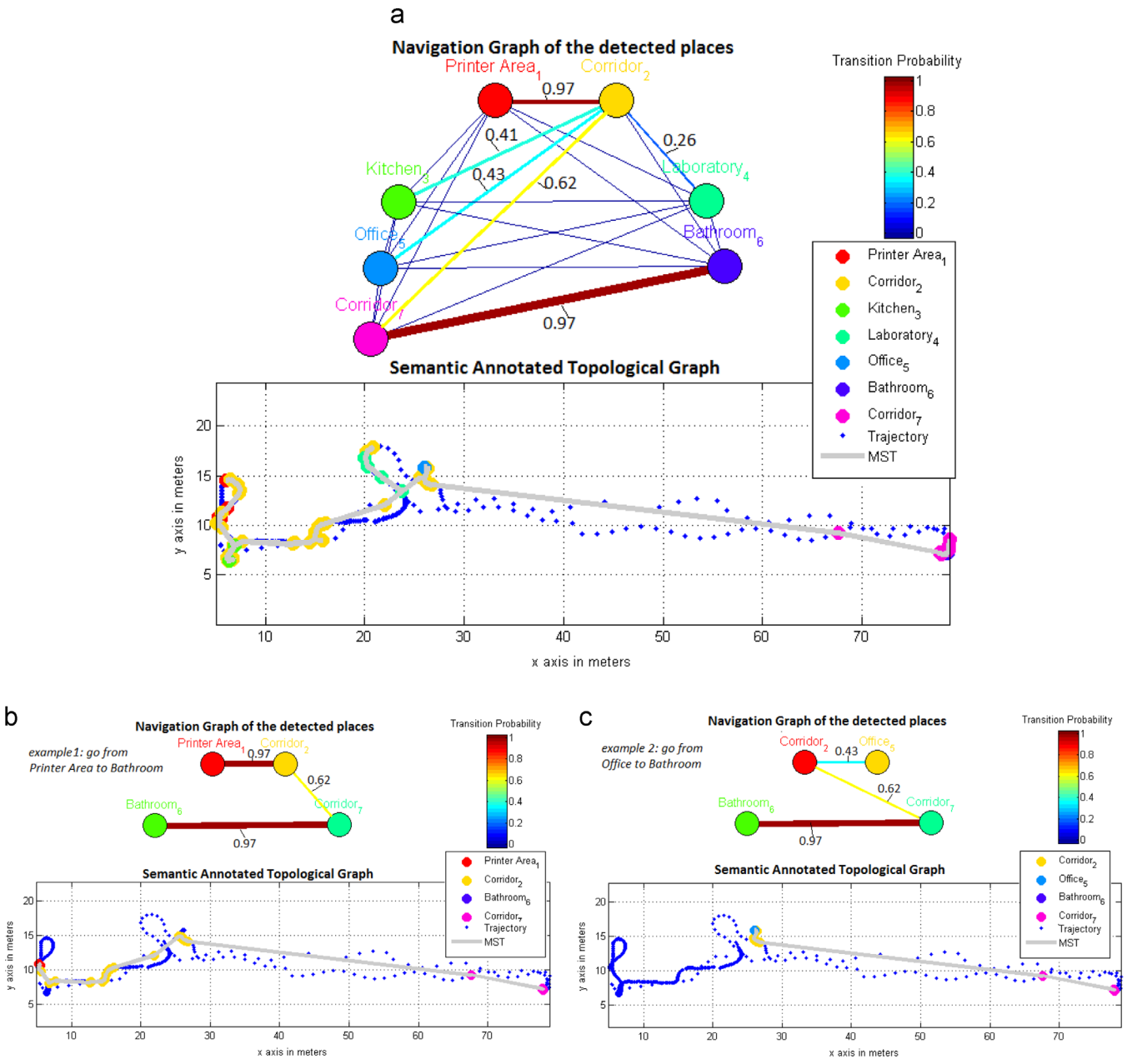

Fig. 10. (a) The ANG and the respective LSTM during the robots exploration in the large sequence $Sunny_1$ of the COLD dataset. The existing places are a "printer area", a "kitchen", a "laboratory", a "corridor", an "office" and a "bathroom". Examples of *go-to* actions (b) from the "*printer area to the* "*bathroom*" and (c) from the "*office to the* "*bathroom*".

