Semantic labeling for indoor topological mapping using a wearable catadioptric system

A. Rituerto, A. C. Murillo, J. J. Guerrero

Departamento de Informática e Ingeniería de Sistemas
Instituto de Investigación en Ingeniería de Aragón
Universidad de Zaragoza, Spain

Abstract

An important part of current research on appearance based mapping goes towards richer semantic representations of the environment, which may allow autonomous systems to perform higher level tasks and provide better human-robot interaction. This work presents a new omnidirectional vision based scene labeling approach for augmented indoor topological mapping. Omnidirectional vision systems are of particular interest because they allow us to have more compact and efficient representation of the environment. Our proposal includes novel ideas in order to augment the semantic information of a typical indoor topological map: we pay special attention to the semantic labels of the different types of transitions between places, and propose a simple way to include this semantic information to build a topological map, as part of the criteria to segment the environment. This work is built on efficient catadioptric image representation based on the Gist descriptor, which is used to classify the acquired views into types of indoor regions. The basic types of indoor regions considered are *Place* and *Transition*, farthest divided into more specific subclasses, e.g., *Transition* into door, stairs and elevator. Besides using the result of this labeling, the proposed mapping approach includes a probabilistic model to account for spatio-temporal consistency. All the proposed ideas have been evaluated in a new indoor dataset presented in this paper. This dataset has been acquired with our wearable catadioptric vision system[1], showing promising results in a realistic prototype.

Keywords:
Semantic maps, Indoor place classification, Wearable sensors, Omnidirectional cameras, Global image descriptor

1. Introduction

For most autonomous tasks, one of the initial steps consists of obtaining a suitable representation of the environment. In order to obtain it, the system interprets the data acquired with different sensors on-line or in exploration phases to build different types of models depending on the tasks to be performed. Focusing on vision sensors, this modeling consists of arranging the acquired images into a visual memory or reference map. Data should be organized efficiently but more importantly, in a way as useful as possible to be used later. In many cases, big and accurate metric maps are not necessary or not informative enough, therefore higher abstraction level maps can be built, such as topological or object-based maps, such as [1, 2, 3, 4], detailed later in Section 2.

The general goal of this work is to achieve a useful semantic-topological map for indoor environments using a wearable catadioptric vision system for personal assistance. We propose how to include interesting semantic information on indoor topological models. In particular we design a simple approach to segment the environment into semantically meaningful clusters using omnidirectional images acquired with our wearable system. We represent the catadioptric images following the approach described in [5], which is based on an adaptation of the global Gist descriptor [6] to omnidirectional images. Our long term goal is

Email addresses: arituerto@unizar.es (A. Rituerto), acm@unizar.es (A. C. Murillo), jguerrer@unizar.es (J. J. Guerrero)

[1] http://robots.unizar.es/omnicam/

to merge the presented semantic model with a set of small metric maps of each topological region, which could be obtained with standard visual odometry or slam algorithms [7]. In this paper we focus on a new scene classification method for topological mapping. Our proposal includes the following two novel ideas with regard to other related works.

First, an improved criteria to define environment clusters. On-line topological mapping approaches usually segment the environment regions by evaluating the similarity within consecutive images and establishing different criteria to decide when and where to segment the trajectory. We define how to easily include the labeling from our semantic indoor region classifier as part of the criteria to organize the topology of the environment. This classifier is based on a model previously built from a few given examples of the different classes to be recognized. We find that most of the approaches for semantic indoor scene labeling try to label types of *Places* [8, 9, 10, 11].

Second, additionally to the types of *Places*, we perform a detailed analysis of the semantic information included in the types of *Transitions* (such as door, elevator or stairs) between these *Places*. This information can be of great interest for autonomous systems working indoors, since depending on the transitions we may be or not be able to traverse from one *Place* to another. Knowing the type of transition may allow us to choose a suitable robot team member to go to a particular destination, or give appropiate instructions in case of human assistance systems.

Two other interesting properties of our method, that are not novel themselves but their integration is essential in our proposal, are the following: first, the fact of using only global descriptors, with the corresponding improvement in efficiency with regard to the use of local features; second, the inclusion of a probabilistic model to keep the spatio-temporal consistency of the labeling along the trajectory. In spite of the simplicity of the image representation, the proposal gets to partition the environment into semantic meaningful areas for humans, as it can be seen later in the experimental validation, using the presented catadioptric dataset.

Additionally to our approach, this paper presents a new indoor dataset, used to evaluate our proposals. This dataset has been acquired indoors with a wearable system composed by an omnidirectional camera, an IMU and a GPS. The use of wearable systems is mostly oriented to create human assistance applications, and adds new difficulties to the work.

The rest of the paper is organized as follows. First we analyze related works on semantic and topological maps in Section 2, and Section 3 details the image representation we use. Section 4 provides the description of the *Places* and *Transitions* classifier and the sequence segmentation approach developed in this work. The experimental validation of the proposed ideas is summarized in Section 5, where we describe the new catadioptric image dataset acquired with a wearable system. Finally we conclude and discuss the future work in Section 6.

2. Related work

Topological modeling of the environment is a subject already studied for long [12, 13]. Initially, these models presented huge possibilities due to their lower computational requirements with regard to accurate metric maps. More recently, these models have gained interest due to the possibilities of augmenting them with semantic concepts [14], such as information about places [1] and/or objects [2]. Topological maps are many times built on top of a hierarchy of different map levels [15], e.g., a global topological map that connects smaller local metric maps [16]. An extensively used solution to achieve different efficiency and accuracy results at the different levels is to use global and local image features through the different steps of the hierarchy [17].

Augmenting topological maps with semantic information makes them more suitable for human-robot interaction [3, 4] and allows us to achieve more complicated goals [18]. Semantic mapping provides new opportunities to increase the autonomy and reasoning skills of intelligent systems, both for outdoor and indoor applications.

In outdoor settings, many of the recent and impressive approaches are achieved by combining multi-sensor information, typically vision and laser sensors [19, 20], to build topological models that include place or object recognition information. In the work [19], which deals with place recognition, the authors present an approach for appearance based mapping using extremely large datasets (1000 km) that efficiently recognizes previously visited places. The work in [20] is focused on objects rather than places, it

recognizes and labels objects in large urban environments proposing a Conditional Random Field based framework.

Focusing on the framework of this paper, indoor environments, we also find proposals using different types of sensors to interpret semantic information that will be included in a topological map. Initial proposals were typically achieved using range data, to learn a room-doorway-hallway structure indoors in [21] or [22]. We also find proposals using a combination of range and vision cues, for example in [3] they combine place and object recognition in exploration and semantic mapping approach. The work in [10] suggests a Support Vector Machine (SVM) scheme that learns how to optimally combine and weight each cue. In [23] boosting is used to learn a classifier with different place labels, using vision and range sensors.

Proposals only based on vision sensors are closer to our approach. Although they usually provide more detailed labels than only range data approaches, most of these approaches still include semantic labels only regarding Places (e.g., office, corridor, kitchen...) [24], considering all transitions as just connections between places. We find different types of approaches that try to classify types of places, with multiple proposals of how to learn the labels to be recognized and how to represent the images.

Regarding how to learn the environment model, some proposals are constantly trained and, sometimes, simultaneously run with human supervision to achieve a representation closer to human concepts [4, 27]. Others use weaker human supervision to learn a model from a few initial labeled samples, such as the work in [9]. This approach learns the representation of problematic locations (e.g. images showing only zoomed wall areas, without any information about the actual indoor region) from a few given examples. This helps to detect when those problematic cases occur and to avoid giving incorrect or noisy labeling.

Our augmented topological mapping approach makes use of human supervision, but only in the initial training phase, to provide sample labeled images of the types of indoor scenes of interest. Besides labels for places, our approach includes semantic information about the types of transitions. This is of particular interest for multiple-floor buildings, where depending on who is using the map a transition (elevator, stairs, closed door,...) may be feasible to traverse or not. We find other recent works also paying attention to transitions [28, 11]. The first work only detects doors, proposing to dynamically model the environment to react if a transition is suddenly closed. The second work applies the generic label *transition* to all areas that are not detected as being of a known type of place, so no knowledge of the kind of transition is included.

