Doctoral Dissertation

Advances in Indoor Semantic Mapping for Mobile Robotics

Jose Luis Matez Bandera
2024

Javier González Jiménez
Javier González Monroy

Tesis doctoral por compendio de publicaciones
Programa de Doctorado en Ingeniería Mecatrónica
Dpt. de Ingeniería de Sistemas y Automática
Universidad de Málaga

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR

D./Dña JOSE LUIS MATEZ BANDERA
Estudiante del programa de doctorado EN INGENIERÍA MECATRÓNICA de la Universidad de Málaga, autor/a de la tesis,presentada para la obtención del título de doctor por la Universidad de Málaga, titulada: ADVANCES IN INDOOR SEMANTIC MAPPING FOR MOBILE ROBOTICS

Realizada bajo la tutorización de JAVIER GONZÁLEZ JIMÉNEZ y dirección de JAVIER GONZÁLEZ JIMÉNEZ Y JAVIER GONZÁLEZ MONROY (si tuviera varios directores deberá hacer constar el nombre de todos)

DECLARO QUE:

La tesis presentada es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, conforme al ordenamiento jurídico vigente (Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), modificado por la Ley 2/2019, de 1 de marzo.

Igualmente asumo, ante a la Universidad de Málaga y ante cualquier otra instancia, la responsabilidad que pudiera derivarse en caso de plagio de contenidos en la tesis presentada, conforme al ordenamiento jurídico vigente.

En Málaga, a 25 de NOVIEMBRE de 2024

Fdo.: JOSE LUIS MATEZ BANDERA Doctorando/a	Fdo.: JAVIER GONZÁLEZ JIMÉNEZ Tutor/a
Fdo.: JAVIER GONZÁLEZ JIMÉNEZ Y JAVIER GONZÁLEZ MONROY Director/es de tesis	

Edificio Pabellón de Gobierno. Campus El Ejido. 29071
Tel.: 952 13 10 28 / 952 13 14 61 / 952 13 71 10
E-mail: doctorado@uma.es

UNIVERSIDAD DE MÁLAGA

DEPARTAMENTO DE
INGENIERÍA DE SISTEMAS Y AUTOMÁTICA

El Dr. D. Javier González Jiménez y el Dr. D. Javier González Monroy, directores de la tesis titulada "Advances in Indoor Semantic Mapping for Mobile Robotics" realizada por D. Jose Luis Matez Bandera, certifican su idoneidad para la obtención del título de Doctor en Ingeniería Mecatrónica.

Málaga, 25 de noviembre de 2024

_____

Dr. D. Javier González Jiménez

_____

Dr. D. Javier González Monroy

Dept. of System Engineering and Automation
University of Málaga
Studies in Mechatronics

Advances in Indoor Semantic Mapping for Mobile Robotics

AUTHOR: Jose Luis Matez Bandera

SUPERVISORS: Javier González Jiménez
Javier González Monroy

A mis cuatro pilares.

Contents

Abstract

For a mobile robot to become truly assistive to humans and perform high-level tasks in complex environments, its understanding of the workspace must extend beyond autonomous navigation and obstacle avoidance. While these capabilities allow a robot to move through an environment, true scene understanding requires perceiving and interpreting elements within the scene to enable meaningful interaction with the environment. Robots need to acquire the knowledge necessary to respond to human-like inquiries, such as: *what is this object for?*, *in which room am I?* or more complex questions like: *I have a headache, where can I find some medicine?*, answering to the latter, for example: *there is a first-aid kit in the bathroom cabinet that contains pain relievers.*

This level of understanding is typically achieved by robots through semantic maps, which are internal representations of the robot's workspace that link spatial information with their semantics. The latter includes the categorization of different scene elements, such as objects, places, and agents, along with the definition of their attributes such as physical characteristics, functionalities, and states. Furthermore, these maps provide not only information about individual entities but also the relationships between them, such as `the cold drinks are inside the fridge` (object-object) or `the teddy bear is in the kids' bedroom` (object-place), among others.

However, while these maps support reasoning over scene elements, their building still relies on fundamental perception tasks. This thesis aims to advance the automatic representation of multiple levels of knowledge within semantic maps, bridging the gap between low-level perception and high-level semantic concepts. Specifically, three fundamental challenges are addressed: i) place categorization, ii) 3D structural reconstruction, and iii)

object-oriented semantic mapping. For the categorization of places within a given environment, this thesis proposes a novel attention mechanism based on active perception that enhances the capabilities of existing place categorization algorithms by improving accuracy and adaptability. The proposed mechanism selectively focuses on relevant features, ensuring that the most informative perspectives are continuously prioritized, thereby refining the method's ability to classify places in complex environments. Addressing the second challenge, this thesis presents an online and incremental plane-based method for structural map construction. The proposal focuses on minimizing data handling and storage requirements, while effectively managing the uncertainties associated with online 3D reconstruction, fundamental aspects for its application on mobile robotics and resource-constrained devices. Lastly, concerning object-oriented semantic mapping, this thesis presents three significant contributions. First, it introduces a novel method aimed at enhancing video object detection in robotics, followed by two approaches for the comprehensive generation of instance-aware semantic maps of scene objects. The first approach emphasizes efficiency and long-term applicability, whereas the second prioritizes maximizing reconstruction accuracy while ensuring robustness.

Last but not least, it is important to emphasize that all previously described contributions share an important assumption in common: the camera that captures the information must be correctly localized in the environment for the appropriate link between semantics and spatial data. In this thesis, we also address this interesting and challenging problem by proposing Cross-Detector Visual Localization, a novel perspective that aims to enable interoperability among devices using different visual feature detectors. The goal is to allow coexisting devices to localize within a unified map built from features different from their own detection algorithms. In this topic, we present a proposal that leverages structural planar information and coplanarity constraints to guide feature matching as an initial solution.

Acknowledgments

Pursuing a PhD is a bit like a roller coaster, a journey of highs and lows, full of moments of both uncontrollable happiness but also of doubts and struggle. Yet, what truly makes a journey memorable is the support of those who walk alongside you. Here are a few words of thanks to all of you.

The first words are for the person who engaged me in this journey, my supervisor, Javier González Jiménez. Thanks for opening me the doors to the exciting field of research in robotics and computer vision, a place from which I never want to leave. Next to him, my co-supervisor, Javier González Monroy aka Javi Monroy. Javi, your extensive Monroy corrections in orange[1] were, at first glance, intimidating even for the bravest, but they soon became one of the key pieces for my growth as a researcher and for the success of this thesis. Together, you both provided me invaluable support and guidance with our endless meetings discussing interesting topics, not only regarding my PhD but also about some off-topic. And of course, I totally agree that someday we should include SLAM in our problem formulation!

Another aspect for which I am grateful to my supervisors is introducing me to MAPIR, which is much more than just a research group. A big thanks to all of you *mapireños* for promoting such a friendly atmosphere that made me hate remote work! Special mention to Raúl, my *semantic* guide, who made sure that I fell more in love, if possible, with research, continuously involving me in exciting projects. It is prohibited to continue without mentioning those colleagues with whom I have shared the laboratory (in person or remotely) in some or the whole time of this journey: Alberto (my Swedish co-adventurer), Mercedes, Andrés, Pepe (the template's chef), Mario, Silvia, Vladys, Goyo, David Fernández, Dominik, Enrique and the young talents Jesús and Jose Antonio. Each of you made my days with our inexcusable, and often long, breakfasts, the times acting as design committees, and the overall lovely atmosphere you generate, sharing laughs but also supporting when necessary

[1] I have seen this decorator countless times throughout this journey, and I believe it deserves to be captured in these pages.

(*e.g.* the inevitable paper rejections). As part of the MAPIR sports section, I would like to thank Paco for the creation of the official *Worst* Padel Tour (WPT) and to all the guys who joined us: Antonio, Borja, Yeray and Cipri. You all kept me moving and active, which is essential for a balanced life. To those of you who will join MAPIR at some point, congratulations, you have made the right decision!

In March 2023 and moving 3.737 km from the MAPIR lab, I landed at the offices of Ericsson Research in Stockholm (Sweden), where I spent four months as a research visitor with the Sensing and Perception team. There, I was under the supervision of José Araújo, whom I owe a special thanks for his invaluable support and guidance during my stay, allowing me to immerse myself in totally innovative topics and bringing fresh insights to my thesis. My gratitude also goes to all the researchers I met there, who made me feel at home from day one, ensuring I fully enjoyed my time, whether through inspiring work or enjoyable afterworks! And, of course, those days wouldn't have been complete without the daily after-lunch ping pong matches with the amazing master students: Davron, Ricardo, José Pedro, Javier, and Rubén. Missing a game was almost unforgivable! Finally, a special mention to Clara Gómez, the person who made all this possible and quickly became an invaluable friend. Clara, huge thanks for your support from the very beginning, and for making this a really unforgettable time!

