Departamento de
Informática e Ingeniería
de Sistemas
Universidad Zaragoza

Trabajo de Fin de Máster

Augmented Indoor Hybrid Maps using Catadioptric Vision

Alejandro Rituerto Sin

Directores: José Jesús Guerrero Campo y Ana Cristina Murillo Arnal

**Máster Universitario en
Ingeniería de Sistemas e Informática**

Departamento de Informática e Ingeniería de Sistemas
Área de Ingeniería de Sistemas y Automática
Centro Politécnico Superior
Universidad de Zaragoza

Septiembre de 2011

Augmented Indoor Hybrid Maps using Catadioptric Vision

RESUMEN

Uno de los principales retos a la hora de diseñar robots y sistemas autónomos es la necesidad de una representación del entorno en el que van a trabajar. Partiendo de los datos recogidos por los diferentes sensores, un sistema autónomo debe ser capaz de construir un modelo de su entorno adecuado para las tareas a realizar. Determinadas tareas requieren altos niveles de detalle, representaciones del entorno en las que las distancias entre los distintos obstáculos y el sistema robótico estén bien definidas. Estos modelos se suelen denominar mapas métricos [4, 40, 8]. En otros casos no es necesario un nivel de detalle alto, son más útiles mapas con mayor nivel de abstracción, como los mapas topológicos, o con mayor información del entorno, como los mapas semánticos [54, 46, 33, 53, 59]. Estos mapas de menor nivel de detalle y mayor abstracción han sido muy utilizados últimamente como niveles superiores en una jerarquía de mapas [22, 57, 29]. Los mapas jerárquicos, o mapas híbridos, ya que unen distintos tipos de mapas, ordenan la información capturada del entorno en distintos niveles: los niveles superiores suelen estar formados por mapas con mayor abstracción y los niveles inferiores por mapas con mayor detalle.

El objetivo a largo plazo en el que se incluye el trabajo aquí presentado es el diseño de un modelo de mapa jerárquico para la ayuda en la navegación de personas (tanto para la asistencia a personas con deficiencias visuales, como asistente visual general en entornos nuevos para el usuario, por ejemplo una ciudad nueva o un edificio público desconocido) utilizando una cámara omnidireccional como sensor de entrada. Al tratarse de una aplicación de asistencia debe incluir información semántica con significado para las personas (mapa semántico), al mismo tiempo que debe ser capaz de dirigir por la trayectoria correcta al usuario (mapa métrico).

En los últimos años el uso de cámaras como sensores principales en tareas de robótica ha aumentado gracias a la mejora de la capacidad de procesamiento de los ordenadores, que permite trabajar con fluidez con la gran cantidad de información que contienen las imágenes. En este trabajo, como ya se ha apuntado, se ha utilizado un tipo particular de cámara: un sistema de visión omnidireccional. La gran ventaja de este tipo de cámaras es que permiten capturar en una sola imagen 360^o de campo de vista. Sin embargo presentan algunos problemas como la gran distorsión de las imágenes y la presencia en la imagen de partes de la cámara el espejo.

En este Trabajo de Fin de Máster se presenta un nuevo método para crear mapas semánticos a partir de secuencias de imágenes omnidireccionales. El objetivo es diseñar el nivel superior de un mapa jerárquico: mapa semántico o mapa topológico aumentado, aprovechando y adaptando este tipo de cámaras. La segmentación de la secuencia de imágenes se realiza distinguiendo entre Lugares y Transiciones, poniendo especial énfasis en la detección de estas Transiciones ya que aportan una información muy útil e importante al mapa. Dentro de los Lugares se hace una clasificación más detallada entre pasillos y habitaciones de distintos tipos. Y dentro de las Transiciones distinguiremos entre puertas, jambas, escaleras y ascensores, que son los principales tipos de Transiciones que aparecen en escenarios de interior. Para la segmentación del espacio en estos tipos de áreas se han utilizado solo descriptores de imagen globales, en concreto Gist [35]. La gran ventaja de usar este tipo de descriptores es la mayor eficiencia y compacidad frente al uso de descriptores locales. Además para mantener la consistencia espacio-temporal de la secuencia de imágenes, se hace uso de un modelo probabilístico: Modelo Oculto de Markov (HMM). A pesar de la simplicidad del método, los resultados muestran cómo es capaz de realizar una segmentación de la secuencia de imágenes en clusters con significado para las personas. Todos los experimentos se han llevado a cabo utilizando nuestro nuevo data set de imágenes omnidireccionales, capturado con una cámara montada en un casco, por lo que la secuencia sigue el movimiento de una persona durante su desplazamiento dentro de un edificio. El data set se encuentra público en Internet [1] para que pueda ser utilizado en otras investigaciones. Este trabajo se encuentra en proceso de revisión en la revista Robotics and Autonomous Systems, en un número especial titulado Semantic Perception, Mapping and Exploration.

Tras este trabajo quedan varias líneas de trabajo abiertas. En primer lugar nuestra propuesta para crear mapas topológicos ha mostrado buenos resultados utilizando tan solo Gist como descriptor, sin embargo para la clasificación semántica de algunos tipos de Transiciones y Lugares el Gist no es suficiente por lo que para mejorar los resultados se está estudiando la inclusión de nuevos descriptores. Del mismo

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

modo en cualquier sistema de navegación es de vital importancia el cerrado de bucle que permite reconocer lugares visitados previamente. La detección de revisitas utilizando tan solo Gist no ha dado resultados suficientemente satisfactorios, por lo que la inclusión de nuevas características y descriptores puede ser de gran utilidad.