Besides the described augmented mapping approaches, we find additional closely related works regarding the more general problem of place or scene recognition indoors [29]. This work also points the idea of plenty of indoor areas usually not considered in classification approaches. Among the big set of types they study we can find elevators and stairs. In this work we include all type of indoor regions that can actually be considered as a transition between places (elevators, stairs and areas under doors or under jambs). Another common point with that work is the use of the Gist descriptor [6].

Attending to the image representation, some works propose to work with local features, such as the robust vision-based robot localization using combinations of local features from [25], or the work in [26] that presents an integration of object detection, using local features, and global image properties for place classification. In this work we use global features. Our global image descriptor is based only on the global Gist descriptor, following the ideas initially presented in [31]. The Gist descriptor was initially presented for classifying outdoor scenes [30] and used in more recent work together with additional cues for indoor scene recognition [29]. Global descriptors are known to be more efficient and compact, but usually less robust and discriminative, than local features. However, in the current work promising results pointed that this weakness can be compensated to a certain extent by the powerful scene representation contained in omnidirectional images.

The use of omnidirectional images is another key characteristic of our proposal. Some proposals take advantage of wide field of view cameras to acquire more compact visual models, e.g., in [32, 33] panoramic cameras are used for indoor topological map building and [34] presents an approach for topological mapping and navigation using a catadioptric vision system. We use this second type of images, acquired with a catadioptric vision system, usually smaller and with lower cost that the panoramic cameras. However, these cameras present additional issues to deal with, such as big image distortion, noise and parts of the vision system self-reflected in the views. These issues together with the fact that we are using

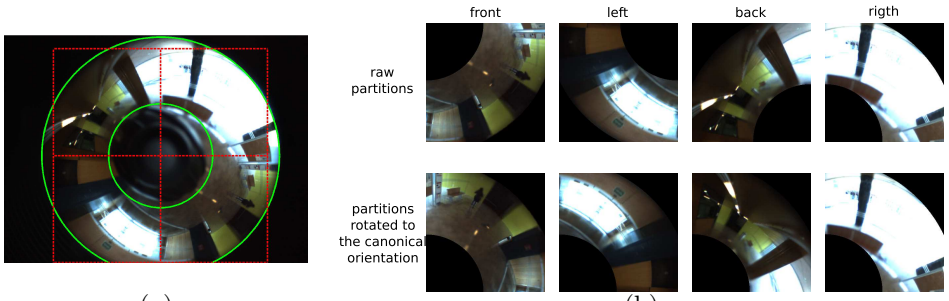


Figure 1: (a) Raw image acquired with our catadioptric vision system. The green line shows the mask limits used to avoid the artifacts of the image. The dashed red line defines the limits of the four image parts. (b) Top row shows these four raw parts, while bottom row shows the four parts rotated to the canonical orientation.

a wearable system, requires a carefully designed image representation detailed in the following section.

3. Image representation and similarity

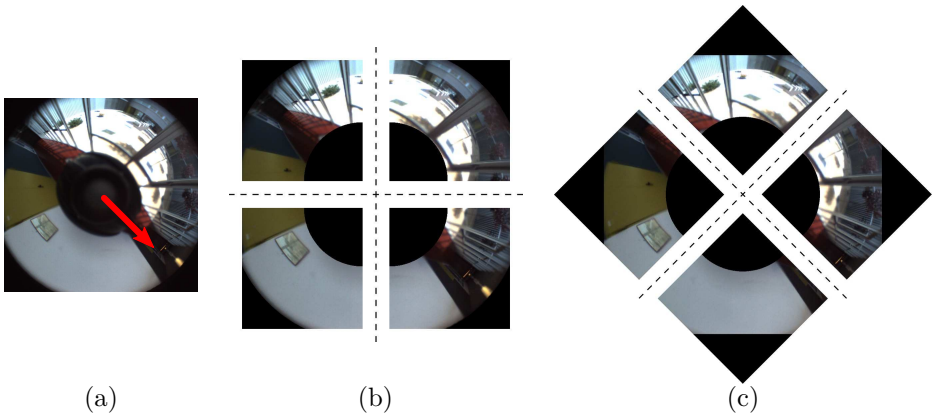
Visual descriptors that capture image information as a whole are known as global descriptors, while those that capture a specific interest region are called local descriptors. It has been typically shown that local descriptors are more accurate for visual localization than global descriptors, but also have much larger memory and processing requirements [35]. Therefore, to deal with large quantities of images for tasks where efficiency is an issue and it is not required a detailed analysis of image content, a global representation is preferred.

In this work we use the Gist descriptor [30], a holistic image representation or global image feature. In particular, it is a low dimensional representation of the scene captured in an image which corresponds to the mean response to steerable filters at different scales and orientations computed over 4x4 sub-windows. The descriptor consists of a vector of 320 components for each color band used, so in a RGB image it has 960 components. This global feature was presented and applied as a successful tool for scene recognition, with the big computational saving of bypassing the segmentation and the processing of individual objects or regions. Approaches using this descriptor typically work with squared conventional images, most of the time assuming frontal scene views acquired with the camera focal axis parallel to the ground plane, since the descriptor is not rotation invariant.

In the case of omnidirectional cameras the image contains 360^o degrees field of view around the camera. This presents a problem when facing the same scene with different direction of travel, i.e., same location but camera rotated around the vertical axis. This situation can generate apparently different scene view, although it is just a matter of re-organization (shift), of the scene parts. To handle this problem and try to make our image representation invariant to the camera vertical rotation we split the omnidirectional images in four parts, similar to the method presented in [5]. Each image is split in four parts, each part is rotated to a canonical orientation (Fig. 1 shows how this rotation is done), and he Gist descriptor is computed for each part. We need to mask out parts of the image where artifacts appear, mostly produced by the reflection of catadioptric system elements in its own mirror. With this representation, the omnidirectional image Gist $\mathbf{g}$ is composed by four conventional Gist descriptors, one computed for each image part (front, left, back and right): $\mathbf{g} = [g_f, g_l, g_b, g_r]$.

We have analyzed how to split the omnidirectional images. Fig. 2 shows two partition possibilities: Direct partition (Fig. 2(a)), that just splits the image into four equal squares, and Rotated partition (Fig. 2(b)), where the parts are extracted from the 45^o rotated image. Due to the camera orientation, the parts extracted in the Direct partition correspond to the main directions of the scene according to the Manhattan World Assumption. The use of these two partitions is analyzed in detail in subsection 3.1.

The similarity between two images using this representation is obtained based on the Euclidean distance between the descriptors. We compute the minimum distance that can be obtained between one image and the four possible permutations of the four sections of the second image. These permutations correspond with the four possible alignments of the sectors of the image and will hopefully provide us



(a) (b) (c)

Figure 2: (a) Omnidirectional image acquired with the Wearable OmniCam system partitioned using the two methods analyzed: (b) Direct and (c) Rotated. The red arrow in the first image shows the front direction of the helmet.

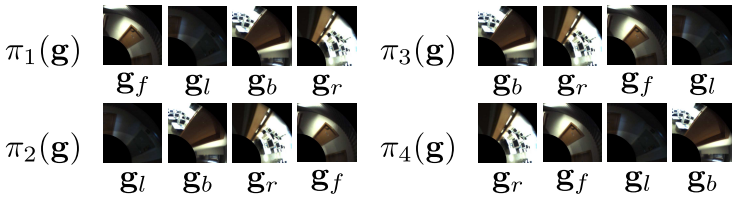


Figure 3: Graphic representation of the circular permutations of the Gist descriptor.

with the best alignment of the two evaluated images. Being $\mathbf{g}$ and $\mathbf{g}'$ the descriptors of two images, the distance between them is:

$$dist(\mathbf{g}, \mathbf{g}') = \min_m (d_e(\mathbf{g}, \pi_m(\mathbf{g}'_{flbr})), \tag{1}$$

where $\pi_m(\mathbf{g}'_{flbr})$ is the m^{th} circular permutation of the descriptor $\mathbf{g}'$ component vectors ($m = 1, 2, 3, 4$) and d_e the Euclidean distance between the Gist descriptors of two omnidirectional images. Fig. 3 shows the circular permutations in a graphic way.