Ahora es el turno de dirigirme a mis amigos. En primer lugar, quiero mencionar a esos chavales que allá por octubre de 2015, nos encontrábamos desperdigados en el salón de grados de la Escuela de Ingeniería Industriales, sin saber que en poco tiempo pasaríamos a ser grandes amigos. Si, hablo de Merino, Alex, Pablo, Sergio, Amador, Alan y Koke, a los que pronto se sumarían Ángel y Jose. A vosotros os debo un enorme gracias por todos esos momentos vividos, en los que vuestras continuas *tonterías* hacen que sea imposible no reírse y desconectar del resto de cosas. Ahora es el momento de celebrarlo con nuestro mítico *rey*. Otra persona que no puedo pasar por alto es mi *partner* Lidia, *mi amiga de toda la vida*. Y es que Lidia, como bien dice nuestra canción, la amistad no entiende el espacio-tiempo, y es que por mucha distancia que nos separe, no habrá nada que pueda con nuestra amistad. Gracias por ser como eres, por todas esas cervezas compartidas mientras hablábamos de nuestras cosas, y por todo lo que hemos disfrutado juntos. Gracias, de verdad. Por último, quiero dirigirme a Fran, Marisa, Berna y Cristi, personas que llegaron a mi vida durante la etapa del doctorado, y que tengo claro que han venido para quedarse. Y es que con vosotros las experiencias se vuelven únicas e inolvidables, desde simplemente estar de *tranquis* sin que falte una cachimba de Fran, hasta pasar un estupendo día dándole de comer a dromedarios y jirafas, o subirnos a un barco en el que nosotros debíamos ser nuestro propio capitán. Muchas gracias a los cuatro, sois increíbles.

Es el momento de hablar de tres de los cuatros pilares de mi vida. Mi familia: mi padre, mi madre y mi hermana. Voy a intentar transmitiros mi inmensa gratitud hacia vosotros, aunque os adelanto que difícilmente mis palabras lograrán describir lo enorme que es.

Papá, eres el ejemplo perfecto de alguien que lo da todo por su familia, sin esperar nada a cambio. Solo con escucharnos, sabes cuando necesitamos ayuda, y no dudas ni un segundo en pararlo todo para ayudarnos. Eres ejemplo de constancia, esfuerzo y sacrificio. Y es que jamás olvidaré tu reacción cuando te llamé para decirte que por fin me habían aceptado mi primera publicación, o esos innumerables momentos

de ir a *andar rápido* por tu Camino Nuevo, mientras me escuchabas quejarme sobre todo lo que me producía desmotivación. Gracias, papá.

Mamá, eres mi mayor ejemplo de lucha en la vida, y es que no hay cirscuntancia adversa que pueda contigo, y así nos lo has demostrado, por desgracia, en innumerables ocasiones. Eres ejemplo de madre incondicional, de madre que solo necesita verte para saber lo que necesitas. Es imposible olvidar cómo, cuando aún vivía en la casa con vosotros, siempre sacabas fuerzas para hacernos la vida más fácil. Mamá, te agradezco ese amor que a día de hoy sigue viajando en cada tupper de *tus croquetas* (esas que, por mucho que intente, jamás conseguiré replicar) o de *boquerones limpios*. Ten clara una cosa, y es que la vida acabará siendo justa contigo.

Y bueno, ahora le toca a mi Hermana, y sí, Hermana con mayúscula porque para mi ese siempre será su nombre, y no Laura. Como hermana mayor, me has demostrado que en la vida no vale rendirse, viendo como con perserverancia y dedicación, conseguías alcanzar ser lo que siempre habías soñado. Y es que no hay nada que hable mejor de ti que la dedicación desinteresada que tienes hacia las personas, con las que generas un cariño mutuo que habla por sí solo. Por último, darte las gracias por estar ahí siempre que lo he necesitado, y por tenerme en cuenta para todo. Siempre juntos, Hermana.

Papá, mamá y hermana, parte de esta tesis lleva vuestro nombre. Extiendo ahora este agradecimiento al resto de mi familia: a quienes ya no están, a quienes siguen conmigo y a quienes aún están por venir, porque todos han dejado, dejan y dejarán huella en mi vida. En especial, parte de esta tesis va dedicada a ti, Patri, que aunque ya no estés entre nosotros, jamás te olvidaremos.

Y ahora llego a mi cuarto pilar, mi compañera de vida, mi *chiquitina* Paula. ¿Quién mejor para aguantar a un doctorando? Pues sí, otra doctoranda. Paula, te has vuelto esencial para mí y no tengo palabras suficientes para agradecerte todo lo que has hecho. A pesar de lidiar con tu propio doctorado y sus dificultades, siempre has estado ahí, dándome el apoyo necesario y entendiéndome cuando ni yo mismo sabía. Pero más allá de eso, me has enseñado a crecer y a entender qué es lo que realmente importa en la vida. Así que solo deseo seguir sumando momentos a tu lado, ya sea haciendo la compra en el super, tumbados en el sofá, o viajando, como tanto nos gusta. Gracias, Paula, gran parte de esta tesis lleva tu nombre. Por último, quiero extender este agradecimiento a su familia, Pilar, Pepe y Lucía, quienes me han hecho sentir como uno más en todo momento.

Jose Luis Matez Bandera
Málaga, November 2024

This thesis has been supported by the grant program FPU19/00704 and the research projects WISER (DPI2017-84827-R), ARPEGGIO (PID2020-117057) and MINDMAPS (PID2023-148191NB-I00), all funded by the Spanish Government, and the research project VOXELAND (JA.B1-09), financed by the University of Malaga.

Resumen

Para que un robot móvil pueda convertirse en un verdadero ayudante de las personas y realizar tareas de alto nivel en entornos complejos, su interpretación del espacio de trabajo debe ir más allá de la navegación autónoma y la evitación de obstáculos. Aunque estas funciones permiten a un robot moverse por un entorno de forma segura, la verdadera interpretación de la escena implica percibir e interpretar los elementos que la componen para poder interaccionar de manera efectiva con el mismo. De esta manera, los robots deben adquirir los conocimientos necesarios para responder a preguntas propias de los humanos, como: *¿para qué sirve este objeto?*, *¿en qué habitación estoy?* o preguntas más complejas como: *me duele la cabeza, ¿dónde puedo encontrar algún medicamento?*, a las que se esperaría que respondiese, por ejemplo, con: *hay un botiquín en el armario del baño en el que hay analgésicos.*

Por lo general, los robots alcanzan este nivel de comprensión a través de los mapas semánticos, que son representaciones internas de su espacio de trabajo que vinculan la información espacial con su semántica. Esta última incluye información como la categorización de los distintos elementos del entorno, tales como objetos, lugares y agentes, junto con la descripción de sus atributos, como pueden ser características físicas, funcionalidades y estados. Además, estos mapas no sólo proporcionan información sobre los elementos individuales, sino también sobre las relaciones entre ellos, como `las bebidas frías están dentro de la nevera` (objeto-objeto) o `el osito de peluche está en el dormitorio de los niños` (objeto-lugar), entre otros.

No obstante, aunque estos mapas permiten el razonamiento sobre los distintos elementos del entorno, su creación sigue basándose fundamentalmente en tareas de percepción. Esta tesis tiene como objetivo

avanzar en la interpretación automatizada de distintos niveles de conocimiento dentro de los mapas semánticos, acortando así la distancia entre la percepción a bajo nivel y los conceptos semánticos a alto nivel. En concreto, se abordan tres retos principales: i) la categorización de lugares, ii) la reconstrucción 3D de la estructura y iii) la elaboración de mapas semánticos orientados a objetos. Para la categorización de lugares de un entorno específico, esta tesis propone un nuevo mecanismo de atención basado en percepción activa que mejora las capacidades de los algoritmos de categorización de lugares existentes, mejorando su precisión y adaptabilidad. El mecanismo propuesto se centra en seleccionar características relevantes del entorno, asegurando elegir de manera continua, los puntos de vista que aportan más información, mejorando así la capacidad de estos métodos para categorizar espacios en entornos complejos. En relación al segundo reto, esta tesis presenta un método que opera en tiempo de ejecución y de manera incremental, el cuál utiliza planos como primitiva geométrica para la construcción de mapas estructurales. La propuesta se centra en minimizar los requisitos de manejo y almacenamiento de datos, al tiempo que se gestiona eficazmente la incertidumbre asociada a la reconstrucción 3D, un aspecto fundamental para su aplicación en robótica móvil y dispositivos con recursos limitados. Por último, con respecto al mapeo semántico orientado a objetos, esta tesis presenta tres contribuciones significativas. En primer lugar, se introduce un método novedoso enfocado a mejorar la detección de objetos en vídeos capturados por robots móviles, así como dos métodos para la creación de mapas semánticos de los objetos de la escena orientados a instancias. De estos dos últimos, el primer enfoque hace especial énfasis en la eficiencia y la aplicabilidad a largo plazo, mientras que el segundo prioriza la maximización de la precisión de la reconstrucción al tiempo que garantiza la robustez de la misma.