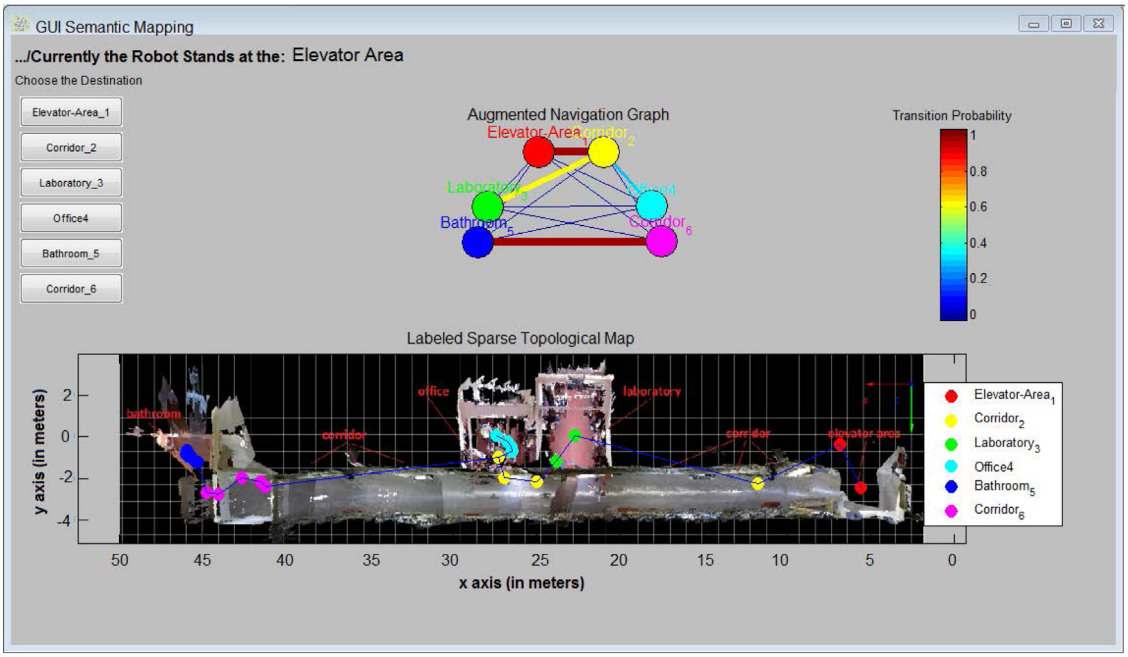

Fig. 11. The GUI which is available for the user after the robot has explored the entire environment. The ANG comprises all the detected places accompanied by the respective edges, which capture the transition probabilities among the detected places indicating their physical connection. The LSTM nodes are fully expanded and superimposed on the metric map of the explored environment. These nodes are connected with the MST revealing the minimum cost distance for the robot to transfer from the first to the last added node in the LSTM.

confusion matrices of which are shown in Fig. 9(a) and (b), respectively. It is exhibited that the appearance based histograms provide substantial generalization capabilities to our system and are capable to precisely deduce the type of a place, even in cases of no visual remembrances on specific scene.