3.1. Rotation invariance analysis

To analyze in detail the rotation invariance issues described above we have performed two experiments.

With the first experiment we want to prove the rotation invariance achieved with this image representation. We get 36 images equally distributed along a 360^o camera rotation movement without translation, around the vertical camera axis. Using the Direct partition we achieve invariance to vertical rotation at angles multiple of 90^o and using both the Direct and the Rotated partition together we get invariance to rotation at angles multiple of 45^o. This rotation invariance is not robust to all kind of movements, but the Manhattan assumption seems a reasonable one to work with man-made environments, where the possible directions of travel on a particular location usually fit these restrictions. Each image from this test set corresponds to a rotation of 10^o with regard to the previous image. We extract the Gist descriptor of all images with the two partitioning methods, so for each image we have two descriptors ($\mathbf{g}_{Direct}$ and $\mathbf{g}_{Rotated}$). We compare the Gist of the Direct partition ($\mathbf{g}_{Direct}$) of the reference 0^o image with both the Direct ($\mathbf{g}_{Direct}$) and the Rotated ($\mathbf{g}_{Rotated}$) Gist descriptors of all the following images. Using a perfect rotation invariant representation all images would get exactly the same descriptor and the distance (1) would be 0. Figure 4(a) shows the results of this test. The red line represents the distance between $\mathbf{g}_{Direct}$ in the test images and the initial reference image. It shows the higher distance values (less similar images according to our representation) at rotations of 45^o, 135^o and 225^o; while, as expected, the minimum distances appear at rotations of 0^o, 90^o, 180^o, 270^o and 360^o. The blue line represents the distance between $\mathbf{g}_{Direct}$ in the test images and $\mathbf{g}_{Rotated}$ in the reference image. The black

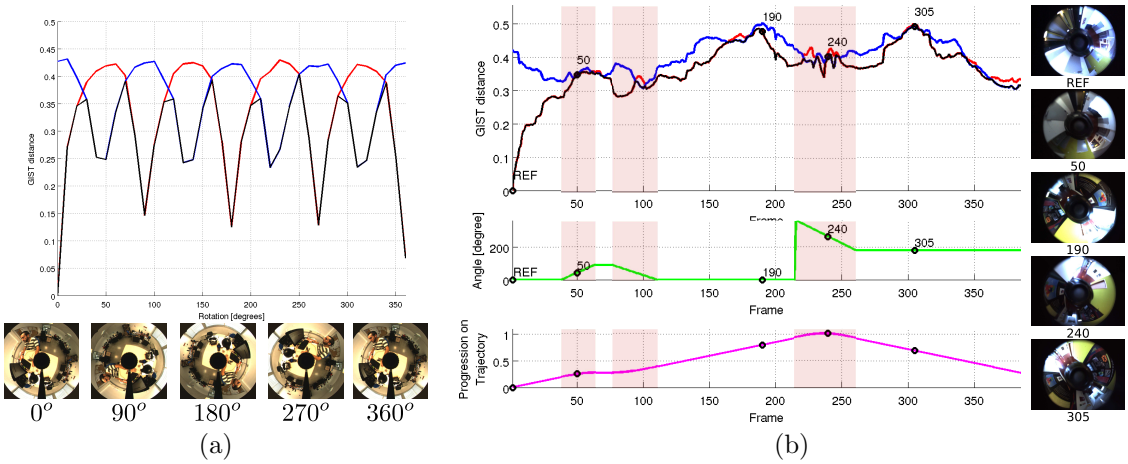


Figure 4: Rotation invariance of the image representation proposed (a), and influence of translation vs. rotation (b). The red line shows the distance $\mathbf{g}_{Direct} - \mathbf{g}_{Direct}$ and the blue one the distance $\mathbf{g}_{Direct} - \mathbf{g}_{Rotated}$. The black line shows the minimum value of both distances. The pink areas in (b) mark the parts of the trajectory where the camera is rotating. The omnidirectional images on the right of the figure correspond to the frames marked in the plots. The green line shows the rotation respect to the reference frame. The pink line shows the translation of the camera in the trajectory, the distance from the reference frame.

line represents the minimum value of red and blue line. This merged result confirms that using both partition methods we achieve better invariance to rotation, in particular to rotation at angles multiple of 45^o.

Second experiment is designed to show the influence of using one or two of the described partition methods while moving indoors. We have chosen a subsequence of the dataset where the camera moves along a corridor and returns the same way but from opposite direction (180^o rotation). The test consists of comparing the $\mathbf{g}_{Direct}$ of all images against the $\mathbf{g}_{Direct}$ and $\mathbf{g}_{Rotated}$ of the reference image. The image used as reference is the first image of the sequence. Ideally, we would like to observe how the distance between images increases as we get the test image farthest from the initial image. Figure 4(b) shows the results of this second experiment. The frames 50 and 240 correspond to frames where the camera is rotating. Points 190 and 305 correspond to the highest Gist distance, points of the trajectory that are almost at the farthest location from the reference. This suggest that the Gist distance variations are bigger due to the translation along the corridor than due to the rotation. The distance $\mathbf{g}_{Direct} - \mathbf{g}_{Rotated}$, blue line, is usually higher than the $\mathbf{g}_{Direct} - \mathbf{g}_{Direct}$, red line. Points where Direct versus Rotated distance is smaller than Direct versus Direct distance correspond to parts of the trajectory where the camera is rotating.

Therefore, as already mentioned, we can see small improvements using the duplicate Gist partition while navigating indoors. Even if it is more robust to rotations (to all angles multiple of 45^o instead of only multiples of 90^o), the increase in the Gist distance due to the camera rotation issues is small compared to those that appear due to translation. Therefore, the experiments in the rest of this work were performed using only the Direct partition method.

4. Augmented topological map with semantic labels of indoor scenes

This section describes all the steps of our augmented topological mapping approach. First, we propose a simple classification to identify basic indoor scene classes of interest (*Places* and *Transitions*) to discover the topology of the environment. Then, we evaluate the classification into more detailed types of scenes and finally integrate it in a proposal for augmented topological map building.

4.1. Labeling of Places and Transitions

Classification of indoor areas into *Places* and *Transitions* is natural and easy for humans when navigating through a building, and they represent the basis to build a topological map of the environment.

Table 1: Classes and subclasses considered in this work.

Classes	Subclasses
Places (P)	Corridor ($P1$) Big Room ($P2$) Medium Room ($P3$) Small Room ($P4$)
Transitions (T)	Door ($T1$) Jamb ($T2$) Stairs ($T3$) Elevator ($T4$)

Places are the nodes of the model and *Transitions* correspond to the edges between nodes. The main objective in this part of the work is to develop a method to automatically classify the images of a sequence into *Places* and *Transitions* to build an initial map with this information. Additionally we evaluate how to label the scene captured in each omnidirectional image into subclasses: the images classified as *Places* are farthest labeled into corridors and rooms of different sizes (big, medium and small rooms) and the images classified as *Transitions* are labeled as doors, jambs, stairs and elevators (Table 1).

Both subclasses provide the model with augmented semantic information, but of particular interest, and different from other approaches, is the fact of analyzing in detail the types of transitions. Indeed, the actions and movements required to traverse each of them are significantly different both for a human or robot navigating the map. For example, climbing stairs is not the same as traversing a door, or the type of movement to be generated may be different in a corridor and in a big room.

We describe next how to perform this classification based on the image representation and similarity evaluation described in previous section.