Otros trabajos realizados muy cercanos al contenido de esta memoria, y con el mismo objetivo común a largo plazo, fueron los realizados durante mi Proyecto de Fin de Carrera [50] y a lo largo de una beca de Iniciación a la Investigación concedida por el I3A. Durante este periodo trabajé en la adaptación de un algoritmo básico de SLAM visual monocular para que pudiera utilizar imágenes omnidireccionales. A partir de una aplicación de SLAM visual para cámaras convencionales basada en el Filtro de Kalman Extendido (EKF) se integró el Modelo de la Esfera [14, 2] como método de proyección para las cámaras omnidireccionales. Además de incluir el Modelo de la Esfera dentro del EKF se realizaron distintas modificaciones para utilizar características SIFT [24] como puntos de interés de la imagen. A partir de este trabajo surgieron dos publicaciones: [44], publicado en International Conference on Pattern Recognition - 2010, el que explica la adaptación del algoritmo de SLAM visual para cámaras omnidireccionales, y [43], premiado como "Best Paper Award" en 10^{th} OMNIVIS, en él se realiza una comparación del funcionamiento de dicha aplicación de SLAM entre visión omnidireccional y visión convencional demostrando las ventajas que supone utilizar cámaras omnidireccionales para SLAM. Este último artículo se incluye como apéndice de este documento, Anexo **??**. Continuando este trabajo, durante este curso he codirigido un Proyecto de Fin de Carrera de Ingeniería Industrial [18] para realizar la adaptación a visión omnidireccional de una aplicación de SLAM visual que trabaja en tiempo real, siguiendo los pasos de los trabajos previos mencionados e incluyendo mejoras en los descriptores de puntos de la imagen específicos para las cámaras omnidireccionales. Éste trabajo ha dado lugar a otro artículo [17] enviado al 11^{th} OMNIVIS, actualmente en proceso de revisión.

Índice de Contenidos

TÍTULO Y ABSTRACT DEL ARTÍCULO:

Augmented topological mapping with semantic indoor labeling using a wearable catadioptric vision system

Abstract-Current research on appearance based mapping goes towards richer semantic models of the environment, which may allow the robots to perform higher level tasks and provide better human-robot interaction. This work presents a new omnidirectional vision based approach for augmented topological mapping. Omnidirectional vision systems are of particular interest for vision based mapping because they allow to have more compact and efficient representation of the environment. Our proposal includes some 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 as part of the criteria to segment the environment. This work is built on efficient catadioptric image representation based on the global Gist descriptor, which is used to classify the acquired views into types of indoor regions. The considered basic types of indoor regions are *Places* and *Transitions*, further divided into more specific subclasses, e.g., *Transitions* into doors, stairs and elevator. Besides using the result of this labeling, the proposed mapping approach includes a probabilistic model to account for temporal consistency. All the proposed ideas have been demonstrated and evaluated in a new indoor dataset acquired with our wearable catadioptric vision system[2] with promising results in a realistic prototype.

[2] `http://robots.unizar.es/omnicam/`

Sección 1

Introduction

For most robotic 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 enough informative, therefore higher abstraction level maps can be built, such as topological or object-based maps [53, 59, 61, 58].

The general goal of this work is to achieve a useful semantic-topological map for indoor environments using a wearable catadioptric vision system. In particular we aim to include interesting semantic information on indoor topological models and to design a simple approach that could be run on-line and be used on the wearable system. We describe the catadioptric images following the approach described in [27], which is based on the global Gist descriptor [35] and adapted to omnidirectional images. The long term goal is to incorporate the presented augmented topological model to a set of small lower level metric maps of each topological region, which could be obtained with standard visual odometry or slam algorithms [44].

This work describes our proposal for a new on-line topological map building method with the following two novel ideas with regard to other related works. On-line approaches for topological mapping, usually build the map evaluating the similarity within consecutive images and establishing different criteria to decide when and where to segment the trajectory into new "clusters" of the topological model. We define a simple way to include the labeling from a semantic 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 examples of the different classes to be recognized. Besides, we find that most of the approaches for semantic indoor scene labeling try to label types of *Places* [51, 47, 37]. We additionally propose to pay attention to 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. For example, we can choose a suitable robot team member to go to a particular destination, or give different instructions in case of human assistance systems. Two other interesting properties of our method, that are not novel themselves but their inclusion is a basic step 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 temporal consistency of the labeling along the trajectory.

In spite of the simplicity of the representation, the proposal gets to partition the environment into semantic meaningful areas for humans, as it can be seen later using the presented catadioptric dataset. The rest of the paper is organized as follows. First we compare our proposal with related works on semantic and topological maps in Section 2, and Section 3 details the image representation used. Section 4 provides an in-depth description of the *Places* and *Transitions* classifier and the mapping 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.

Sección 2

Related Work

Topological modeling of the environment is a subject already studied for long [54, 46]. Initially, these models presented huge possibilities due to its 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 [33]. For example adding to the topology information about places [53] and/or object information [59]. Topological maps are many times built on the top of a hierarchy of different levels [22], e.g., a global topological map that connects smaller local metric maps [57]. An extensively used solution to achieve different efficiency and accuracy results at the different levels is to include global and local image features through the different steps of the hierarchy [29].