Semantic mapping: According to the literature survey conducted in Section 2 the semantic mapping methods are typically complex and modular frameworks, thus a direct comparison between these methods is not easy mainly due to the fact that each method employs diverse modules to solve the same issue e.g. SLAM, or place recognition. Moreover, to the best of our knowledge there is not a common evaluation method for the semantic mapping algorithms. However, a direct qualitative comparison of our semantic mapping method with other contemporary approaches, reveals that the suggested one exhibits extra capabilities as summarized in Table 2. Specifically, in their majority semantic mapping methods retain metric mapping, place recognition and topological maps however, concepts such as space segmentation and temporal coherency are rarely met (Kostavelis and Gasteratos, 2015). The experimental evaluation of the suggested semantic mapping framework has been assessed on the Ljubljana subdataset as shown in Fig. 10(a) and comprises six place categories. The LSTM has been correctly formed by detecting precisely all the existing places observed during the robot's exploration. One attribute that is spotted in this figure is that the places labeled as "corridor" have been sufficiently memorized and the LSTM will not create additional nodes until the appearance based histograms differentiate significantly. Another attribute also highlighted here is that the "corridor" is split into two different places according to the MST partitioning of the LSTM. The spatial connectivity in all cases is correctly formed indicating the precise formation of the groups in LSTM, given their geometrical attributes. Regarding the respective ANGs of the places visited, these have also been correctly evolved during the robot's exploration. The edges with zero transition probability are omitted, indicating thus the absence of direct connections among non-adjacent places. As an example, traversing from the "printer area" to the "bathroom" is physically impossible and the robot has to pass through the "$corridor_2$" and "$corridor_5$", thus optimizing its route towards its target by including only the essential, known places, as depicted in Fig. 10 (b) and (c). This is an important attribute of the proposed method since the robot is able to use optimal shortcuts so as to reach its target location without roving about.

Robot navigation: After concluding the construction of the semantic map the robot navigation module has also been evaluated. The ability of the system to correctly build the ANG and the LSTM has been discussed in the previous section; here we evaluate the entire framework, i.e. the capacity of a robot to receive and execute *go-to* commands given the semantic map. Initially, the robot was manually guided through the environment under exploration and the metric and topological maps were evolved progressively. Simultaneously, the SVM models (trained on the Ljubljana data) were queried at each frame and produced semantic inferences for the visited places. During the robot's travel, the LSTM and the ANG are constructed covering a 70 m route with different places, namely "elevator area", "corridor", "laboratory", "two-persons office" and a "bathroom". The semantic map is constructed, as depicted in Fig. 11, where the GUI is illustrated containing information on places recognized in the environment, as well as the ANG and the LSTM superimposed over the metric map. By observing the ANG, it is clearly shown that the algorithm has detected all existing places, annotating them suitable with the respective label. Given the fact that the corridor is a repetitive pattern, the LSTM adds very sparse nodes indicating that the robot explores the same area along many frames. In some cases, where this pattern alters, additional nodes with the same label, e.g. "$corridor_2$" and "$corridor_6$, are created in the LSTM. The increased transition probability expressed as an edge in the ANG indicates a massive physical connection among "$corridor_2$" and "$corridor_6$". Moreover, the present location of the robot is also demonstrated on the GUI, which in the case of Fig. 11. By utilizing the buttons in the left hand side of the GUI, the user may order the robot to navigate towards any of the detected places.