4.1.1. The Environment Model

A basic step in our method is obtaining the environment model, to use it later as reference to classify new occurrences. This environment model is created using a set of training reference images. It is composed of representative descriptors of each class and subclass given with this training data. In the dataset used in this work, detailed in next section, all the images have been manually classified and grouped in clusters of consecutive images belonging to the same semantic class/subclass. To build the model in a systematic way we consider as training data the first n_i clusters of each class. The value n_i for each class i is computed as $\frac{C_i}{4}$, being C_i the number of clusters of class i in the environment. To obtain a more homogeneous sampling, the value n_i is quantized into the following set of values: $n_i \in [1, 2, 5, 10]$. The model of each class, M_i, initially consists of all $\mathbf{g}_{Direct}$ descriptors of the training images. Note that typically *Place* clusters would include more images than *Transition* clusters, since the time spent traversing a corridor is longer than the time spent crossing a door, so more images of that type are acquired. To avoid that this fact leads to unbalanced models towards *Place*, we use a standard k-means method to find the k Gist descriptors that better represent each class. Then, all classes have the same amount of reference data in the model, the k Gist descriptors that correspond to the centroids of the obtained clusters. More formally, the environment model is:

$$M_i = \{M_{i,j} | j = 1..m\} \text{ with } M_{i,j} = \{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_{k_{Model}}\} \tag{2}$$

where M_i is the model of class i ($i \in [P, T]$), m is the number of subclasses of class i, and $M_{i,j}$ is the model of subclass j from class i that is composed by its k representative descriptors.

4.1.2. Label new occurrences according to the environment model

To label new images, we use a simple nearest neighbor based classifier. To measure the likelihood of a new image being of a particular class, we compute the following likelihood function (3) based on the Gist descriptor distance, and assign the maximum likelihood solution as label for the new image.

$$p(I_t | S_t = i) = \frac{K\, e^{\frac{-d_i}{\sigma^2}}}{\sum_{j=P,T} K\, e^{\frac{-d_j}{\sigma^2}}} \tag{3}$$

$S_t = i$ is the event of being in an area of class i at time t, when the image I_t is acquired. Then $p(I_t|S_t = i)$ is the likelihood of acquiring image I_t at time t being in an area of class i. Parameters K and σ^2 are user defined gain and variance respectively, in this work we use $K = 1$ and $\sigma^2 = 0.2$, adjusted after some initial experiments. d_i and d_j are computed comparing the current image Gist with the environment model:

$$d_c = \min_{\mathbf{g}' \in M_c} (dist(\mathbf{g}_t, \mathbf{g}')) \tag{4}$$

d_c is the similarity between image I_t and the class c. The Gist descriptor $\mathbf{g}_t$ is the descriptor of the image I_t and $\mathbf{g}'$ is one of the representative descriptors of class c included in the model (M_c). The distance between these descriptors is computed following equation (1).

4.2. Semantic sequence segmentation with temporal consistency

This section describes our complete approach for labeling sequential information. It is based on the image classifier described in previous subsection. The idea is that this semantic labeling of a sequence can be of great help to obtain more meaningful topological representations of the environment captured in that sequence.

Generally, with catadioptric images, if we only pay attention to consecutive image descriptors distance to decide where to split the topological regions, what we obtain is far from a semantic segmentation that a human would do, containing for example lots of small clusters. This is because even consecutive catadioptric images may present big visual differences, due to big image distortions and image changes. This effect is specially pronounced when objects and scene elements are close to the camera (as it usually happens indoors). Our proposal uses semantic labels as basic criteria to obtain semantic meaningful clusters in the topological model as detailed next.

First, we enclose the classifier described before in a framework that allows us to include spatio-temporal coherence in the model. We expect this coherence to improve the classification on sequential data: if the current image is very likely to belong to a transition area, next image is also likely to be part of it. We model these ideas using a Hidden Markov Model (HMM) following the approach presented in [36]. A HMM is a dynamic Bayesian network that represents a sequence of variables. At each instance of time the state is a random variable which can take one of the just two values: P (*Place*) or T (*Transition*). Let S_t be the random variable that represents the event of being in *Place* or *Transition* area at time t and I_t the image at this time. Then, the problem of detecting the kind of area j being crossed can be formulated as the search of j that satisfies:

$$j = \arg\max_{i \in \{T,P\}} \; p(S_t = i|I_t). \tag{5}$$

The posterior probability $p(S_t = i|I_t)$ is the probability of the event $S_t = i$ given the image I_t, which can be decomposed using the Bayes rule and the Markov property:

$$\begin{aligned} p(S_t = i|I_t) =& \alpha \;\; p(I_t|S_t = i)p(S_t = i|I_{t-1}) = \\ =& \alpha \;\; p(I_t|S_t = i) \sum_{j=T,P} p(S_t = i|S_{t-1} = j)p(S_{t-1} = j|I_{t-1}), \end{aligned} \tag{6}$$

where α is a normalization term, and the conditional probability $p(I_t|S_t = i)$ is the likelihood function (eq. (3)) modeling the likelihood of the current image I_t being of type i. The term $p_(S_t = i|S_{t-1} = j)$ is the state transition probability for observing the event $S_t = i$ given $S_{t-1} = j$, i.e., having an image of type i when previous image was of type j. This term models the probability of all possible changes in the state from time $t-1$ to t. We need to model four possible state transitions: $p(S_t = i|S_{t-1} = j)$, with $i, j \in T, P$. In practice, we set empirically the value of the probabilities of repeating the same event occurred at time $t-1$ in time t, $p(S_t = i|S_{t-1} = i)$, so the rest can be computed as $p(S_t = j|S_{t-1} = i) = 1 - p(S_t = i|S_{t-1} = i)$, with $j \neq i$.

Algorithm 1 details the proposed semantic trajectory segmentation method. For each new image the probability of being *Transition* or *Place* is estimated using the HMM. Consecutive images of the same class are grouped into the same cluster until the criteria to start a new cluster is fired. This criteria is based on the likelihoods estimated from the described HMM, but to prevent the appearance of too small clusters a criteria based on the similarity with the first image of the current cluster ($minSize$ filter) is

Algorithm 1 Semantic sequence segmentation method.

Input: Omnidirectional image sequence and environment model
Output: Semantic sequence segmentation

```
n = Number of the current cluster
th = Similarity threshold of the minSize filter
M_i = Model of class i, with i ∈ [P, T]
g_n = Gist of the first image of the current cluster n
g_t = Gist of the new image I_t
P_{t-1} = Probabilities at previous step
while not end of sequence do
  // New image I_t
  g_t = OmnidirectionalGist(I_t)
  // Compute similarity with the current cluster
  d = dist(g_t, g_n);
  // Compute probability of being transition or place
  [p_P, p_T] = HMMEnvironmentModel(g_t, M_P, M_T, P_{t-1})
  if p_P > p_T then
    state = P
  else
    state = T
  end if
  if d > th & state ≠ state_{ncluster} then
    CreateNewCluster(I_t,n + 1,state)
    n = n + 1
  else
    IncludeImageInCluster(I_t,n)
  end if
  P_{t-1} = [p_P, p_T]_{t-1}
end while
```

included. If this distance is below the similarity threshold (th) established, the new image is included in the current cluster, even if classification results according to HMM likelihood would label it as a different class than current images in the cluster. We will see the differences of using one or both of this criterion to build the topological map in next section.

This first step just segments the input sequence into clusters of images labeled as *Places* or *Transitions*, the classification into subclasses is performed next. We try to take advantage of this first level classification as a prior for the more detailed classification into subclasses. Once an image is labeled as *Transition* or *Place*, we look for the subclass with the reference descriptor most similar to the current image. We already know the class assigned to the image, so we only evaluate the subclasses of that class.

Note that at the end of the process, we want to assign a unique class and subclass to all members of each cluster. However, during the process images labeled as different subclasses may had ended up together. We consider this is noise due to the fact that descriptors of some subclasses are pretty close to each other and difficult to separate sometimes (doors and jambs for instance, are hard to distinguish for a human observer as well). Then, to assign the most likely subclass label to the whole cluster, we compute the mode of the subclass label assigned to each image in the cluster.