Por último, pero no por ello menos importante, es fundamental destacar que todas las contribuciones descritas anteriormente comparten un importante requisito en común: la cámara con la que se captura la información debe ser correctamente localizada en el entorno, a fin de poder asociar la información semántica con la geometría del entorno. En esta tesis, también abordamos este interesante problema proponiendo la localización visual con detectores diferentes a los empleados en la construcción del mapa (*Cross-Detector Visual Localization*), un problema muy poco tratado en la literatura que busca la interoperabilidad entre dispositivos que utilizan diferentes detectores de características visuales. En este ámbito, presentamos una propuesta que aprovecha la información estructural planar e impone restricciones de coplanaridad para guiar el emparejamiento de características.

Prólogo

En el interior de un concurrido hospital, un equipo de robots trabaja en estrecha colaboración con cirujanos, enfermeros y personal auxiliar, integrándose a la perfección en el acelerado mundo de la asistencia sanitaria. Uno de estos robots, Sapin, se mueve por el quirófano, observando en silencio lo que ocurre mientras el equipo de cirujanos se prepara para una compleja operación de corazón. La cirujana jefa, la Dra. García-Fuentes, cruza una mirada de aprobación con su compañera, la Dra. Heersche, que asiente con la cabeza en señal de que Sapin está presente. A medida que el equipo se pone en marcha, Sapin se anticipa rápidamente a sus necesidades (véase la viñeta de la izquierda en la Figura 1): le acerca el bisturí al cirujano cuando éste lo necesita, ajusta la iluminación para asegurarse de que la zona del paciente esté perfectamente iluminada y se asegura de que los campos quirúrgicos estén meticulosamente colocados. Cuando el enfermero le pide las agujas de sutura, Sapin ya está en camino y regresa con el material antes de que él tenga tiempo de preocuparse. En ese momento, Sapin se convierte en algo más que un robot: es parte del equipo médico que, sin sustituir la labor de los expertos, les facilita su actividad.

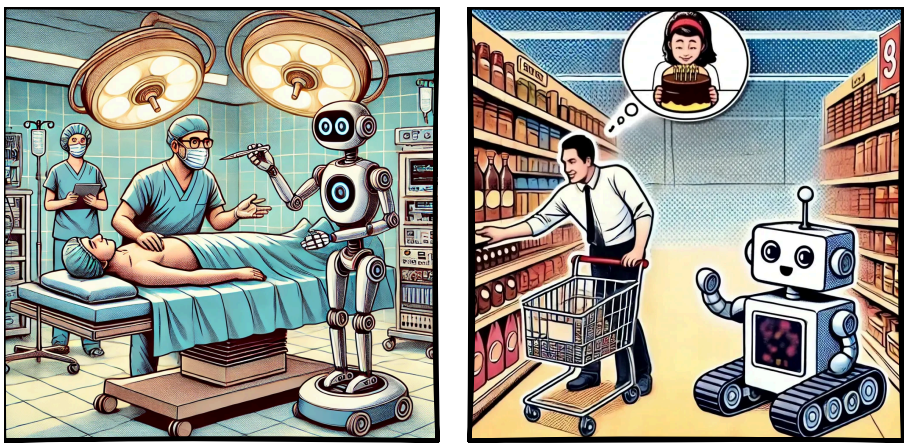


Figure 1: Ilustración de dos escenarios paralelos en una misma ciudad, en los que los robots facilitan la vida a los humanos. A la izquierda, un robot móvil colabora con el equipo médico en una operación, mientras que a la derecha, un robot asistente de compras guía a un padre en la elección de los ingredientes para la tarta de cumpleaños de su hija. Imágenes creadas con el modelo de IA generativa DALL-E.

Mientras tanto, al otro lado de la ciudad, un padre entra apresurado en su supermercado de confianza tras un largo día en la oficina. El cumpleaños de su hija Paula está a la vuelta de la esquina y tiene poco tiempo para prepararle una deliciosa tarta. Se siente desbordado, hasta que Malagueto, el

simpático robot asistente de la tienda, aparece a su lado. Malagueto analiza rápidamente la situación y empieza a guiarle por los pasillos, sugiriéndole ingredientes que no sólo encajan con una deliciosa receta de tarta, sino que también coinciden con los sabores favoritos de Paula (ver viñeta de la derecha en la Figura 1). Mientras recorren la tienda, Malagueto recuerda que a Paula le gusta especialmente un determinado chocolate, el cual siempre elige, así que le sugiere combinarlo con fresas frescas de temporada. ¡Será una cobertura deliciosa! -exclama Malagueto en tono alegre. Continúan su recorrido por la tienda, mientras Malagueto sugiere otras ideas, como añadir virutas de colores y una nota de *"¡Feliz cumpleaños, Paula!"* hecha con nata. El padre se siente ahora mucho más tranquilo y respira aliviado.

Estos dos robots, Sapin y Malagueto, ilustran cómo la robótica puede desempeñar un papel vital en la mejora de la calidad de vida de las personas. Aunque el rápido avance de la robótica y la inteligencia artificial últimamente está suscitando preocupación por el riesgo de destrucción de puestos de trabajo, es crucial aclarar que estas tecnologías no están aquí para sustituirnos, sino para facilitar nuestro día a día y enriquecer nuestras vidas. En lugar de percibir a los robots como nuestra competencia, deberíamos considerarlos como aliados que pueden aliviar nuestra carga de trabajo, permitiéndonos centrarnos en las tareas que realmente importan. Sin embargo, para llegar a este punto, independientemente de su área de aplicación (hospitales, fábricas, invernaderos, etc.), hay un requisito común que comparten todos los robots: la necesidad de una representación interna de su espacio de trabajo para lograr una interpretación exhaustiva del entorno, que permita a los robots asimilar conceptos semánticos a través de la percepción y realizar razonamientos complejos sobre su entorno.

Introducción y motivación

Los mapas semánticos representan información de alto nivel sobre los elementos individuales de un entorno, integrando no solo las categorías de objetos y tipos de lugares[2], sino también sus atributos, como características físicas, funcionalidades y estados. Además, estos mapas capturan las relaciones entre los distintos elementos, describiendo cómo interactúan o se conectan entre sí. Por ejemplo, una relación entre objetos podría expresarse como `los calcetines limpios suelen estar en la mesita de noche o en el armario`, mientras que una relación objeto-lugar podría formularse como `el desodorante suele encontrarse en el baño`.

[2]Nótese que, en esta tesis, el término *lugar* no se restringe a habitaciones completas, sino que también encapsula áreas con una función específica, como una cocina y una sala de estar dentro de la misma habitación de un estudio.

Aunque el mapeo semántico ha sido objeto de especial interés en los últimos años, como demuestra el elevado número de contribuciones en esta área de investigación [1–7], sigue existiendo una notable brecha entre la teoría y la aplicación práctica. A continuación, presentaremos los retos particulares que contribuyen a esta brecha y dificultan la construcción y el mantenimiento de mapas semánticos fiables en escenarios reales.

Uno de los principales problemas radica en las limitaciones de los algoritmos existentes para la interpretación del entorno, como la categorización de lugares y la detección de objetos, cuando se aplican a contextos robóticos. Estos algoritmos muestran un rendimiento notable en condiciones ideales (como por ejemplo, con buenas condiciones de iluminación, imágenes de alta calidad y entornos ordenados), pero rara vez reflejan la realidad a la que se enfrentan los robots móviles. Habitualmente, las cámaras montadas en los robots suelen ser de baja resolución y campo de visión limitado, a menudo debido al uso de sensores de bajo coste o capacidades de procesamiento limitadas. Como resultado, las imágenes capturadas durante la exploración del robot pueden ser borrosas a causa del movimiento, incluir partes ocluidas de objetos o mostrar escenas poco informativas, como paredes vacías, entre otras condiciones desfavorables. Estos problemas merman considerablemente el rendimiento de los algoritmos y pueden provocar errores que se propagan a lo largo del proceso de mapeo y dan lugar a un funcionamiento inadecuado e incluso inseguro del robot. Imaginemos que un robot que trabaja en un hospital clasifica erróneamente un quirófano como una habitación normal para pacientes. Más tarde, cuando el robot reciba la orden de comprobar como están los pacientes, podría interrumpir y distraer a los cirujanos en un momento crítico pensando que estaba visitando una habitación normal, y no un quirófano.