In the last years, research on topological models works to augment them with semantic labels and conceptual models of the environment. These semantic models, rather than pure metric models, are more suitable for human-robot interaction [61, 58] and allow us to achieve more complicated goals [12]. These provides new opportunities to increase the autonomy and reasoning skills of our 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 [7, 10], to build topological models that include place or object recognition information. In [7], which deals with place recognition, the authors present an approach for appearance based mapping in extremely large datasets (1000 km) that efficiently recognizes the revisit of known places. The work in [10] 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 many proposals using different types of sensors to interpret semantic information that will be included in a topological map. Initial related proposals were typically achieved using range data, to learn a room-doorway-hallway structure indoors in [25] or [11]. We find proposals also using a combination of range and vision cues, for example in [61] they combine place and object recognition in exploration and semantic mapping approach. The work in [37] suggests a Support Vector Machine (SVM) scheme that learns how to optimally combine and weight each cue.

Other recent 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 specific semantic labels only regarding Places, e.g., office, corridor, kitchen... [36], considering all transitions as just connections between places. We find different types of approaches that then try to classify these types of places, with multiple proposals both for the way of learning the labels to be recognized and for the way the images are represented. Some works propose to work with local features, such as the robust vision-based robot localization using combinations of local features from [42], or the work in [60] that presents an integration of object detection, using local features, and global image properties for place classification. Some of these proposals for semantic indoor mapping are constantly trained, or even simultaneously run, with human supervision to achieve a representation closer to human concepts [58, 31]. Others use weaker human supervision to obtain a few labeled samples, such as the work in [47]. It tries to learn the representation on 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 them detect when those cases occur and 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. 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 recent works [32]. also paying attention to transitions that may be or not valid, in their case only doors, proposing to dynamically model the environment to react if a transition is suddenly closed. If we have a model with the transition information we can decide in advance where we may have problems if we can not open doors for instance, and plan alternative routes.

Besides the augmented mapping approaches, we find additional closely related works regarding the more general problem of place or scene recognition indoors [41]. This works also points the idea that plenty of indoor areas are usually not considered, among the big set we can find elevators and stairs, at different levels of difficulty to be detected. In this work we include indoor places that can actually be considered as a transition between places (besides elevators and stairs we include areas under a door and under jambs). Another common point with that work is the use of the GIST descriptor [35], initially presented for classifying outdoor scenes [34] and used in more recent work together with additional cues for indoor scene recognition [41].

Our global image descriptor is based only on the global GIST descriptor, following the ideas initially presented in [28]. Global descriptors are known to be more efficient and compact, but usually less robust and discriminative, than local features. However in this work [28] 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 [45, 39] panoramic cameras are used for indoor topological map building and [15] 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. This together with the fact that we want to use a wearable system, requires a carefully designed image representation detailed in the following section.

Sección 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 [9]. Therefore, to deal with large quantities of images for tasks where efficiency is an issue and is not required a detailed analysis of image content, a global representation is preferred.

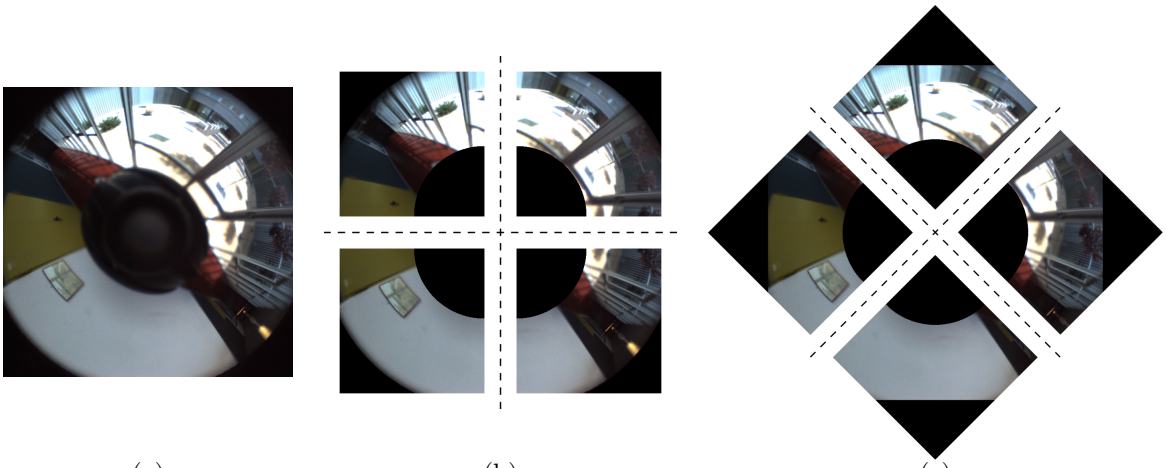
In this work we use the Gist descriptor [34], a holistic image representation or global image feature. In particular, it is a low dimensional representation of the scene captured in an image which corresponds to the mean response to steerable filters at different scales and orientations computed over 4x4 sub-windows. The descriptor consists of a vector of 320 components for each color band used, so in a RGB image it is formed by 960 components.

This global feature was presented and extensively 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 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 (or robot). 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 [27]. As explained in this work, each image part is rotated to a canonical orientation before compute the Gist descriptor. Additionally we need to mask parts of the image where appear artifacts mostly produced by the reflection of parts of the catadioptric system in its own mirror. With this representation, the omnidirectional image Gist is composed by four conventional Gist descriptors, one computed for each image part (front, left, back and right): $\mathbf{g} = [g_f, g_l, g_b, g_r]$.