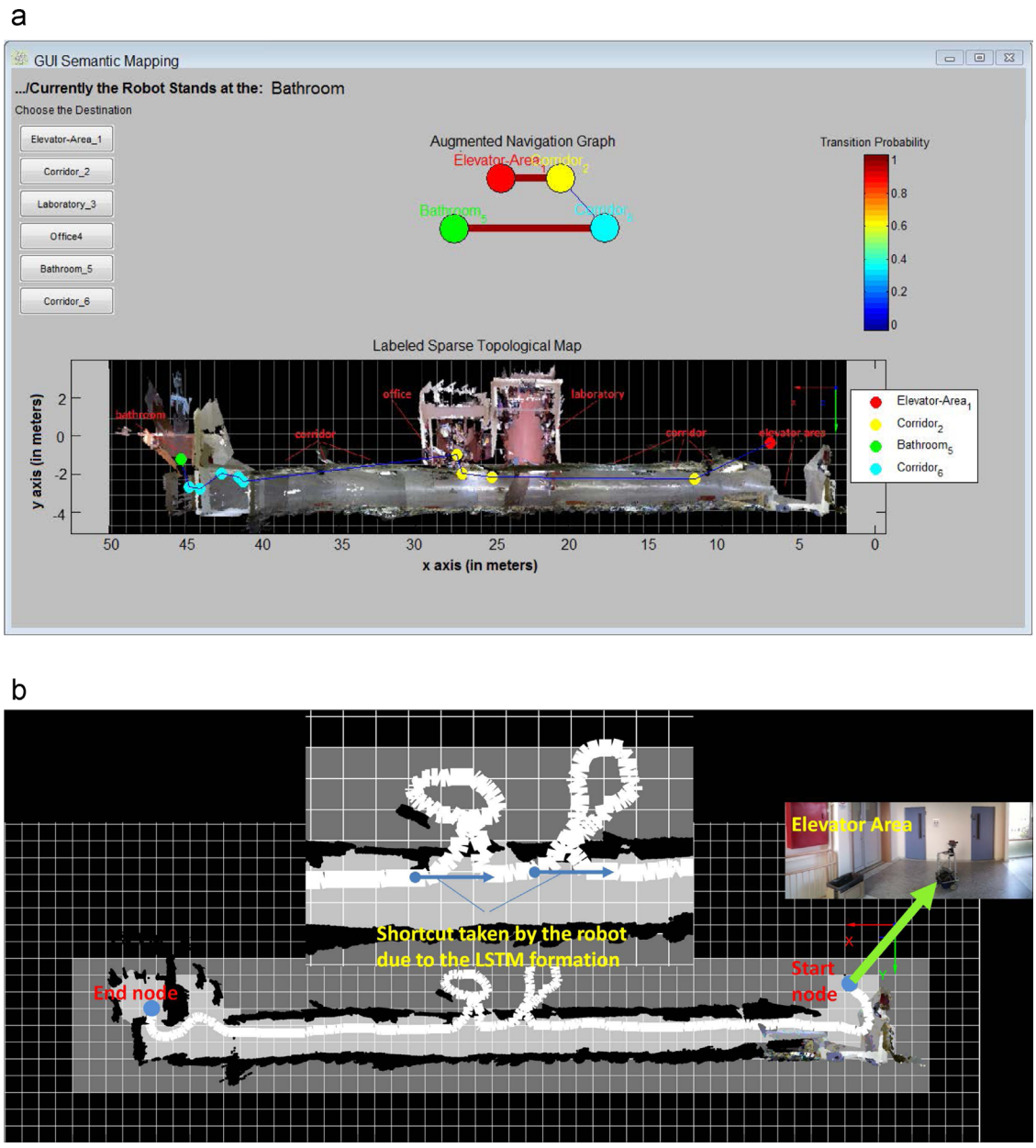

Fig. 12. (a) The GUI exhibiting the ANG as evolved following the user's request; it conceptually identifies the places the robot has to traverse in order to go from the elevator area to the bathroom. (b) The resulting occupancy grid with a fraction of 3D point clouds superimposed on it, along the robot's local navigation from one place to the other.

In order to prove the navigation capacities of the proposed method we have examined a representative navigation scenario. It comprises the navigation from the "elevator area$_1$" to the "bathroom$_5$". After the selection of the target location button in the GUI, the respective labeled nodes on the LSTM are activated and at the same time the robot progressively recalls the appropriate nodes on the topological map. Hence, in course of the robot's navigation its two nearest consecutive nodes of the topometric map are examined, by means of their associated point clouds. The target location is considered to be the first detected node of the group of LSTM vertices corresponding to the selected target place. The resulting forward path, as calculated from the Dijkstra algorithm is illustrated in Fig. 12(a). In this figure the triggered nodes on the LSTM are also highlighted and superimposed on the metric map. During the robot's navigation from one place to another, the nodes in the topological map that bear the respective semantic label are triggered, so that the temporary 2D occupancy grid to be formed. The result of this procedure is exhibited in Fig. 12(b) where the recalled metric map is sufficiently utilized by the robot to reach the target location. Note that in these scenarios the robot took the shortcuts thus avoiding the pass from unnecessary locations. Note that once the robot has reached its target location the ANG switches on the initial state.

6. Discussion

In this paper an integrated framework that seamlessly brings together a semantic mapping method with a robot navigation strategy has been presented. The ultimate aim of the proposed work is to bridge the gap between robot navigation and human cognizance through a progressive mapping methodology, which draws semantic inferences about the robot's surroundings based solely on geometrical representations and on a memorized visual vocabulary. To achieve this goal, we aimed for a stacked map hierarchy of four different kind of maps, namely a metric, a

topological, a labeled sparse topological and an augmented navigation one. The proposed method involves the construction of a dense 3D map, organized in a hierarchical manner in terms of a topological and a labeled sparse topological map. Consequently, during navigation the robot recalls in a recursive manner the memorized nodes, which are registered with point clouds of high density and performs the navigation task safely and with low computational burden. An additional advantage of our work is that it facilitates the unsupervised partitioning of the explored environment into labeled places.