Por otro lado, la mayoría de los algoritmos de interpretación de escenas empleados están basados en aprendizaje (*e.g.* redes neuronales convolucionales), que generalmente se entrenan sobre un conjunto cerrado de categorías. Esto último se convierte en una limitación cuando estas redes se emplean en escenarios donde se encuentran con elementos cuyas categorías no fueron consideradas en el entrenamiento, lo que se traduce en una importante reducción del rendimiento. A modo ilustrativo, en detección de objetos, estas redes pueden producir clasificaciones con alta incertidumbre, como por ejemplo, clasificar un mismo objeto como `frigorífico` con un 47% de confianza y como `armario` con un 53% de confianza. También pueden producir clasificaciones inconsistentes pero excesivamente confiadas para detecciones consecutivas del mismo objeto físico, como por ejemplo, identificarlo como una `tarta` con un 72% de confianza en un fotograma, y en el siguiente, clasificarlo como una `mochila` con un 78% de confianza. Estos dos casos ilustran un escenario en el que no está clara la categoría correcta del objeto, es decir, existe una gran incertidumbre en su clasificación. Por desgracia, esta incertidumbre suele ser ignorada por los métodos de mapeo

Figure 2: Ejemplo ilustrativo del rendimiento reducido de la red de detección de objetos Detectron2 [8] en una imagen que contiene objetos cuyas categorías no fueron consideradas en el entrenamiento. Los resultados demuestran que la red detecta correctamente objetos de estas categorías, como el frigorífico (84%) y el microondas (96%). Sin embargo, produce clasificaciones excesivamente seguras pero incorrectas cuando encuentra objetos que no estaban en el conjunto de entrenamiento, como robots móviles, etiquetándolos como un pastel (72%), un tren (82%) y un teléfono móvil (52%).

semántico, que a menudo se limitan en asignar la clase más votada sin tener en cuenta la probabilidad de las demás clases. Sin embargo, la clase más votada no siempre es la correcta, ya que podría tratarse de un objeto fuera de distribución o, simplemente, no haber sido clasificado correctamente por haber sido observado desde puntos de vista engañosos. Ignorar esta incertidumbre puede llevar a un funcionamiento erróneo del robot, como por ejemplo se ilustra en la Figura 2, en el que un objeto cuya categoría no fue considerada durante el entrenamiento se clasifica erróneamente como `tarta`. Ahora, continuando con la historia descrita en el prólogo de este trabajo, imaginemos que estamos en el cumpleaños de Paula, y su padre le ordena al robot *"¡oye Malagueto, por favor, tráenos la tarta de cumpleaños!"*. El resultado probablemente dejaría a Paula y a los invitados decepcionados, y con razón, a causa de la inesperada "tarta". Esta falta de consideración de la incertidumbre se extiende a otras partes del proceso de mapeo, siendo la localización del robot una de las más importantes. Aunque los enfoques de mapeo semántico a menudo asumen que conocen la localización del robot sin error, esto no es una suposición realista en escenarios reales. En

consecuencia, los enfoques que no estén preparados para manejar estas incertidumbres inevitablemente darán lugar a un funcionamiento erróneo.

Estos son algunos de los principales retos a los que se ha de hacer frente a la hora de desplegar algoritmos de mapeo semántico en la práctica. Sin embargo, hay otras limitaciones que también requieren atención, como la necesidad de datos de entrada muy preprocesados, que a menudo son difíciles de obtener, o el manejo y almacenamiento de grandes volúmenes de datos generados durante el mapeo, lo cual no se puede garantizar, especialmente cuando se trabaja con robots móviles con recursos limitados.

Contribuciones

Esta tesis contribuye a mejorar la generación de mapas semánticos con robots móviles, centrándose particularmente en la búsqueda de soluciones a los principales obstáculos a los que se enfrentan los algoritmos tradicionales de mapas semánticos cuando se despliegan en entornos reales. Sin embargo, es importante destacar que las contribuciones de esta tesis se centran en los problemas relacionados con la percepción y la representación del conocimiento y no abordan la explotación de este conocimiento para el razonamiento semántico de alto nivel. Además, se explora cómo la información almacenada en estos mapas semánticos puede aprovecharse para resolver el problema de la localización visual de detectores cruzados, un nuevo reto introducido y abordado por primera vez en la literatura a través de esta tesis. Al considerar tanto el mapeo semántico como la localización, la tesis contribuye a incrementar la robustez y aplicabilidad de los sistemas robóticos móviles en entornos dinámicos.

Las contribuciones de esta tesis pueden dividirse en cuatro bloques principales. En las siguientes secciones se presentan los problemas más importantes asociados a estos cuatro bloques y se presentan las contribuciones específicas de esta tesis en cada uno de ellos.

Categorización de lugares

La mayoría de los algoritmos de categorización de lugares intentan determinar la categoría de un lugar a partir de una sola imagen o de una secuencia de imágenes [1, 9–12]. Cuando se aplican en el campo de la robótica, donde las imágenes suelen incluir paredes o regiones vacías, el rendimiento de estos algoritmos se degrada considerablemente. En este contexto, un problema común a estos enfoques es determinar el punto de vista óptimo desde donde capturar las imágenes. Esto es relevante para maximizar la información contenida en estas imágenes sobre la categoría del

lugar (por ejemplo, priorizando observar objetos que típicamente sólo se encuentran en lugares específicos como los inodoros en los baños), con el fin de garantizar la correcta categorización.

En nuestro trabajo [13], introducimos un mecanismo de atención basado en visión activa que pretende mejorar la eficiencia y precisión de los algoritmos existentes para categorización de lugares. Este mecanismo selecciona dinámicamente la línea de visión de la cámara utilizando una unidad de movimiento panorámico para maximizar, en cada momento, la ganancia de información. En concreto, se formaliza como un problema de siguiente mejor punto de vista (*next-best-view*) dentro de un modelo de Proceso de Decisión de Markov (*Markov Decision Process*, MDP). Nuestro enfoque se evalúa con algoritmos de los dos principales paradigmas de categorización de lugares (métodos basados en objetos y en imágenes) y demuestra su eficacia mejorando las configuraciones de cámara habituales en robótica.

Reconstrucción estructural 3D

La generación de mapas estructurales, también conocida como reconstrucción 3D de la planta, implica la recopilación de datos obtenidos a través de los sensores de a bordo, como cámaras RGB-D y sensores de alcance, mientras el robot explora el entorno. A pesar de los alentadores resultados obtenidos en los últimos años, la mayoría de los enfoques existentes se enfrentan a una serie de limitaciones que dificultan su aplicación en escenarios del mundo real. En primer lugar, la necesidad de disponer de una nube de puntos densa de todo el entorno como entrada, lo que requiere almacenar y manipular grandes volúmenes de datos. En segundo lugar, estos métodos suelen ignorar la incertidumbre inherente a la localización de los robots, que puede dar lugar a nubes de puntos incorrectamente alineadas durante su registro. Este problema se suele ignorar asumiendo que las nubes de puntos ya están alineadas. En tercer lugar, algunos enfoques se basan en el supuesto poco realista de un mundo Manhattan, en el que todos los elementos estructurales son ortogonales.

En [14] proponemos *Sigma-FP*, un método para la reconstrucción de plantas 3D a partir de secuencias RGB-D. La principal ventaja de *Sigma-FP* reside en su capacidad de operar en tiempo de ejecución, que elimina la necesidad de almacenar grandes volúmenes de datos, procesando la información sobre la marcha y reteniendo únicamente una representación compacta de los elementos estructurales (*i.e.* planos). *Sigma-FP* emplea un enfoque probabilístico para tener en cuenta las incertidumbres inherentes tanto a la localización del robot como a la estimación de planos, mejorando así la precisión de la reconstrucción. Además, *Sigma-FP* relaja la hipótesis de configuración geométrica al mundo de Atlanta, en el que la única restricción es que los elementos estructurales deben ser verticales. Por último, *Sigma-FP*

no sólo reconstruye las paredes, sino también aperturas como puertas y ventanas, proporcionando una reconstrucción más precisa y detallada. Nuestro enfoque se evalúa en diversos entornos, superando a los métodos más recientes, que tienen dificultades cuando se enfrentan a condiciones del mundo real como la incertidumbre en la localización.

Mapeo semántico orientado a objetos

La detección de objetos es una tarea clave para poblar los mapas semánticos con los objetos que hay en el espacio de trabajo de un robot. Aunque los recientes avances en aprendizaje profundo han mejorado significativamente el rendimiento de la detección de objetos, estos métodos a menudo presentan un rendimiento inferior en aplicaciones robóticas. Ello se debe, en gran medida, a las peculiaridades de las imágenes captadas por los sensores montados en los robots, que por lo general son de baja resolución y tienen un campo de visión reducido. Además, el propio movimiento del robot puede causar imágenes borrosas, lo que hace que los resultados de la detección varíen considerablemente entre fotogramas consecutivos, provocando una falta de consistencia. Otro problema surge por las oclusiones, vistas parciales y cambios en la iluminación, habituales durante la adquisión de las imágenes. Estos factores dificultan la construcción de mapas semánticos fiables, lo que a menudo da lugar a múltiples instancias del mismo objeto físico o a objetos mal clasificados.

El trabajo presentado en [15] propone un método para mejorar la detección de objetos en secuencias de imágenes capturadas por robots. Aprovechando el conocimiento del movimiento de la cámara entre fotogramas consecutivos, el método propaga las detecciones entre fotogramas y las integra utilizando un filtro Bayesiano recursivo. Este método garantiza la coherencia espacio-temporal de las predicciones de la red para aplicaciones robóticas. Los resultados demuestran su eficacia, con un aumento mínimo del tiempo de procesamiento.