Fig. 3.1 shows an example of omnidirectional image and two possibilities to partition it: Direct partition (Fig. 3.1(a)) and Rotated partition (Fig. 3.1(b)), where the parts have been 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. Using the Direct partition we achieve invariance to vertical rotation angles multiple of 90^o and using both the Direct and the Rotated partition together we get invariance to rotation 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.

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 from the four possible permutations of the four sections of the image, which would hopefully provide us with the best alignment of the two evaluated images. Being $\mathbf{g}$ and $\mathbf{g}'$ the descriptors of two images, the distance



(a) (b) (c)

Figure 3.1: (a) Omnidirectional image acquired with the Wearable OmniCam system partitioned with the two partition methods analyzed: (b) Direct and (c) Rotated.

between them is:

$$dist(\mathbf{g}, \mathbf{g}') = \min_m(d_e(\mathbf{g}, \pi_m(\mathbf{g}'_{flbr})), \tag{3.1}$$

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

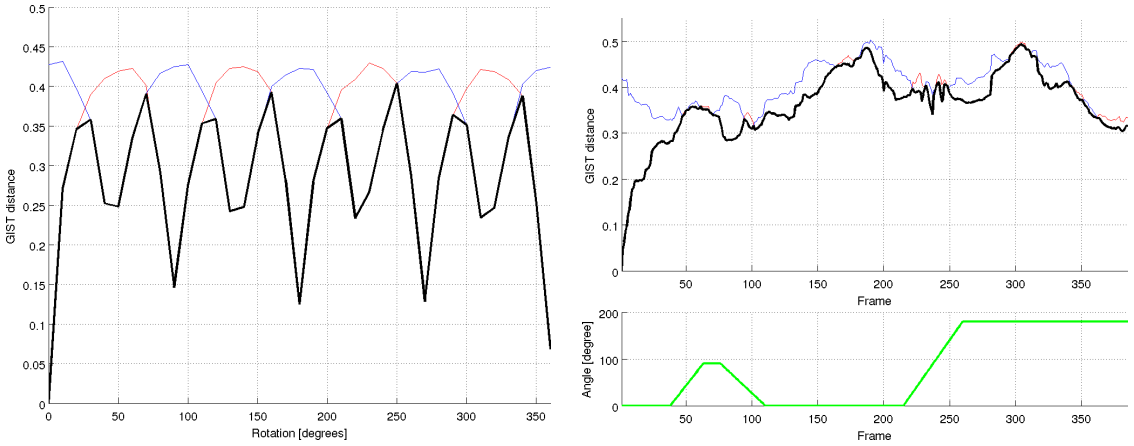
3.1 Rotation invariance analysis

To analyze in more detail the rotation invariance issues described above we have performed two experiments. For the first experiment we get 36 images equally distributed along a 360^o camera rotation movement without translation, around the vertical camera axis. 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 all the images with both the Direct ($\mathbf{g}_{Direct}$) and the Rotated ($\mathbf{g}_{Rotated}$) Gist descriptors of the reference 0^o image. Using a perfect rotation invariant representation all images would get exactly the same descriptor. Figure 3.2(a) shows the results of this test. The red line represents the distance between $\mathbf{g}_{Direct}$ in both the test and the 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 $\mathbf{g}_{Direct}$ of the test images and $\mathbf{g}_{Rotated}$ of the reference image. The black line represents the minimum value of red and blue line and confirms that using both partition methods we achieve invariance to rotation of angles multiple of 45^o.

The 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 further from the initial image. Figure 3.2(b) shows the results of this second test. The green line shows the orientation of the camera with respect to the initial frame of the sequence. The distance between test image descriptors $\mathbf{g}_{Direct}$ and reference image $\mathbf{g}_{Rotated}$ is most of the times bigger than the distance between $\mathbf{g}_{Direct}$ in both test and reference image.

When the orientation of the images change, i.e., the camera has been rotated, both lines have similar values. We see that the Gist distance variations are bigger due to the translation among the corridor than due to the rotation.

Therefore, as already mentioned, we can conclude that using the Direct partition method to compute the descriptor, we achieve rotation invariance to rotation of angles multiple of 90^o. Using a double



(a) (b)

Figure 3.2: Rotation invariance of the image representation proposed (a), and influence in the Gist distance in a trajectory section, (b). The red line shows the distance between the Direct partition Gists and the blue one the distance between the Direct partition Gist and the Rotated partition Gist. The black line shows the minimum value of both distances. The green line shows the angle of orientation respect to the reference frame.

representation for the reference images, storing both the descriptors obtained through Direct and Rotated partition methods, rotation invariance to angles multiple of 45^o is achieved. However, we can see there are not worth improvements using this duplicate representation while navigating indoors. Even if it is more robust to rotations, 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.

Sección 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 (*Places* and *Transitions*) of interest to discover the topology of the environment. Then, we evaluate the extension of 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 model of the environment. *Places* are the nodes of the model and the *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 further 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 4.1).

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 that 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.

The Environment Model

A basic step in our method is obtaining the Environment Model, to use it later as reference to classify new occurrences. The model is generated from a set of reference images from each class and subclass. 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 we consider the first n_i clusters of each class, where n_i is chosen depending on the frequency of occurrence of the class i. The model of each class initially consists of all $\mathbf{g}_{Direct}$ descriptors of the reference example images picked. 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. To avoid that this fact leads to unbalanced models towards *Place*, we use a standard k-means method to find the k Gists 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.