One might wonder whether it is necessary to utilize so many maps for the implementation of the navigation task, yet indeed each of the distinctive representations fulfills a separate unique purpose. More precisely, the topometric (metric and topological) one preserves the geometrical information of the environment in terms of point clouds registered in a graph of nodes, facilitating thus the robot navigation in a global coordinate system. The LSTM establishes the spatiotemporal coherence by associating the respective nodes in the topometric maps via place labels and geometrical transformations, enabling bidirectional exchange of information among the conceptual and metric maps. Taking advantage of this spatiotemporal coherence the ANG encloses the semantic attributes of the detected places as well as their connectivity relationships, expressed in terms of their transition probability. Thus, the high level semantic information is amalgamated with the low level geometrical one along the robot's perambulation. Consequently, the proposed ANG reveals qualitative navigation characteristics similar to the ones apprehended by a user, whilst preserving also tokens of quantitative ones, in terms of traversability quantification among places. To this end, the ANG can be further exploited by the user to issue high level commands to the robot directly. Execution of such orders is performed hierarchically, firstly by planning an abstract route and secondly by triggering the appropriate sequence of nodes in the LSTM. Subsequently, the respective nodes in the topometric map are recalled by the robot to perform local navigation routines, so as to reach the target location.

At this point it should be mentioned that the proposed method assumes static environments during both the formation of the semantic map and robot navigation. However, the goal of this work is to establish a method for robot navigation within an explored environment in terms of human concepts. Finally, the proposed method has been thoroughly examined on long range indoor datasets involving multiple scenarios on semantic maps creation and robot navigation through them. The overall system, built upon our previous work for topometric mapping and place recognition, has been shown to yield remarkable performance in terms of semantic mapping and navigation within the explored environment. In our future work we intend to integrate more semantic information considering also object recognition and tracking capabilities. This will assist both the accuracy of the robot localization, while the constructed semantic map will be more explicit by adding in the ANG further details. Another functional component that can be built upon this work is the dynamic update of the semantic and metric map with the aim to take into consideration changes in the explored environment. Moreover, in order to enhance the domestic behavior of the robot, human tracking and action recognition capacities can be added to our system. Thus, exploiting the already existing temporal proximity mechanism, a social map can be constructed on top of the semantic one, bringing our system even closer to actual and operation human robot interaction.

References

Anand, A., Koppula, H.S., Joachims, T., Saxena, A., 2013. Contextually guided semantic labeling and search for three-dimensional point clouds. Int. J. Robot. Res. 32 (1), 19–34.

Aydemir, A., Göbelbecker, M., Pronobis, A., Sjöö, K., Jensfelt, P., 2011. Plan-based object search and exploration using semantic spatial knowledge in the real world, In: Proceedings of the European Conference on Mobile Robotics (ECMR11), Orebro, Sweden.

Bailey, T., Durrant-Whyte, H., 2006. Simultaneous localization and mapping (SLAM): part ii. IEEE Robot. Autom. Mag. 13 (3), 108–117.

Blodow, N., Goron, L.C., Marton, Z.-C., Pangercic, D., Ruhr, T., Tenorth, M., Beetz, M., 2011. Autonomous semantic mapping for robots performing everyday manipulation tasks in kitchen environments. In: Intelligent Robots and Systems International Conference. IEEE, pp. 4263–4270.

Cadena, C., Gálvez-López, D., Tardós, J.D., Neira, J., 2012. Robust place recognition with stereo sequences. IEEE Trans. Robot. 28 (4), 871–885.

Case, C., Suresh, B., Coates, A., Ng, A.Y., 2011. Autonomous sign reading for semantic mapping. In: IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 3297–3303.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2 27:1–27. Software available at ⟨http://www.csie.ntu.edu.tw/∼cjlin/libsvm⟩..

Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C., 2004. Visual categorization with bags of keypoints. In: Workshop on statistical learning in computer vision (ECCV) vol. 1, pp. 22.