En [16], presentamos *LTC-Mapping*, un método para construir mapas semánticos orientados a objetos que emplea *bounding boxes* 3D como primitiva geométrica, con el objetivo principal de garantizar que estos mapas sean consistentes a largo plazo. Este método aborda dos retos cruciales: la duplicidad de instancias y los entornos dinámicos. Para mitigar la duplicidad de instancias, *LTC-Mapping* etiqueta los vértices de cada *bounding box* con indicadores de visibilidad, indicando si han sido observados o no. Esta información es crucial para determinar si un objeto sólo se ha observado parcialmente debido a observaciones incompletas u oclusiones, lo que nos permite identificar varias instancias procedentes de observaciones parciales del mismo objeto físico e integrarlas en una única instancia completa. Además, para gestionar entornos dinámicos, presentamos el concepto de no detección, que facilita la eliminación de objetos del mapa cuando ya no se

observan más en el entorno, garantizando que el mapa se mantiene actualizado al tener en cuenta tanto las nuevas observaciones como los objetos que ya no se vuelven a detectar. Se ha demostrado que estas técnicas mejoran la fiabilidad y robustez de los mapas semánticos a lo largo del tiempo.

Por último, en [17], presentamos *Voxeland*, una alternativa para el mapeo semántico orientado a objetos que utiliza voxeles como primitiva geométrica. Este método aborda problemas habituales en las predicciones de redes neuronales, como las predicciones incorrectas con objetos fuera de la distribución de entrenamiento y la generación de máscaras imprecisas. Para solucionar estos problemas, proponemos un marco probabilístico inspirado en la Teoría de la Evidencia, en el que cada predicción de la red se trata como una opinión subjetiva tanto a nivel geométrico como semántico. Estas opiniones se integran a lo largo del tiempo en evidencias mediante un modelo probabilístico que permite cuantificar la incertidumbre en ambos niveles. Esto último nos permite identificar los objetos que requieren una reobservación o reclasificación. En este trabajo, proponemos una estrategia para aprovechar esta información implementando la desambiguación semántica mediante la integración de un *Large Vision-Language Model* (LVLM). Este modelo genera una opinión más completa para objetos que presentan una elevada incertidumbre semántica. La evaluación pone de manifiesto la importancia de incorporar y explotar la incertidumbre para mejorar la robustez y fiabilidad de los mapas resultantes.

Localización visual con detectores heterogéneos para garantizar la interoperabilidad a largo plazo

Un requisito imprescindible para construir mapas semánticos es poder conocer, en cada momento, la localización del robot, lo cual es esencial para ubicar los distintos elementos de la escena con respecto a un sistema de referencia. Además, lo ideal sería que el método de localización fuera compatible con múltiples robots operando en el mismo entorno, permitiendo la interoperabilidad entre ellos, como colaborar en la construcción y mantenimiento del mapa semántico. Asimismo, este enfoque de localización también debería ser aplicable a largo plazo, con el fin de mantener el mapa actualizado con los cambios que se produzcan en el mundo real (*e.g.* reubicación de un objeto, o la eliminación de un elemento que ya no se encuentra en el entorno). Una solución a este problema es la Localización Visual (*Visual Localization*, VL), que utiliza sensores visuales a bordo tales como cámaras, para extraer características de una imagen de referencia para, posteriormente, establecer correspondencias con características disponibles en un mapa 3D existente. Sin embargo, este enfoque tiene una limitación importante: funciona bajo el supuesto de que el algoritmo de detección de

características utilizado tanto para la imagen de referencia como para el mapa es el mismo.

En nuestro trabajo [18], ampliamos el problema de la localización visual para que sea aplicable a detectores heterogéneos, lo que denominamos *Cross-Detector Visual Localization*. El principal reto de este problema radica en la discrepancia espacial inherente a los puntos representativos de referencia heterogéneos que representan entidades físicas diferentes pero cercanas, lo que dificulta el proceso de establecer correspondencias correctas. Para abordar este problema, aprovechamos la información sobre los elementos estructurales disponibles en el mapa semántico para imponer restricciones de coplanaridad durante el proceso de correspondencia. Los resultados demuestran que, a pesar de una pequeña sobrecarga computacional, la consideración de estas restricciones de coplanaridad guían eficazmente el proceso de emparejamiento, permitiendo establecer correspondencias correctas entre puntos representativos heterogéneos.

Publicaciones

Esta tesis se presenta bajo la modalidad de *tesis avalada por un compendio de trabajos relevantes publicados*. Concretamente, la presente tesis engloba las siguientes publicaciones:

Revistas

- *Jose-Luis Matez-Bandera, Pepe Ojeda, Javier Monroy, Javier Gonzalez-Jimenez and Jose-Raul Ruiz-Sarmiento.* **Voxeland: Probabilistic Instance-Aware Semantic Mapping with Evidence-based Uncertainty Quantification.** Enviado y en revisión, (2024).

- *Jose-Luis Matez-Bandera, Alberto Jaenal, Clara Gomez, Alejandra C. Hernandez, Javier Monroy, José Araújo and Javier Gonzalez-Jimenez.* **Cross-Detector Visual Localization with Coplanarity Constraints for Indoor Environments.** Enviado y en revisión, (2024).

- *Jose-Luis Matez-Bandera, Javier Monroy and Javier Gonzalez-Jimenez.* **Sigma-FP: Robot Mapping of 3D Floor Plans with an RGB-D Camera Under Uncertainty.** En IEEE Robotics and Automation Letters, vol. 7, no. 4, pp. 12539-12546, (2022). DOI: 10.1109/LRA.2022.3220156

- *Jose-Luis Matez-Bandera, David Fernandez-Chaves, Jose-Raul Ruiz-Sarmiento, Javier Monroy, Nicolai Petkov and Javier Gonzalez-Jimenez.* **LTC-Mapping, Enhancing Long-Term Consistency of Object-Oriented Semantic Maps in Robotics.** En MDPI Sensors, vol. 22, no. 14, (2022).
  DOI: 10.3390/s22145308

- *Jose-Luis Matez-Bandera, Javier Monroy and Javier Gonzalez-Jimenez.* **Efficient Semantic Place Categorization by a Robot through Active Line-of-Sight Selection** En Knowledge-Based Systems, vol. 240, pp. 108022-108034, (2022).
  DOI: 10.1016/j.knosys.2021.108022

Conferencias

- *David Fernandez-Chaves, Jose-Luis Matez-Bandera, Jose-Raul Ruiz-Sarmiento, Javier Monroy, Nicolai Petkov and Javier Gonzalez-Jimenez.* **Exploiting Spatio-Temporal Coherence for Video Object Detection in Robotics.** En International Conference on Computer Analysis of Images and Patterns, (2021).
  DOI: 10.1007/978-3-030-89131-2_17

Marco de la tesis

Esta tesis es el resultado de cuatro años de trabajo como investigador en el grupo de investigación *Machine Perception and Intelligent Robotics* (MAPIR[3]), perteneciente al Departamento de Ingeniería de Sistemas y Automática de la Universidad de Málaga. El autor ha disfrutado de una beca FPU (Formación de Profesorado Universitario) (FPU19/00704) del Ministerio de Ciencia, Innovación y Universidades, que ha sido la principal fuente de financiación de esta investigación. Esta investigación también ha sido financiada por los proyectos de investigación del grupo, en particular, los tres proyectos nacionales WISER (DPI2017-84827-R), ARPEGGIO (PID2020-117057GB-I00) y MINDMAPS (PID2023-148191NB-I00), y el proyecto regional VOXELAND (JA.B1-09).

Durante este periodo, el autor completó con éxito el programa de doctorado en Ingeniería Mecatrónica, coordinado por el Departamento de Ingeniería de Sistemas y Automática de la Universidad de Málaga. Dentro de este programa multidisciplinar, el autor adquirió una amplia experiencia en los diferentes ámbitos que conforman la base de la mecatrónica, incluyendo la

[3] mapir.isa.uma.es

mecánica, la electrónica, la programación y la automatización. Además, el autor complementó su formación académica con cursos técnicos de gran relevancia para su investigación como el *Artificial Intelligence* de Samsung Innovation Campus (septiembre-diciembre 2020), *Getting Started with AI on Jetson Nano* (febrero 2021) y *Fundamentals of Deep Learning for Multi-GPUs* (junio 2021), siendo estos dos últimos impartidos por NVIDIA.

A lo largo de esta tesis, el autor participó activamente con la comunidad científica de las dos áreas principales de su tesis, robótica y visión por computador, asistiendo y presentando sus trabajo en múltiples congresos nacionales e internacionales como *International Conference on Computer Analysis of Images and Patterns* (Chipre, en remoto, 2021), *International Conference on Robotics and Automation* (Londres, 2023), y las *Jornadas de Automática* (Málaga, 2024). Además, el autor participó como voluntario en la organización del *European Robotics Forum* en 2020. También cabe destacar su participación activa como revisor –algo sumamente necesario en el ámbito de la investigación científica– de prestigiosas conferencias y revistas, entre las que destacan IEEE International Conference on Robotics and Automation (ICRA), IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), y IEEE Transactions on Robotics (T-RO), entre otras.