5. Experiments

In this section we present a new dataset acquired as part of this work and the results of the experimental validation of our proposed method performed with it.

5.1. The Wearable OmniCam Dataset

The catadioptric image dataset presented in this work has been acquired with our Wearable OmniCam acquisition system. This system includes a small hyper-catadioptric camera mounted on the top of a helmet (Fig. 5(a)), a 3-axis IMU (compass, gyroscope and accelerometer) and a GPS device. The three sensors are synchronized and the camera has been calibrated using the approach described in [37].

Table 2: Number of clusters of each class in the dataset. Values between parentheses are the number of images of that class/subclass.

Places	TOTAL	Corridor	Big Room	Medium Room	Small Room
	56 (16522)	38 (12577)	7 (1559)	3 (1021)	8 (1365)
Transitions	TOTAL	Door	Jamb	Stairs	Elevator
	55 (4382)	40 (1268)	9 (514)	4 (1933)	2 (667)

However, the presented data has been acquired indoors, so GPS is deactivated. IMU data is also not used in this work, that is a purely vision based approach, but could be used for future works and included in the published data.

The use of wearable sensing is mainly intended for applications of human assistance. There are a lot of sensors used in different ways to help persons: GPS for localization and guidance, IMU for movement supervision, cameras for object or place recognition, range sensors for obstacle avoidance. In this work we have focused in omnidirectional vision as a wearable system, so we find some problems that do not exist when using omnidirectional cameras for robots. In our case, placing the camera in a helmet, make us face the head movements of a person walking. Up to our our knowledge this is the first dataset published form indoor environments with a wearable omnidirectional camera.

The dataset acquisition has been performed inside one building at our Campus at the University of Zaragoza, Spain. The building has three floors and includes areas of different types: corridors, research laboratories, offices, classes, etc. The acquisition has been performed by a person wearing the helmet, so the dataset suffers the typical motion of a person walking. A long trajectory covering as much areas as possible was performed (many areas are locked or with restricted access so it was not possible to cover all regions in the building). Figure 5(b) shows the map of the three floors of the building highlighted with different colors, depending on the type of area traversed during the acquisition. The gray areas are parts not included in the dataset.

The visual part of the dataset consists of 20905 omnidirectional images at 1024x768 pixels resolution acquired at a frame rate of 10 FPS. The ground truth labeling of the building areas has been made according to our objective of separating *Places* and *Transitions*. We consider the main spaces of a building, like corridors or rooms, as *Places*. *Transitions* label comprises all the areas joining different *Places*: doors, stairs, elevators, etc. The more detailed classification in type of *Places* or type of *Transitions* has been chosen to adequately describe the environment of acquisition. *Places* are classified as Big, Medium and Small Rooms and Corridors. Typically small rooms correspond to offices, medium to classes and big to halls or laboratories, for simplicity we classify them according to their size despite their different uses. *Transitions* are classified as Doors, Jambs, Stairs and Elevators. The areas labeled as *Transitions* starts about 0.5 meters before and ends about 0.5 meters after the *Transition* has been crossed.

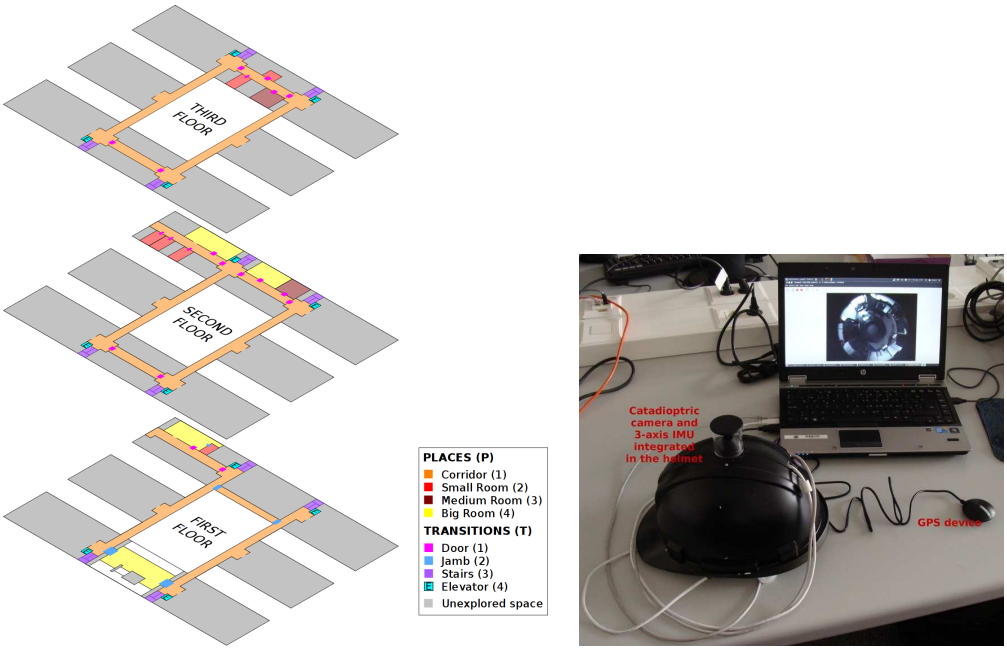
All images have been manually labeled with the type of area where acquired and its position. Consecutive images labeled with the same type of area have been grouped into clusters. Table 2 shows the number of clusters and between parentheses the number of images of each type.

5.2. Image representation evaluation

This first set of experiments is designed to evaluate how suitable and discriminative for our problem the image representation described is. These experiments evaluate different environment models and how they work classifying the rest of the images into *Places* and *Transitions*, as well as the detailed classification into subclasses.

As said before a key element of our labeling process is the reference model used. Then, we have tried to build this model automatically to avoid any bias with hand made selections. The basic model created from the dataset, let us name it *One*-Cluster-Model, includes only the first cluster of each subclass found in the sequence. The second model evaluated, named n-Cluster-Model, includes a variable amount of clusters considered as reference for each subclass, depending on the occurrence of each class and subclass.

In practice, the value of n_i used for the n-Cluster-Model is 5 for Corridor, 2 for Big room, 1 for Medium room and 1 for Small room in the case of *Places*. In the case of *Transitions* the value of n_i is 10 for Door, 2 for Jamb, 1 for Stairs and 1 for Elevator. 4351 images are used to build this model, while the amount of images used to create the *One*-Cluster-Model is about half of that value (2208 images). The



(a) (b)

Figure 5: (a) Map of the building where the dataset has been acquired, different colors mean different type of area traversed. (b) Acquisition system: an omnidirectional catadioptric camera mounted on a helmet.

amount of representative descriptors per class k is set to 25 after some initial evaluations of the approach. Then, the environment model is formed by 200 representative Gists (2 Classes $\times$ 4 Subclasses/Class $\times$ k).

The test images used to evaluate the approach in the following experiments are all images in the dataset not used to create the model. We run a Naive Bayes Classifier based on the likelihood function described in eq. (3), that assigns a label to each image independently of the rest of images. It is a simple probabilistic classifier based on applying Bayes' theorem under independence assumptions. The formulation of the Naive Bayes Classifier in our case and following the nomenclature used for the formulation of the Hidden Markov Model is:

$$p(S_t = i|I_t) = \alpha \;\; p(S_t = i) \prod_{j=1}^{n} p(I_t|S_t = i), \tag{7}$$

again, $p(S_t = i|I_t)$ is the posterior probability of the event $S_t = i$ given the image I_t, $p(S_t = i)$ is the prior probability of the class i and $p(I_t|S_t = i)$ is the likelihood function (eq. 3). We set the same prior probability for each class: $p(S_t = i) = 0.5$ with $i \in [P, T]$.