Dayoub, F., Morris, T., Upcroft, B., Corke, P., 2013. Vision-only autonomous navigation using topometric maps. In: International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 1923–1929.

Durrant-Whyte, H., Bailey, T., 2006. Simultaneous localization and mapping: part i. IEEE Robot. Autom. Mag. 13 (2), 99–110.

Ekvall, S., Jensfelt, P., Kragic, D., 2006. Integrating active mobile robot object recognition and SLAM in natural environments. In: International Conference on Intelligent Robots and Systems. IEEE, pp. 5792–5797.

Fazl-Ersi, E., Tsotsos, J.K., 2012. Histogram of oriented uniform patterns for robust place recognition and categorization. Int. J. Robot. Res. 31 (4), 468–483.

Filliat, D., Meyer, J.-A., 2003. Map-based navigation in mobile robots: I. A review of localization strategies. Cogn. Syst. Res. 4 (4), 243–282.

Günther, M., Wiemann, T., Albrecht, S., Hertzberg, J., 2013. Building semantic object maps from sparse and noisy 3d data. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2228-2233, http://dx.doi.org/10.1109/IROS.2013.6696668.

Galindo, C., Fernández-Madrigal, J.-A., González, J., Saffiotti, A., 2008. Robot task planning using semantic maps. Robot. Auton. Syst. 56 (11), 955–966.

Girdhar, Y., Giguère, P., Dudek, G., 2013. Autonomous adaptive exploration using realtime online spatiotemporal topic modeling. Int. J. Robot. Res. 0278364913507325

Ko, Dong Wook, Chuho Yi, Il Hong Suh., 2013. Semantic mapping and navigation: A Bayesian approach. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2630-2636, http://dx.doi.org/10.1109/IROS.2013.6696727.

Kostavelis, I., Gasteratos, A., 2013. Learning spatially semantic representations for cognitive robot navigation. Robot. Auton. Syst. 61 (12), 1460–1475.

Kostavelis, I., Gasteratos, A., 2015. Semantic mapping for mobile robotics tasks: a survey. Robot. Auton. Syst. 66 (0), 86–103.

Kostavelis, I., Boukas, E., Nalpantidis, L., Gasteratos, A., 2013. Visual odometry for autonomous robot navigation through efficient outlier rejection. In: International Conference on Imaging Systems and Techniques. IEEE.

Krishnan, A.K., Krishna, K.M., 2010. A visual exploration algorithm using semantic cues that constructs image based hybrid maps. In: International Conference on Intelligent Robots and Systems. IEEE, pp. 1316–1321.

Li, Z., Shi, Z., Zhao, W., Li, Z., Tang, Z., 2013. Learning semantic concepts from image database with hybrid generative/discriminative approach. Eng. Appl. Artif. Intell. 26 (9), 2143–2152.

Linde, O., Lindeberg, T., 2004. Object recognition using composed receptive field histograms of higher dimensionality. In: International Conference on Pattern Recognition, vol. 2. IEEE, pp. 1–6.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60 (2), 91–110.

Maohai, L., Han, W., Lining, S., Zesu, C., 2013. Robust omnidirectional mobile robot topological navigation system using omnidirectional vision. Eng. Appl. Artif. Intell. 26 (8), 1942–1952.

Martinetz, T., Schulten, K., et al., 1991. A Neural-Gas Network Learns Topologies. University of Illinois at Urbana-Champaign.

Martinez Mozos, O., Triebel, R., Jensfelt, P., Rottmann, A., Burgard, W., 2007. Supervised semantic labeling of places using information extracted from sensor data. Robot. Auton. Syst. 55 (5), 391–402.

Meger, D., Forssén, P.-E., Lai, K., Helmer, S., McCann, S., Southey, T., Baumann, M., Little, J.J., Lowe, D.G., 2008. Curious george: an attentive semantic robot. Robot. Auton. Syst. 56 (6), 503–511.

Meyer, J.-A., Filliat, D., 2003. Map-based navigation in mobile robots: II. A review of map-learning and path-planning strategies. Cogn. Syst. Res. 4 (4), 283–317.

Milford, M., 2013. Vision-based place recognition: how low can you go?. Int. J. Robot. Res. 32 (7), 766–789.