El autor también tuvo la oportunidad de realizar una estancia de investigación de cuatro meses en Ericsson AB, en las oficinas situadas en Estocolmo (Suecia), dentro del equipo de Sensing and Perception, bajo la supervisión del Dr. José Araújo. Durante estas prácticas, el autor adquirió conocimientos sobre temas tangenciales al tema principal de su tesis, centrándose principalmente en la localización visual con detectores heterogéneos. Esta experiencia le permitió incorporar nuevos conocimientos a su investigación al tiempo que contribuía al progreso del equipo. No menos importante, esta estancia le brindó al autor la oportunidad de establecer estrechas colaboraciones con otros investigadores del equipo.

La beca FPU también ofreció la oportunidad de ser colaborador docente en el Departamento de Ingeniería de Sistemas y Automática de la Universidad de Málaga. En concreto, el autor impartió docencia en la Escuela Técnica Superior de Ingeniería Informática en las asignaturas de *Modelado de Sistemas Biomédicos* (2021/2022, 2022/2023 y 2023/2024), *Programación de Robots* (2021/2022) y *Control Automático* (2023/2024).

La colaboración del autor con miembros del grupo de investigación MAPIR e investigadores internacionales ha dado lugar a una serie de publicaciones adicionales:

Conferencias

- *Jesús Moncada Ramírez, Jose-Raul Ruiz-Sarmiento, Jose-Luis Matez-Bandera, Javier Gonzalez-Jimenez.* **Modelos a gran escala para mapeo semántico en robótica móvil.** En XLV Jornadas de

Automática, (2024).
DOI: 10.17979/ja-cea.2024.45.10940

- *Gregorio Ambrosio-Cestero, Jose-Luis Matez-Bandera, Jose-Raul Ruiz-Sarmiento, Javier Gonzalez-Jimenez.* **Entorno basado en contenedores Linux para el desarrollo de aplicaciones robóticas.** En XLV Jornadas de Automática, (2024).
DOI: 10.17979/ja-cea.2024.45.10943

Workshops

- *Matteo Luperto, Jose-Luis Matez-Bandera, Tomasz Piotr Kucner, Michele Antonazzi, Gabriele Somaschini, Javier Monroy, Javier Gonzalez-Jimenez, Nicola Basilico.* **Ensuring the Consistency of Heterogeneous World Representations Using Structural Features.** En ICRA 2023 Workshop on Unconventional spatial representations. Opportunities for robotics, (2023).

Patentes

- *Jose-Luis Matez-Bandera, Alberto Jaenal, Clara Gomez, Alejandra C. Hernandez, and José Araújo.* **Determining Location of a Device within an Environment Comprising Planar Surfaces.** En revisión.

Estructura de la tesis

Los siguientes capítulos de la presente tesis se organizan de la siguiente manera:

Capítulo 1: Introducción ofrece un resumen de la motivación principal de esta tesis doctoral, junto con una visión general de su desarrollo y las contribuciones a las que ha dado lugar.

Capítulo 2: Fundamentos teóricos introduce los conceptos básicos que sustentan la presente tesis, centrándose en los tres pilares principales: matemáticas, visión por computador e interpretación del entorno. Cada sección abarca los aspectos elementales de estas áreas para así facilitar su entendimiento, indicando para cada una de ellas, cuál ha sido su relevancia y sus aplicaciones en el contexto de este trabajo.

Capítulo 3: Categorización de lugares presenta un mecanismo de atención para robots móviles, cuyo objetivo es mejorar la eficiencia de

los métodos existentes para categorización de lugares. Para ello, este mecanismo selecciona de forma continua el punto de vista óptimo desde el cual capturar imágenes, maximizando así la ganancia de información.

Capítulo 4: Reconstrucción 3D de elementos estructurales introduce un método para la reconstrucción automática de planos de planta tridimensionales de entornos con múltiples habitaciones, utilizando una secuencia de imágenes RGB-D. Además, el método propuesto considera la naturaleza probabilística tanto de la localización del robot como de la etapa de estimación de planos.

Capítulo 5: Mapeo semántico orientado a objetos inicialmente propone una técnica que, aprovechando el conocimiento del movimiento de la cámara entre fotogramas consecutivos, otorga coherencia espacio-temporal a la detección de objetos en vídeos capturados por robots. A continuación, este capítulo describe dos métodos para el mapeo semántico orientado a objetos. En concreto, el primer método utiliza *bounding boxes* 3D como primitiva geométrica para representar a los objetos, buscando así una mayor eficiencia. El segundo método se centra en mejorar la precisión en la reconstrucción, empleando para ello voxeles como primitiva. Ambos métodos son aptos para operar en tiempo de ejecución, y consideran a los objetos como entidades individuales, lo que es de vital relevancia para interactuar con los elementos del entorno.

Capítulo 6: Localización visual con detectores cruzados para facilitar una interoperabilidad a largo plazo introduce y formaliza el problema de localización visual con detectores cruzados, un problema complejo que busca asegurar la interoperabilidad de cámaras que utilizan diferentes algoritmos de detección de características, utilizando para localizarse un único mapa común. Posteriormente, este capítulo presenta un método que aprovecha información sobre planos estructurales para afrontar el problema.

Capítulo 7: Conclusiones y líneas de trabajo futuras concluye esta tesis doctoral, resumiendo los resultados principales de este trabajo, así como proponiendo posibles líneas futuras de investigación.

Es importante señalar que no hay una sección general de *trabajos relacionados*. En su lugar, cada sección de los capítulos siguientes ofrece una revisión de la literatura concreta de cada uno de los problemas abordados.

Conclusiones

Inicialmente, esta tesis tenía como objetivo explorar el uso de técnicas de visión activa para avanzar en el mapeo semántico de interiores con robots móviles. Sin embargo, durante los primeros pasos en esta línea nos encontramos con una serie de limitaciones y/o requisitos que debían ser abordados antes de seguir avanzando en dicha línea. Principalmente, estos desafíos estaban relacionados con el despliegue de métodos de mapeo semántico en escenarios reales, donde a menudo la información que tiene el robot es limitada, y en muchos casos, está sujeta a una alta incertidumbre. En consecuencia, el enfoque principal de la presente tesis evolucionó al estudio de estos problemas, como un paso previo necesario antes de poder aplicar técnicas de visión activa para mejorar la eficiencia de los métodos de mapeo semántico. Por último, como resultado de la estancia realizada en Ericsson Research, surgió un trabajo complementario centrado en la localización visual, fruto de la unión entre el área de investigación de dicha compañía y los objetivos de esta tesis. Este resultado no solo enriquece la temática general, sino que también destaca la relevancia transversal de los avances logrados, fundamentales tanto para el mapeo semántico como para otras aplicaciones robóticas. Es por ello que, las contribuciones de esta tesis tienen cierta diversidad, aunque principalmente enfocadas al avance del mapeo semántico. A continuación, se resumen las principales conclusiones extraídas en los diferentes problemas abordados.

El primer bloque abordado fue la categorización de lugares, cuyo objetivo es proporcionar un contexto esencial al robot que le permite interpretar la funcionalidad de cada zona del entorno, como por ejemplo reconocer que una cocina es el lugar apropiado para calentar comida. Aunque los algoritmos de categorización de lugares existentes, tanto los basados en objetos como los basados en imágenes, funcionan bien en condiciones ideales (en las que todos los objetos se detectan con claridad y las imágenes son muy informativas), su aplicación a escenarios robóticos suele presentar dificultades, como que las imágenes captadas por los robots son propensas a contener oclusiones, vistas parciales de objetos o escenas vacías. Para hacer frente a estas limitaciones, en el Capítulo 3 hemos propuesto un mecanismo de atención para maximizar la información disponible en las imágenes tomadas por los robots mediante la selección, en cada momento, de la línea de visión de la cámara más informativa a través de una unidad de movimiento panorámico. Además, también hemos tenido en cuenta la planificación de la trayectoria de navegación del robot a corto plazo para maximizar aún más la ganancia de información a lo largo de la trayectoria prevista del robot. Los resultados obtenidos han demostrado la importancia de seleccionar el punto de vista adecuado para maximizar la precisión de los métodos de categorización de lugares, a la vez que se reduce el tiempo necesario para una categorización correcta.