Table 4.1: 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 ($T4$)

Classify new occurrences according to the model

To classify new images we use a simple nearest neighbor like approach. To measure the likelihood of a single image being of a particular class, we compute the following likelihood function (4.1) based on the gist descriptor distance, and assign the maximum likelihood solution as label for the new image.

$$p(I^t|S_t = i) = \frac{k\, e^{\frac{-d_i}{\sigma^2}}}{\sum_{j=P,T} k\, e^{\frac{-d_j}{\sigma^2}}} \tag{4.1}$$

$S_t = i$ is the event of being in an area of class i at time t, when the image I^t is acquired, so $p(I^t|S_t = i)$ is the likelihood of the image I^t of being of class i. d_i is the minimum distance, using eq.(3.1), between the current image and the reference images from the model labeled as i. k and σ^2 are user defined gain and variance respectively.

The same evaluation is done to classify into the basic classes, *Places* (P) and *Transitions* (T), and classify into specific subclasses ($P1 \dots P4, T1 \dots T4$).

4.2 Augmented Topological map building

We develop our augmented map building method based on the image classifier described in the previous subsection.

Generally with catadioptric images, if we pay too much attention to consecutive image similarity, the topological segmentation is far from a semantic segmentation a human would manually do, containing lots of small clusters. This is because even consecutive catadioptric may present big visual differences, due to big image distortions and image changes specially with objects and scene elements close to the camera (as usually happens indoors). Our intention is using semantic labels as basic criteria to obtain semantic meaningful clusters in the topological model.

First, we integrate 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 [1]. 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:

$$j = \arg\max_{i \in \{T,P\}} \; p(S_t = i|I^t). \tag{4.2}$$

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:

$$\begin{aligned} p(S_t = i|I^t) =& \alpha \;\; p(I^t|S_t = i)p(S_t = i|I^{t-1}) = \\ =& \alpha \;\; p(I^t|S_t = i) \sum_{j=T,P} p(S_t = i|S_{t-1} = j)p(S_{t-1} = j|I^{t-1}), \end{aligned} \tag{4.3}$$

where α is a normalization term, and the conditional probability $p(I^t|S_t = i)$ is the likelihood function (eq. 4.1) 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 $i, j \in T, P$. We set the values to the probabilities of repeating the same event occurred at time $t-1$ in time t. 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 Augmented topological mapping method.

Input: Omnidirectional image sequence and environment model
Output: Augmented topological Map
  n = Number of the current cluster
  th = Similarity threshold
  $\mathbf{g}^i$ = Gist of component i from the model
  $\mathbf{g}^n$ = Gist of the first image of the current cluster n
  P_{t-1} = Probabilities at previous step
  while not end of sequence **do**
    // New image I^t
    $\mathbf{g}^t$ = OmnidirectionalGist(I^t)
    // Compute similarity with the current cluster
    $d = dist(\mathbf{g}^t, \mathbf{g}^n)$;
    // Compute probability of being transition or place
    $[p_P, p_T]$ = HMMEnvironmentModel($\mathbf{g}^t, \mathbf{g}^i$,$P_{t-1}$)
    if $p_P > p_T$ **then**
      $state = P$
    else
      $state = T$
    end if
    if $d > th$ & $state \neq state_{ncluster}$ **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

Algorithm 1 details the mapping 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 we fit a criteria to start a new cluster. This criteria is basically the likelihoods estimated from the HMM, but to prevent the appearance of too small clusters a criteria based on the similarity with the first image ($minSizefilter$) of the current cluster is included. If the Gist distance between the new image and the first image of the current cluster is below the similarity threshold established, the new image is kept in the current cluster, even if classification results according to HMM likelihood would put it in a different class. We will see the differences of using one or both of this criterion to build the topological map in next section.

This first mapping step just segments the environment model into *Places* or *Transitions*, the classification into subclasses is performed separately, and we try to take advantage of this first level classification to facilitate the more detailed classification into subclasses. Once an image is labeled as *Transition* or *Place*, we evaluate the subclass according to the reference image from that class most similar to the current image.

A cluster is composed by images of the same class, but initially may contain images labeled as different subclasses. We consider this is noise due to the fact that actually descriptors of some of the subclasses are pretty close and difficult to separate sometimes (doors and jambs for instance, are sometimes hard to distinguish for a human observer as well). Then, to assign the most likely subclass to the whole cluster, we compute the mode of the subclass label assigned to each image in the cluster. Finally, all the images in a cluster are labeled with the dominant subclass to compose the final augmented topological model of

the environment.

Sección 5

Experiments

In this section we present a new dataset acquired for 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. 5.1(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 [38]. However, the presented data has been acquired indoors, so GPS is deactivated. IMU data is also not used in this work, that is purely vision based approach, but could be used for future works.