Nüchter, A., Hertzberg, J., 2008. Towards semantic maps for mobile robots. Robot. Auton. Syst. 56 (11), 915–926.
Nüchter, A., Lingemann, K., Hertzberg, J., Surmann, H., 2007. 6D SLAM3D mapping outdoor environments. J. Field Robot. 24 (8–9), 699–722.
Nielsen, C.W., Ricks, B., Goodrich, M.A., Bruemmer, D., Few, D., Few, M., 2004. Snapshots for semantic maps. In: International Conference on Systems, Man and Cybernetics, vol. 3. IEEE, pp. 2853–2858.
Pronobis, A., Caputo, B., 2007. Confidence-based cue integration for visual place recognition. In: International Conference on Intelligent Robots and Systems. IEEE, pp. 2394–2401.
Pronobis, A., Jensfelt, P., 2011. Hierarchical multi-modal place categorization. In: Proceedings of the 5th European Conference on Mobile Robots.
Pronobis, A., Jensfelt, P., 2012. Large-scale semantic mapping and reasoning with heterogeneous modalities. In: International Conference on Robotics and Automation. IEEE, pp. 3515–3522.
Pronobis, A., Caputo, B., Jensfelt, P., Christensen, H.I., 2006. A discriminative approach to robust visual place recognition. In: RSJ International Conference on Intelligent Robots and Systems. IEEE. pp. 3829–3836.
Pronobis, A., Jie, L., Caputo, B., 2010a. The more you learn, the less you store: memory-controlled incremental svm for visual place recognition. Image Vis. Comput. 28 (7), 1080–1097.
Pronobis, A., Martínez Mozos, O., Caputo, B., Jensfelt, P., 2010b. Multi-modal semantic place classification. Int. J. Robot. Res. 29 (2–3), 298–320.
Ranganathan, A., Dellaert, F., 2007. Semantic modeling of places using objects. In: Proceedings of the 2007 Robotics: Science and Systems Conference, vol. 3, pp. 27–30.
Rottmann, A., Mozos, O.M., Stachniss, C., Burgard, W., 2005. Semantic place classification of indoor environments with mobile robots using boosting. In: AAAI, vol. 5, pp. 1306–1311.
Rusu, R.B., Marton, Z.C., Blodow, N., Dolha, M., Beetz, M., 2008. Towards 3D point cloud based object maps for household environments. Robot. Auton. Syst. 56 (11), 927–941.
Rusu, R.B., Marton, Z.C., Blodow, N., Holzbach, A., Beetz, M., 2009. Model-based and learned semantic object labeling in 3D point cloud maps of kitchen environments. In: International Conference on Intelligent Robots and Systems. IEEE, pp. 3601–3608.
Tamas, L., Goron, L.C., 2014. 3D semantic interpretation for robot perception inside office environments. Eng. Appl. Artif. Intell. 32, 76–87.
Theodoridis, S., Koutroumbas, K., 2008. Pattern Recognition, 4th ed. Academic Press.
Thrun, S., Burgard, W., Fox, D., et al., 2005. Probabilistic Robotics, vol. 1. MIT Press, Cambridge.
Trevor, A., Gedikli, S., Rusu, R., Christensen l.H., 2013. Efficient organized point cloud segmentation with connected components. Semantic Perception Mapping and Exploration (SPME) workshop in conjunction with IEEE International Conference on Robotics and Automation (available at http://www.spme.ws/2013).
Ullah, M., Pronobis, A., Caputo, B., Luo, J., Jensfelt, R., Christensen, H., 2008. Towards robust place recognition for robot localization. In: International Conference on Robotics and Automation. IEEE, pp. 530–537.
Viswanathan, P., Meger, D., Southey, T., Little, J.J., Mackworth, A.K., 2009. Automated spatial-semantic modeling with applications to place labeling and informed search. In: Canadian Conference on Computer and Robot Vision, pp. 284–291.
Zender, H., Martínez Mozos, O., Jensfelt, P., Kruijff, G.-J., Burgard, W., 2008. Conceptual spatial representations for indoor mobile robots. Robot. Auton. Syst. 56 (6), 493–502.