La siguiente problemática abordada en esta tesis, fruto de uno de los principales requisitos de los algoritmos de categorización de lugares (como es conocer la delimitación de los distintos lugares del entorno), es la reconstrucción 3D de los elementos estructurales. En el Capítulo 4, hemos presentado un método para abordar tres limitaciones claves encontradas en escenarios del mundo real: i) minimizar la necesidad de almacenar y manejar gran cantidad de datos, ii) tener en cuenta las incertidumbres inherentes al proceso de reconstrucción, y iii) adoptar hipótesis realistas en lugar de simplificar en exceso el problema. Nuestra propuesta, *Sigma-FP*, opera fotograma a fotograma, extrayendo una representación compacta de la estructura de la escena en forma de planos. Además, *Sigma-FP* se ha formulado en términos probabilísticos, lo que ha permitido incorporar incertidumbres procedentes tanto de la localización del robot como de la etapa de estimación de planos, a la vez que se propagan estas incertidumbres al modelo. Esto último es crucial para ponderar las distintas observaciones y determinar qué partes del modelo son más fiables y cuáles requieren ser re-observadas. Por último, hemos relajado la hipótesis típica del mundo de Manhattan al mundo Atlanta, la cuál es más realista, al tiempo que hemos incorporado a nuestro modelo detalles como el grosor de las paredes y las aperturas (es decir, puertas y ventanas).

El tercer bloque se centra en el mapeo semántico orientado a objetos, una de las áreas más relevantes –si no la más importante– dentro del ámbito del mapeo semántico. Esto se aborda en el Capítulo 5, en el que primero propusimos un método para dotar de consistencia espacio-temporal a las predicciones de las redes neuronales para detección de objetos de dos etapas, aplicadas a vídeos capturados por robots. Después, con el objetivo de obtener un mapeo completo de los objetos de la escena, propusimos dos enfoques: *LTC-Mapping* y *Voxeland*. El primero se centró en resolver las principales limitaciones que impiden la usabilidad de los mapas semánticos a largo plazo: i) la falta de mantenimiento del mapa, que no captura los cambios dinámicos que se producen en el mundo real, y ii) el problema de duplicidad de instancias, en el que se mapean múltiples instancias del mismo objeto físico. Además, *LTC-Mapping* se diseñó para priorizar la eficiencia, eligiendo para ello una primitiva liviana como son los *bounding boxes* 3D. En cambio, *Voxeland* se orientó más hacia mejorar la precisión de la reconstrucción, a la vez que pretendía tener en cuenta las limitaciones habituales de las redes de detección de objetos, que a menudo producen detecciones excesivamente seguras pero incorrectas que, al propagarse, comprometen la fiabilidad del mapa. Los enfoques propuestos han demostrado avances significativos para permitir un despliegue satisfactorio en el mundo real.

Por último, y como resultado de la estancia en Ericsson Research tal y como se ha mencionado anteriormente, esta tesis también aborda el problema de localización, el cuál es de gran relevancia para vincular correctamente la información semántica adquirida con su lugar correspondiente en el mapa

geométrico. Dicha localización idealmente debe ser relativa a un sistema de referencia común para todos los robots, permitiendo así su colaboración. Asimismo, debe ser aplicable a largo plazo, garantizando que el mapa pueda mantenerse actualizado. En el Capítulo 6 de la presente tesis se ha introducido por primera vez en la literatura el problema de *Cross-Detector Visual Localization*, cuyo objetivo es permitir la interoperabilidad entre robots equipados con sensores visuales. En particular, el propósito era garantizar que las cámaras que utilizan diferentes detectores de características puedan ser localizadas frente a un único mapa común, el cual contiene características 3D de diferente índole. Para abordar este problema, como primera alternativa se ha propuesto *CoplaMatch*, un enfoque que aprovecha la información de los planos estructurales para guiar el emparejamiento de puntos clave heterogéneos. Los resultados han mostrado que, en este escenario, no es recomendable depender en gran medida de los descriptores de características para extraer correspondencias, ya que carecen de la unicidad que se suele explotar en la localización visual tradicional. Por lo tanto, se hace necesario explorar alternativas para complementar y guiar el proceso de establecimiento de correspondencias.

Líneas de trabajo futuras

Aunque esta tesis representa un gran avance en nuestra área –y en mi vida personal–, esto no acaba aquí. Aún queda mucho trabajo interesante por delante mientras sigamos explorando nuevas tecnologías, las cuales aparecen cada vez con mayor frecuencia. Aquí presentamos algunas de las líneas de investigación abiertas que nos parecen particularmente interesantes para continuar en un futuro próximo.

Sistema de mapeo semántico integral y modular

Para facilitar el despliegue de los mapas semánticos en escenarios reales, es crucial desarrollar un paquete integral y modular de mapeo semántico, el cual integre múltiples niveles de información en un único sistema unificado. Los sistemas existentes, como Kimera e Hydra, generan gráficos de escenas 3D con capas semánticas básicas, como objetos y lugares, pero carecen de modularidad y flexibilidad. Una solución idónea, preferiblemente implementada en ROS2 y/o en cualquier otro *framework* robótico ampliamente utilizado por la comunidad, permitiría a los usuarios mapear selectivamente determinados tipos de información en función de las necesidades de la aplicación. El paquete debería ser totalmente configurable, permitiendo a los usuarios elegir la primitiva geométrica que deseen para la

reconstrucción, seleccionar las redes neuronales que quieran utilizar para las diferentes tareas de percepción y adaptar el sistema a distintos niveles de requisitos computacionales. Esta flexibilidad permitiría una amplia gama de configuraciones, pudiendo ser configurados para dispositivos con recursos limitados optimizados para la eficiencia, o realizando configuraciones más avanzadas que utilicen la computación en el borde (*edge computing*) para aplicaciones más precisas y con un uso más intensivo de recursos.

Establecimiento de estándares de evaluación

La falta de estandarización en la evaluación de los métodos de mapeo semántico es un claro obstáculo para el progreso en este campo, lo que ha dado lugar hasta ahora a la utilización de una gran variedad de métricas, adoptadas cada una de ellas por diferentes autores en sus respectivos trabajos. Un ejemplo es la utilización de métricas diferentes para evaluar la reconstrucción de objetos, en función del tipo de primitiva utilizada. Proporcionar un sistema de evaluación estandarizado que trascienda de los tipos concretos de primitivas o datos utilizados sería extremadamente beneficioso para comprender mejor los avances en este campo. Además, un sistema unificado de evaluación facilitaría llevar a cabo comparaciones más exhaustivas entre los distintos métodos, fomentando la colaboración y promoviendo la innovación en el campo del mapeo semántico.

Adopción de técnicas de IA generativa

Las técnicas tradicionales de mapeo semántico se valen de redes de detección de objetos entrenadas en conjuntos de datos concretos, lo que las limita a un conjunto cerrado de categorías (*vocabulario cerrado*). Por ejemplo, los modelos entrenados en COCO, que incluye 80 categorías de objetos, pueden detectar electrodomésticos como microondas y hornos, pero no reconocen objetos como lavavajillas. Esto limita su aplicabilidad en entornos reales, donde muchos objetos permanecerían indetectables. La llegada de la IA generativa, en concreto los modelos de visión y lenguaje de gran tamaño (*Large Vision-Language Models*, LVLMs), ofrecen una solución al permitir la detección de objetos sin limitación de categorías detectables (*vocabulario abierto*), lo que permite la detección de una gama más amplia de objetos. Del mismo modo, los grandes modelos lingüísticos (*Large Language Model*, LLM) pueden sustituir a las bases de conocimientos semánticos tradicionales, que normalmente se construyen y mantienen mediante la *elicitación humana*. Este proceso implica la codificación manual de las propiedades, funcionalidades y relaciones de los elementos de la escena, lo cual requiere mucho tiempo y da lugar a una base de conocimientos cerrada que sólo refleja la información aportada por el experto. Aprovechando los LLM, esta etapa puede ser mejorada, creando una base de conocimiento semántico

abierta que reduce la necesidad de intervención continua del experto y amplía el conocimiento semántico. Estas son sólo dos aplicaciones potenciales, y la integración de la IA generativa en el mapeo semántico robótico abre innumerables posibilidades de avance en este campo.

Explotación de mapas semánticos en casos reales

El principal objetivo de la construcción de mapas semánticos es permitir su uso práctico por parte de los robots móviles, con el fin de realizar tareas de alto nivel en entornos con personas. En la actualidad, la mayor parte de la investigación se centra en la construcción de dichos mapas más que en su explotación práctica, y algunos trabajos sólo presentan pequeños casos de uso en escenarios controlados. Para avanzar en este campo, es esencial ir más allá de estas aplicaciones limitadas y facilitar la explotación integral de los mapas semánticos. En el futuro, los robots deberán ser capaces de utilizar estos mapas para razonar y tomar decisiones complejas, por ejemplo, utilizando los mapas semánticos como un gemelo digital de su espacio de trabajo, para ayudar a los humanos en una gama más amplia de tareas. Si fomentamos un enfoque más holístico de la explotación de los mapas semánticos, podremos liberar todo el potencial de los robots móviles para mejorar la productividad de las personas en situaciones cotidianas.