The results of this classification using the *One*-Cluster-Model can be seen in Table 3a and Table 3b shows the results using the n-Cluster-Model. Each row contains the percentage of tests corresponding to a label correctly classified or wrong labeled with the other type. The accuracy is computed as the sum of all the correct classifications divided by the total number of classifications. The classification using any of the models works better for *Places* (P) than for *Transitions* (T) and, as it could be expected, the simple model is less powerful to represent the environment that the n-Cluster-Model, with around 5% higher accuracy. There are additional reasons to use the second model: first, indoor environments use to include more areas of some classes than others, e.g., in the building of the tests there are more doors than stairs or elevators; second, some areas of the same subclass can be very different, e.g., the hall of the building and a research laboratory are both classified as Big Rooms. The n-Cluster-Model is kept for the rest of the experiments as reference model.

Besides the basic *Place*/*Transition* segmentation, we want to test how the proposed image representation works to classify the images into the considered subclasses. Following a similar approach, using our hand labeled ground truth, we classify all the images from each class (P or T) into the corresponding

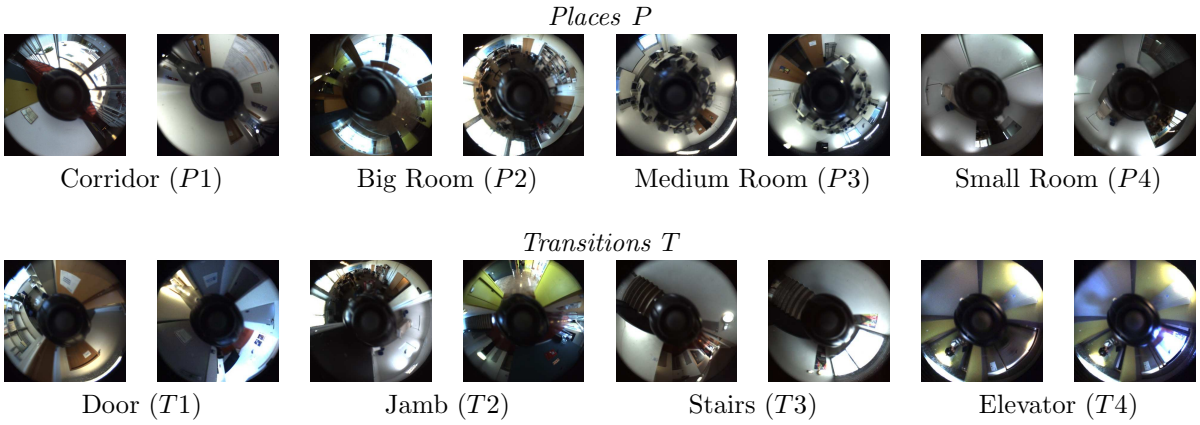


Figure 6: Examples of images labeled in the ground truth as elements of the different classes and subclasses.

Table 3: Labeling results evaluating each test independently from the rest of the sequence with a Naive-Bayes Classifier. Top: results for *Place* (P) - *Transition* (T) classification using two different reference models. Bottom: subclasses classification results using the best performing reference model.

(a) *One*-Cluster-Model

	P	**T**
P	75.02	24.98
T	43.79	56.21
Accuracy: 71.51		

(b) n-Cluster-Model

	P	**T**
P	80.89	19.11
T	39.89	60.11
Accuracy: 76.82		

(c) *n-Cluster-Model for Place subclasses*

	P1	**P2**	**P3**	**P4**
P1	93.04	1.78	0.46	4.72
P2	28.48	61.13	10.31	0.08
P3	32.26	0.98	42.46	24.30
P4	43.90	20.00	2.05	34.05
Places Accuracy: 82.95				

(d) *n-Cluster-Model for Transition subclasses*

	T1	**T2**	**T3**	**T4**
T1	69.31	2.01	3.41	25.28
T2	15.42	44.71	25.33	14.54
T3	0.00	0.00	100.00	0.00
T4	0.54	0.00	1.09	98.37
Transitions Accuracy: 82.41				

subclasses ($P1/P2/P3/P4$ or $T1/T2/T3/T4$). Tables 3c and 3d show the results of this experiment. Looking to the results for *Places* we can observe acceptable average values for the accuracy in the labeling, however there are big differences in the results at different subclasses (almost all corridor images ($P1$) are well classified, but only 34.05% of small rooms ($P4$) were labeled correctly. The misclassified rooms (about 30% for each room subclass) are usually classified as corridors. The poor results obtained for the classification among different rooms means that the descriptor is not discriminative enough to distinguish well between these subclasses. Table 3d shows the results for the classification of *Transitions* with also heterogeneous results for different subclasses but acceptable average accuracy above 80%. Conclusion after these results is that the representation proposed gives acceptable results to augment topological representations, but there are chances of better performance if we achieve a more discriminative representation for particular subclasses.

5.3. Testing the mapping method

Previous subsection shows the accuracy of the labeling classifier: around 76% when classifying into *Places* or *Transitions* and around 82% when labeling one of the basic classes into one of its subclasses. This subsection summarizes our experiments to validate the whole mapping method proposed in Section 4.2.

First we evaluate the effect of including temporal consistency on the label assignment along the sequence. We compare results using the Hidden Markov Model (HMM) to decide the most likely class/-

Table 4: Labeling results evaluating the probability of each class according to the HMM including or not the *minSize* filter. Top: results for *Place* (P) - *Transition* (T) classification. Bottom: subclasses classification results.

(a) P/T Classification without *minSize* filter

	P	**T**
P	82.87	17.13
T	39.86	60.14
Accuracy: 78.42		

(b) P/T Classification including *minSize* filter

	P	**T**
P	78.07	21.93
T	32.27	67.73
Accuracy: 76.04		

(c) Subclasses classification including *minSize* filter

	P1	**P2**	**P3**	**P4**	**T1**	**T2**	**T3**	**T4**
P1	76.61	0.70	0.09	1.63	7.48	0.56	4.13	8.79
P2	22.18	43.94	0.00	0.00	15.06	18.82	0.00	0.00
P3	47.77	0.00	49.16	0.00	0.56	0.00	2.51	0.00
P4	21.95	16.10	0.00	30.97	20.41	9.74	0.00	0.82
T1	33.30	0.00	10.53	1.71	41.12	1.71	0.00	11.63
T2	60.35	11.01	0.00	8.37	0.00	20.26	0.00	0.00
T3	2.31	0.00	0.00	13.38	0.00	0.00	84.31	0.00
T4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00

Table 5: Number of clusters generated with the mapping approach with and without *minSize* filter.

	HMM	HMM + *minSize*	GT
# clusters	267	180	111
Minimum cluster size (# images)	1	7	19

subclass instead of the Naive Bayes Classifier evaluation. The HMM requires to adjust the probability of a transition to happen. In [23] the authors propose a system to automatically adjust the value of this probability based in the training data. We test this system, what give us values of $p(S_t = i|S_{t-1} = j)$ higher than 0.99 when $j = i$. The effect of this values is a benefit in the detection of *Places* to the detriment of the detection of *Transistions*. We set the value of $p(S_t = i|S_{t-1} = j)$ to 0.9 when $j = i$ and $p(S_t = i|S_{t-1} = j)$ to 0.1 when $j \neq i$ to obtain results more adequate to our objective. Using the HMM probability evaluation to assign the labels, *Places* and *Transitions*, we can see a slight improvement, as can be seen in Table 4a compared to previous results in Table 3b.

Secondly, experiments summarized in table 4, compare results including or not the *minSize* filter explained in Section 4.2, in the tables 4a and 4b respectively. This filter compares the Gists distance between the first and the last images on the current cluster and check it with a similarity threshold, the distance must be over this threshold (th) to create a new cluster.

The images are classified with HMM and, as explained, they are grouped into clusters according to the assigned class: consecutive images fitting the conditions are grouped together. We can appreciate similar average accuracy in the P/T classification with or without taking into account the *minSize* filter. However the fact of avoiding too small clusters turns into a more meaningful semantic partition of the environment as detailed later. Detailed results of classification into subclasses are only shown for the complete approach, including the *minSize* filter, in table 4c. Results without using this filter were very similar, slightly better for subclasses of Places but slightly worse for subclasses of Transitions.