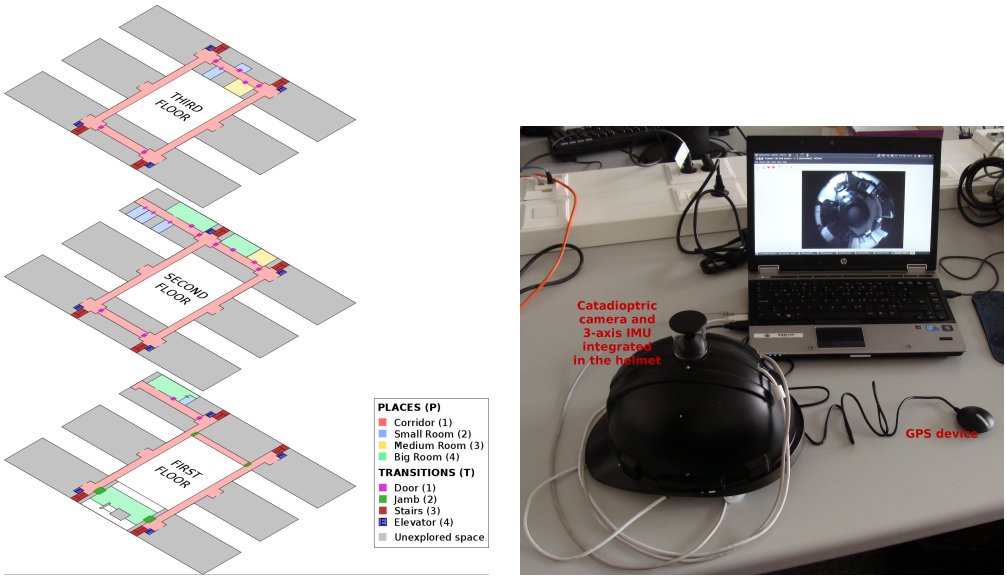
The dataset acquisition has been performed inside one building in Campus Río Ebro located in Zaragoza, Spain. The building has three floors and includes areas of different types: corridors, research laboratories, offices, classes, etc. The acquisition has been performed by a person wearing the helmet, so the 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 5.1(b) shows the planes 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 image 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 5.1 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 the image representation described is for our problem. 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.



(a) (b)

Figure 5.1: (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.

Table 5.1: Number of clusters of each class. The Values between parentheses are the number of images labeled with that class.

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

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, as explained previously in Section 4.1.

The tests 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. (4.1), 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:

$$p(S_t = i|I^t) = \alpha \;\; p(S_t = i) \prod_{j=1}^{n} p(I^t|S_t = i), \tag{5.1}$$

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. 4.1). 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 5.2a and Table 5.2b shows the results using the n-Cluster-Model. Each row contains the percentage of tests corresponding to a label correctly classified or confused 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**). However, as it could be expected,

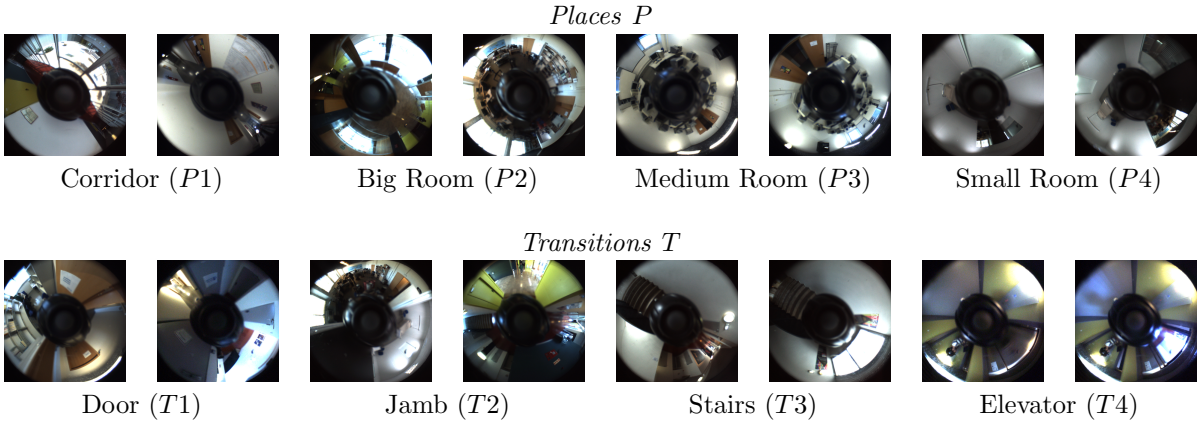


Figure 5.2: Examples of images labeled in the ground truth as elements of the different classes and subclasses.

we can see the simple model is weaker to represent the environment, while n-Cluster-Model seems a bit more robust. There are additional reasons to keep the second model: First, indoor environments use to include more areas of some classes than others, e.g., in the building of the test there are more doors than stairs or elevators, and secondly, 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. It includes images belonging to the following clusters: 5 corridor, 2 big rooms, 1 medium room, 2 small rooms, 10 Doors, 2 Jambs, 1 stairs and 1 elevator.

Table 5.2: 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				

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 subclasses ($P1/P2/P3/P4$ or $T1/T2/T3/T4$). Tables 5.2c and 5.2d show the results of this experiment. Looking to the diagonal of the confusion matrix 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 are usually classified as corridors, about 30% for each room subclass. 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 5.2d 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 gives acceptable results for our goals 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 test the whole mapping method proposed in Section 5.3. 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/subclass instead of the Naive Bayes Classifier evaluation. The HMM requires to adjust the probability of a transition to happen. We set high probability to the transition that implies repeating the same event occurred at previous step, time $t-1$, at time t: $p(S_t = i|S_{t-1} = j) = 0.9$ when $j = i$ and $p(S_t = i|S_{t-1} = j) = 1-0.9$ when $j \neq i$. Using the HMM probability evaluation to assign the labels, *Places* and *Transitions*, brings a slight improvement, as can be seen in Table 5.3 compared to previous results in Table 5.2.