Table 5 shows the size and number of clusters we generate with the two options and the cluster arrangement done manually as ground truth when labeling the images. We have set empirically the *minSize* threshold to 0.275 for all tests. Note that using the *minSize* threshold most of the extremely small clusters are eliminated and the map obtained is more similar to the manually labeled map. Besides, as described before, this option provided better accuracy for Transitions, that is the particular labels we are interested the most.

Table 6: Map areas automatically detected (of # in the Ground Truth)

Places	$P1$	$P2$	$P3$	$P4$	TOTAL
	30 (of 38)	5 (of 7)	2 (of 3)	4 (of 8)	41 (of 56)
Transitions	$T1$	$T2$	$T3$	$T4$	TOTAL
	22 (of 40)	3 (of 9)	4 (of 4)	2 (of 2)	31 (of 55)

Another interesting comparison run was to analyze the usefulness of doing jointly the semantic labeling and the topological clustering. We evaluated the results of the individual location labeling with or without getting a common sub-class label for all images in each cluster. We obtained improvements in the labeling results running both steps simultaneously and assigning a common label to all components in a topological cluster. This is not surprising, since by grouping images we take into account the subclass of all the images in the cluster as a group, so we filter some misclassification errors.

Finally, summarizing the experimental validation, Fig. 7 shows the trajectory of the sequence with the mapping results. This result is obtained with the whole sequence to obtain a representation of the whole environment. Then as the images used to estimate the model are included now, we observe higher accuracy values: 81.83% for the classification into *Places* and *Transitions*, 71.70% for the classification into *Places* subclasses and 74.37% for the classification into *Transitions* subclasses. Fig. 7(a) shows the manual segmentation into clusters and their ground truth class label, and Fig. 7(b) shows the segmentation after running our approach. Comparing both segmentations we can see where errors occur. Regarding *Places* detection, as previously observed, corridors are much clearly recognized than the different types of rooms. In the case of *Transitions*, the higher errors occur for Jambs (blue), that are present only in the first floor and are not detected, so the corridors that should be separated by them are joined in one cluster. Some errors also occur in the classification of corridors due to the creation of inexistent transitions. These errors may be happening because of rapid illumination changes that produce big appearance changes and artifacts in the images.

All previous classification evaluations have been estimated considering the individual labeling of each image. However, the objective when creating a semantic map is to correctly detect the different areas of the environment. Despite some mistakes, the map created captures the distribution of the areas of the building. Table 6 shows the number of areas detected according to their class and subclass. We consider an area detected by our approach when 50% of the images in that area have been correctly labeled. Usually the problem is that the generated clusters are still smaller than the ground truth annotated ones, that is why we consider correct detections even if only a part of the hand labeled images in the region are correctly classified.

6. Conclusions and future work

This work presents a novel indoor semantic place labeling method that includes information of the basic indoor scenes. The method uses catadioptric images and the adaptation of the Gist global descriptor to represent these images. The general idea proposed is to simultaneously run a topological map building approach and the classifier to label the different types of indoor scenes considered. We have described an approach to label different types of *Places* and *Transitions*. The result of our method is a semantic-topological model, where the nodes are *Places* and the edges are *Transitions* between *Places*, including information about different types of *Places* (Big, Medium or Small room, Corridors) and *Transitions* (Door, Jamb, Stairs, Elevator). A detailed semantic analysis of the types of transitions is not common although could provide important information to later uses of the map. Our approach is based on this semantic classification of the images, using a simple environment model, integrated with a Hidden Markov Model framework to add spatio-temporal consistency.

We performed the experimental validation of our approach using the new Wearable OmniCam dataset acquired for this work. First, we show good accuracy in the labeling of a sequence into classes and subclasses. A second group of experiments evaluates qualitatively the approach. They demonstrate the advantages of including the spatio-temporal framework and show the type of indoor topological models that can be obtained. Despite the simple and efficient image representation proposed and the difficulty of the dataset, acquired from a camera on a helmet while the person walks normally, the map obtained is quite close to the ground truth manually generated.

For future work, it is necessary to evaluate if this representation is valid not only for class labeling but also for loop detection, to provide a more consistent representation of the environment when revisiting a certain place. The step from our proposal that could be improved is the subclass classification. Using only the Gist based representation seems not enough sometimes, e.g., when trying to distinguish small rooms from other rooms or jambs from doors. Therefore, future work should include additional image features that allow us to distinguish among indoor scenes with similar structure but small details that may provide important differences in the semantic label.

Acknowledgement

This work has been supported by the Spanish project DPI2009-14664-C02-01 including FEDER founds.

References

[1] A. Tapus, R. Siegwart, Incremental robot mapping with fingerprints of places, Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2005) 2429–2434.

[2] S. Vasuvedan, S. Gachter, V. Nguyen, R. Siegwart, Cognitive maps for mobile robots - an object based approach, Robotics and Autonomous Systems 55 (5) (2007) 359–371.

[3] H. Zender, O. M. Mozos, P. Jensfelt, G.-J. Kruijff, W. Burgard, Conceptual spatial representations for indoor mobile robots, Robotics and Autonomous Systems 56 (6) (2008) 493–502.

[4] E. A. Topp, H. I. Christensen, Detecting region transitions for human-augmented mapping, IEEE Transactions on Robotics 26 (4) (2010) 715–720.

[5] A. C. Murillo, P. Campos, J. Kosecka, J. J. Guerrero, Gist vocabularies in omnidirectional images for appearance based mapping and localization, in: 10th OMNIVIS, held with Robotics: Science and Systems (RSS), 2010.

[6] A. Oliva, A. Torralba, Building the Gist of a scene: The role of global image features in recognition, Progress in brain research 155 (2006) 23–36.

[7] A. Rituerto, L. Puig, J. J. Guerrero, Visual slam with an omnidirectional camera, in: 20th International Conference on Pattern Recognition, 2010, pp. 348–351.

[8] C. Stachniss, O. Martinez-Mozos, A. Rottman, W. Burgard, Semantic labeling of places, in: Proc. of the International Symposium of Robotics Research (ISRR), 2005.

[9] O. Saurer, F. Fraundorfer, M. Pollefeys, Visual localization using global visual features and vanishing points, in: Conference on Multilingual and Multimodal Information Access Evaluation (CLEF 2010), 2010.

[10] A. Pronobis, O. M. Mozos, B. Caputo, P. Jensfelt, Multi-modal semantic place classification, The International Journal of Robotics Research (IJRR), Special Issue on Robotic Vision 29 (2–3) (2010) 298–320.

[11] A. Ranganathan, Pliss: Detecting and labeling places using online change-point detection, in: Proceedings of Robotics: Science and Systems, 2010.

[12] S. Thrun, A. Bucken, Integrating grid-based and topological maps for mobile robot navigation, in: Proc. of the National Conference on Artificial Intelligence, 1996, pp. 944–950.

[13] J. Santos-Victor, R. Vassallo, H. Schneebeli, Topological maps for visual navigation, in: International Conference on Computer Vision Systems, 1999, pp. 21–36.

[14] A. Nüchter, J. Hertzberg, Towards semantic maps for mobile robots, Robotics and Autonomous Systems 56 (11) (2008) 915–926.

[15] B. Kuipers, The Spatial Semantic Hierarchy, Artificial Intelligence 119 (1-2) (2000) 191–233.

[16] N. Tomatis, I. Nourbakhsh, R. Siegwart, Hybrid simultaneous localization and map building: a natural integration of topological and metric, Robotics and Autonomous systems 44 (1) (2003) 3–14.

[17] A. C. Murillo, C. Sagüés, J. J. Guerrero, T. Goedemé, T. Tuytelaars, L. Van Gool, From omnidirectional images to hierarchical localization, Robotics and Autonomous Systems 55 (5) (2007) 372–382.

[18] C. Galindo, J. Fernández-Madrigal, J. González, A. Saffiotti, Robot task planning using semantic maps, Robotics and Autonomous Systems 56 (11) (2008) 955–966.

[19] M. Cummins, P. Newman, Highly scalable appearance-only slam - fab-map 2.0, in: Proc. of Robotics: Science and Systems (RSS), 2009.