Table 5.3: Labeling results evaluating the probability of each class according to the HMM including or not the min-size 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 5.3 summarizes the labeling results after running the proposed augmented mapping approach. It compares results including the minSize criteria explained in Section or not including it, in the subtables 5.3a and 5.3b respectively. To create a new cluster a class change is needed. We analyze here the influence of the minSize filter. This filter checks the Gist distance between the first and the last images on the current cluster and compares it with a similarity threshold, the distance must be greater than this threshold to create a new cluster.

The images are classified with HMM and as explained they are grouped in 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, as detailed later, the fact of avoiding too small clusters turns into a more meaningful semantic partition of the environment. Detailed results of classification into subclasses are only shown for the complete approach, including the min-size filter, in table 5.3c. Results without using this filter were very similar, slightly better for subclasses of Places but slightly worse for subclasses of Transitions. Table 5.4 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. 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 5.4: Number of clusters generated with the mapping approach with and without min-size filter.

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

Another interesting comparison we 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 things 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. 5.3 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. 5.3(a) shows the manual segmentation into clusters and their ground truth class label, and Fig. 5.3(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 5.5 shows the number of areas detected according to its class and subclass. We consider an area detected by our approach when 50% of the images in that area have been correctly classified. Usually the problem is that the generated clusters are still smaller than the ground truth annotated ones, that is why we consider correct detections only with a part of the hand labeled region found.

Table 5.5: Map areas detected

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

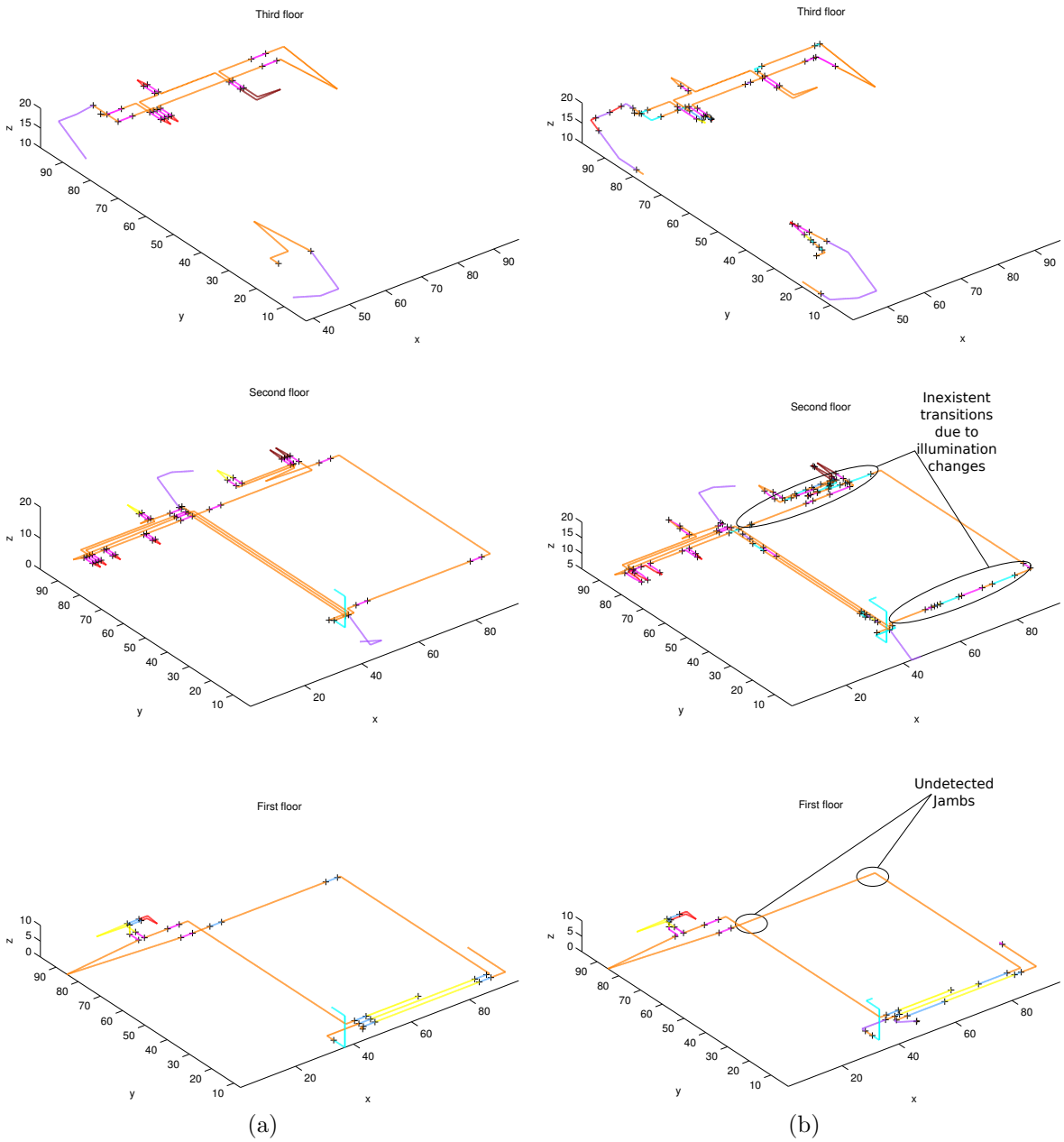


Figure 5.3: Segmentation of the trajectory in clusters of subclasses: (a) Manual, (b) Complete approach. The start position of each cluster is marked with a black cross. One color for each subclass: *Places*: (Orange) Corridor, (Yellow) Big Room, (Brown) Medium Room, (Red) Small Room
Transitions: (Pink) Door, (Blue) Jamb, (Purple) Stairs, (Light Blue) Elevator

Sección 6

Conclusions and Future Works

This works presents a novel indoor topological mapping method augmented with different labels of the basic indoor scenes. The mapping method uses catadioptric images and the adaptation of the Gist global descriptor to this kind of images. The general idea proposed is to simultaneously run the topological map building and a classifier to label the different types of indoor scenes considered. We have described a simple approach to label different types of *Places* and *Transitions*. The result of our mapping 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 for different use of the map. Our topological mapping method is based on this semantic classification of the images, using a previously built model of the environment, 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. A second group of experiments was run to evaluate 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 model of the environment when revisit to a certain place occurs. The step from our proposal that needs most improvements 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 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 project DPI2009-14664-C02-01.