[20] B. Douillard, D. Fox, F. Ramos, H. Durrant-Whyte, Classification and semantic mapping of urban environments, International Journal of Robotics Research 30 (1) (2011) 5–32.

[21] O. Martínez Mozos, W. Burgard, Supervised learning of topological maps using semantic information extracted from range data, in: Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2006, pp. 2772–2777.

[22] S. Friedman, H. Pasula, D. Fox, Voronoi random fields: extracting the topological structure of indoor environments via place labeling, in: Proc. of the 20th International Joint Conference on Artifical intelligence, 2007, pp. 2109–2114.

[23] A. Rottmann, O. Mozos, C. Stachniss, W. Burgard, Semantic place classification of indoor environments with mobile robots using boosting, in: Proceedings of the 20th national conference on Artificial intelligence, 2005, pp. 1306–1311.

[24] A. Pronobis, B. Caputo, P. Jensfelt, H. I. Christensen, A realistic benchmark for visual indoor place recognition, Robotics and Autonomous Systems 58 (1) (2010) 81–96.

[25] A. Ramisa, A. Tapus, D. Aldavert, R. Toledo, R. Lopez De Mantaras, Robust vision-based robot localization using combinations of local feature region detectors, Autonomous Robots 27 (4) (2009) 373–385.

[26] P. Viswanathan, T. Southey, J. Little, A. Mackworth, Place classification using visual object categorization and global information, in: 2011 Canadian Conference on Computer and Robot Vision, 2011, pp. 1–7.

[27] C. Nieto-Granda, J. Rogers, A. Trevor, H. Christensen, Semantic map partitioning in indoor environments using regional analysis, in: Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2010, pp. 1451–1456.

[28] M. Nieuwenhuisen, J. Stückler, S. Behnke, Improving indoor navigation of autonomous robots by an explicit representation of doors, in: International Conference on Robotics and Automation, 2010, pp. 4895–4901.

[29] A. Quattoni, A. Torralba, Recognizing indoor scenes, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 413–420.

[30] A. Oliva, A. Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope, International Journal of Computer Vision 42 (3) (2001) 145–175.

[31] A. C. Murillo, J. Kosecka, Experiments in place recognition using Gist panoramas, in: 9th OMNIVIS, held with ICCV, 2009, pp. 2196–2203.

[32] P. E. Rybski, F. Zacharias, J.-F. Lett, O. Masoud, M. Gini, N. Papanikolopoulos, Using visual features to build topological maps of indoor environments, in: IEEE International Conference on Robotics and Automation (ICRA), 2003, pp. 850–855.

[33] L. Puig, J. J. Guerrero, K. Daniilidis, Topological map from only visual orientation information using omnidirectional cameras, in: IEEE International Conference on Robotics and Automation (ICRA), Workshop on Omnidirectional Robot Vision, 2010.

[34] T. Goedemé, M. Nuttin, T. Tuytelaars, L. Van Gool, Omnidirectional vision based topological navigation, International Journal of Computer Vision 74 (3) (2007) 219–236.

[35] T. Deselaers, D. Keysers, H. Ney, Features for image retrieval: an experimental comparison, Information Retrieval 11 (2) (2008) 77–107.

[36] A. Angeli, D. Filliat, S. Doncieux, J.-A. Meyer, A fast and incremental method for loop-closure detection using bags of visual words, IEEE Transactions On Robotics, Special Issue on Visual SLAM 24 (5) (2008) 1027–1037.

[37] L. Puig, Y. Bastanlar, P. Sturm, J. Guerrero, J. Barreto, Calibration of central catadioptric cameras using a dlt-like approach, International Journal of Computer Vision, IJCV 93 (1) (2011) 101–114.

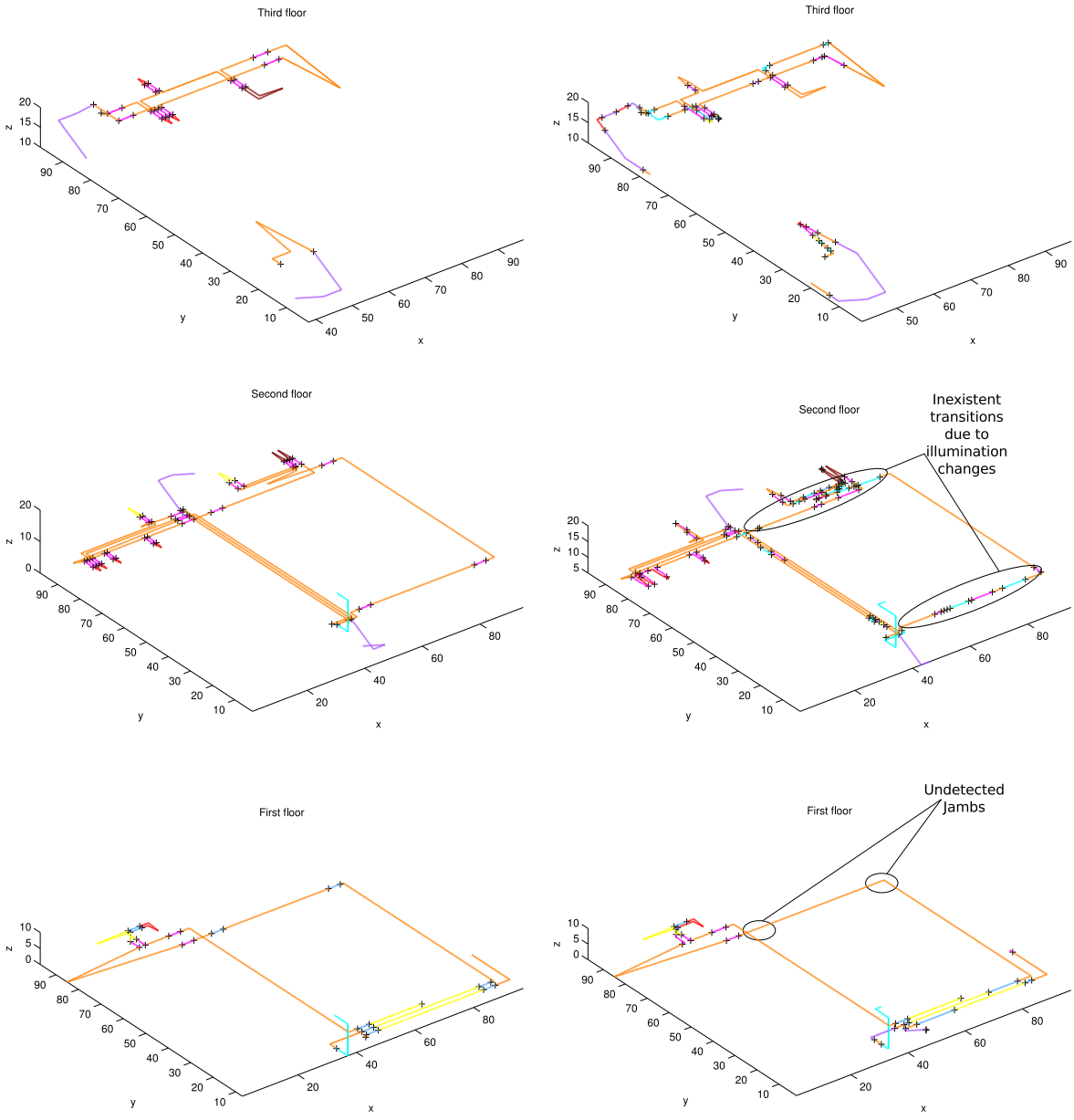


Figure 7: Segmentation of the trajectory in clusters of subclasses: (a) Manual, (b) Complete approach. The start position of each cluster is marked with a black cross. One color for each subclass: *Places*: (Orange) Corridor, (Yellow) Big Room, (Brown) Medium Room, (Red) Small Room
Transitions: (Pink) Door, (Blue) Jamb, (Purple) Stairs, (Light Blue) Elevator