Índice de figuras

Índice de tablas

Bibliografía

[1] Adrien Angeli, David Filliat, Stéphane Doncieux, and Jean-Arcady 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):1027–1037, 2008.

[2] João P. Barreto and Helder Araujo. Issues on the geometry of central catadioptric image formation. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 422–427, 2001.

[3] J.A. Castellanos, J. Neira, and J.D. Tardos. Multisensor fusion for simultaneous localization and map building. *Robotics and Automation, IEEE Transactions on*, 17(6):908–914, 2001.

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

[5] Andrew J. Davison, Ian D. Reid, Nicholas D. Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):1052–1067, 2007.

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

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

[8] Stephen Friedman, Hanna Pasula, and Dieter 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*, pages 2109–2114, 2007.

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

[10] Christopher Geyer and Konstantinos Daniilidis. A unifying theory for central panoramic systems and practical applications. In *Proc. of the European Conference on Computer Vision*, pages 445–461, June 2000.

[11] Toon Goedemé, Marnix Nuttin, Tinne Tuytelaars, and Luc Van Gool. Omnidirectional vision based topological navigation. *International Journal of Computer Vision*, 74(3):219–236, 2007.

[12] D. Gutiérrez, A. Rituerto, J. M. M. Montiel, and J. J. Guerrero. Adapting a real-time monocular visual slam from conventional to omnidirectional cameras. In *submitted to 11th OMNIVIS, held with International Conference on Computer Vision (ICCV)*, 2011.

[13] Daniel Gutiérrez Gómez. Slam omnidireccional en tiempo real para asistencia personal.

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

[15] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[16] O. Martínez Mozos and 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)*, pages 2772–2777, 2006.

[17] A. C. Murillo, P. Campos, J. Kosecka, and 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.

[18] A. C. Murillo and J. Kosecka. Experiments in place recognition using gist panoramas. In *9th OMNIVIS, held with ICCV*, pages 2196–2203, 2009.

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

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

[21] Matthias Nieuwenhuisen, Jörg Stückler, and Sven Behnke. Improving indoor navigation of autonomous robots by an explicit representation of doors. In *International Conference on Robotics and Automation*, pages 4895–4901, 2010.

[22] Andreas Nüchter and Joachim Hertzberg. Towards semantic maps for mobile robots. *Robotics and Autonomous Systems*, 56(11):915–926, 2008.

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

[24] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

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

[26] Andrzej Pronobis, Oscar M. Mozos, Barbara Caputo, and Patric Jensfelt. Multi-modal semantic place classification. *The International Journal of Robotics Research (IJRR), Special Issue on Robotic Vision*, 29(2–3):298–320, 2010.

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

[28] L. Puig, J. J. Guerrero, and 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.

[29] M. Pupilli and A. Calway. Real-time visual slam with resilience to erratic motion. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 1244–1249, June 2006.

[30] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 413–420, June 2009.

[31] Arnau Ramisa, Adriana Tapus, David Aldavert, Ricardo Toledo, and Ramon Lopez De Mantaras. Robust vision-based robot localization using combinations of local feature region detectors. *Autonomous Robots*, 27(4):373–385, 2009.

[32] A. Rituerto, L. Puig, and J. J. Guerrero. Comparison of omnidirectional and conventional monocular systems for visual slam (best paper award). In *10th OMNIVIS, held with Robotics Science and Systems (RSS)*, 2010.

[33] Alejandro Rituerto, Luis Puig, and J. J. Guerrero. Visual slam with an omnidirectional camera. In *20th International Conference on Pattern Recognition*, pages 348–351, 2010.

[34] Paul E. Rybski, Franziska Zacharias, Jean-Francois Lett, Osama Masoud, Maria Gini, and Nikolaos Papanikolopoulos. Using visual features to build topological maps of indoor environments. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 850–855, 2003.

[35] J. Santos-Victor, R. Vassallo, and H. Schneebeli. Topological maps for visual navigation. In *International Conffference on Computer Vision Systems*, pages 21–36, 1999.

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

[37] Alejandro Rituerto Sin. Localización y reconstrucción 3d con visión monocular omnidireccional.

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

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

[40] S. Thrun and A. Bucken. Integrating grid-based and topological maps for mobile robot navigation. In *Proc. of the National Conference on Artificial Intelligence*, pages 944–950, 1996.

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

[42] Elin Anna Topp and Henrik I. Christensen. Detecting region transitions for human-augmented mapping. *IEEE Transactions on Robotics*, 26(4):715–720, 2010.

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

[44] P. Viswanathan, T. Southey, J. Little, and A. Mackworth. Place classification using visual object categorization and global information. In *2011 Canadian Conference on Computer and Robot Vision*, pages 1–7, 2011.

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