1 Introduction

1.1 Prologue

In the heart of a bustling hospital, a team of robots work closely with surgeons, nurses, and auxiliary staff, integrating seamlessly into the fast-paced world of healthcare. One of these robots, Sapin, moves around the operating theatre, quietly observing the scene as the surgical team prepares for a complex heart operation. The head surgeon, Dr. Garcia-Fuentes, exchanges a confident glance with her colleague, Dr. Heersche, who nods in acknowledgment of Sapin's presence. As the team begins, Sapin quickly anticipates their needs (see the left vignette in Figure 1.1): he approaches the scalpel to the surgeon when he needs it, adjusts the lighting to make sure the patient area is perfectly illuminated and ensures that the surgical drapes are meticulously in place. When the nurse asks for the suture needles, Sapin is already on his way and returns with the supply before he has time to worry. At that moment, Sapin becomes more than just a robot: he is a trusted colleague in the operating theatre, who, while not replacing the expertise of the medical team, facilitates their activity.

Meanwhile, across the city, a dad rushes into his local supermarket after a long day at the office. His daughter Paula's birthday is just around the corner, and he has little time to bake her a delicious cake. He feels overwhelmed, until Malagueto, the shop's friendly robot assistant, shows up at his side. Malagueto quickly assesses the situation and begins to guide him through the aisles, suggesting ingredients that not only fit a delicious cake recipe, but also match Paula's favorite flavors (see right vignette in Figure 1.1). As they walk, Malagueto remembers that Paula has a special love for a particular type of chocolate that she always chooses, so he suggests combining it with

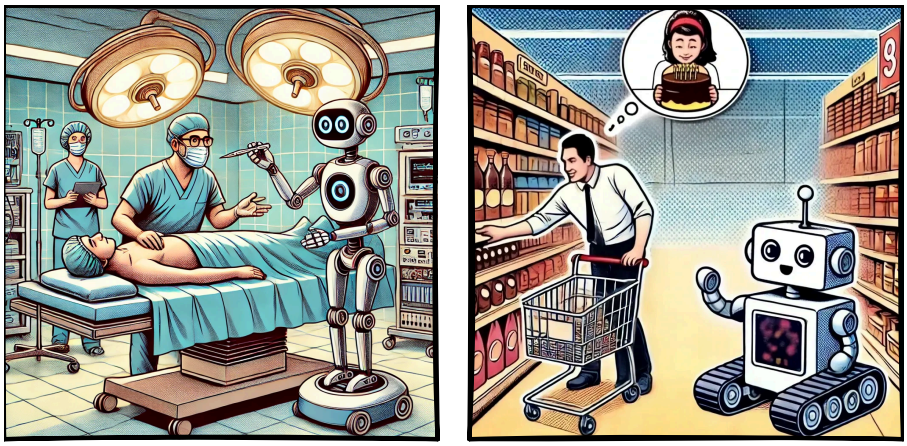


Figure 1.1: Illustration of two simultaneous scenarios in a city, in which robots make life easier for humans. On the left, a mobile robot collaborates with the medical team in surgery, while on the right, a shopping assistant robot guides a father in choosing the ingredients for his daughter's birthday cake. Images created with the generative AI model DALL·E.

fresh, seasonal strawberries. This will make a delicious topping! –Malagueto exclaims in a cheerful tone. They continue their journey through the store, while Malagueto suggests other ideas, such as adding colored sprinkles and a *"Happy Birthday, Paula!"* note made of cream. The dad now feels much more relaxed and breathes a sigh of relief.

These two robots, Sapin and Malagueto, illustrate how robotics can play a vital role in improving people's quality of life. While the rapid advancement of robotics and artificial intelligence often raises concerns about job displacement, it is crucial to clarify that these technologies are not here to replace us but to facilitate our daily routines and enrich our lives. Rather than perceiving robots as competitors, we should embrace them as supportive partners who can alleviate our workload, allowing us to focus on the tasks that truly matter. Yet, to reach this point, regardless of their application area (*e.g.* hospitals, factories, greenhouses, etc.), there is a common requirement shared by all robots: the need for an internal representation of their workspace to achieve comprehensive scene understanding, enabling robots to process semantic concepts through perception and perform complex reasoning on their environment.

1.2 Introduction and motivation

Semantic maps represent high-level information about the individual elements present in an environment, incorporating not only the categories of objects and types of places[1] but also their attributes, such as physical characteristics, functionalities, and states. In addition, these maps capture the relationships between different elements, describing how they interact or connect with each other. For instance, a relationship between objects could be expressed as `clean socks are usually on the nightstand or in the wardrobe`, while an object-place relationship could be formulated as `the deodorant is generally found in the bathroom`.

Despite the significant attention that semantic mapping has received in recent years, as evidenced by the number of contributions in this research area [1–7], a notable gap remains between theory and practical application. Next, we will introduce the specific challenges that contribute to this gap and hinder the building and maintenance of reliable semantic maps in real-world scenarios.

A primary challenge stems from the limitations of off-the-shelf scene understanding algorithms, such as place categorization and object detection, when applied to robotic contexts. These algorithms demonstrate remarkable performance under ideal conditions (*e.g.* good lighting conditions, high-quality images and uncluttered environments) but this rarely reflects the reality faced by mobile robots. Onboard robot cameras typically have low resolution and a narrow field-of-view, often due to the use of low-cost sensors or limited processing capabilities. As a result, images captured during the robot's exploration may be blurred due to motion, including occluded views of objects, or depict non-informative scenes, such as empty walls, among other unfavorable conditions. Such practical issues significantly degrade the performance of these algorithms, leading to potential errors that could propagate through the mapping process, resulting in impractical and even unsafe robot operations. Imagine a robot operating in a hospital, which misclassified an operating theater as a regular patient room. Later, when the robot is commanded with a mission to check patients, it could disrupt and distract surgeons in a critical moment.

Another challenge arises from the unavoidable uncertainty inherent to the robots' perception of the world through their sensors, where factors such as sensor noise and environmental variations lead to the presence of disturbances in the acquired data. Similarly, methods and algorithms that operate on this raw data to detect, recognize, or localize elements –allowing the robot to perceive some aspects of the world– are also subject to uncertainty, making it essential to correctly propagate uncertainty through

[1] Note that in this thesis, the term *place* is not restricted to enclosed rooms but also includes meaningful sub-areas within a space such as a kitchen and a living area in a studio apartment.

Figure 1.2: Illustrative example of the degraded performance of the Detectron2 [8] object detection network on an image containing out-of-distribution objects. The results demonstrate that the network correctly identifies in-distribution objects like the refrigerator (84%) and microwave (96%). However, it produces overconfident yet incorrect classifications when encountering out-of-distribution objects such as mobile robots, labeling them as a cake (72%), a train (82%), and a cell phone (52%).

subsequent processing stages, a requirement that is not always met. This is particularly evident in the widely adopted and highly effective neural networks, where a major source of uncertainty arises from their typical training on a fixed, closed set of categories. The latter becomes a limitation when these networks encounter out-of-distribution categories (*i.e.* those not present in their training data), resulting in significant performance degradation. For instance, in object detection, such networks may produce ambiguous classifications, such as identifying an object as a `refrigerator` with 47% confidence and a `cabinet` with 53% confidence. They may also yield inconsistent yet overconfident classifications across consecutive detections of the same object, such as recognizing it as a `cake` with 72% confidence in one frame and, in the next frame, labeling it a `backpack` with 78% confidence. These cases illustrate scenarios of high classification uncertainty, where the correct category remains unclear (probably because it is not any of the detectable categories). Unfortunately, this uncertainty is frequently ignored by semantic mapping approaches, often just assigning the highest-scored class without considering how possible the other classes are. Yet, the highest-scored class is not always the correct one, as it could be an out-of-distribution object or simply be misclassified due to being observed

from challenging viewpoints. Ignoring this uncertainty may result in erroneous robot operation as, for instance, illustrated in Figure 1.2, in which an out-of-distribution object is misclassified as `cake`. Now, following the story described in the prologue section of this thesis, imagine we are on Paula's birthday, and his dad commands the robot: *"hey Malagueto, please bring us the birthday cake!"* (of course, Malagueto was also invited to the birthday party). The result would likely leave Paula and the guests disappointed, and rightly so, because of the disgusting "cake". This neglection of uncertainty extends to other parts of the mapping process, being robot localization one of the most important ones. While semantic mapping approaches often assume ground-truth localization, it is not realistic for real-world settings. Consequently, approaches that are not prepared to handle these uncertainties will unavoidably result in an erroneous operation.

These are some of the main challenges faced when deploying semantic mapping approaches in real-world environments. Yet, there are still other limitations that also require attention, such as the need for highly pre-processed input data, which is often difficult to obtain, or handling and storing large volumes of data generated during mapping, which can not be guaranteed especially when working with resource-constrained mobile robots.

1.3 Contributions

This thesis contributes to enhancing the generation of semantic maps with mobile robots, focusing specifically on overcoming the common challenges encountered when deploying traditional semantic mapping algorithms in real-world environments. However, it is important to note that the contributions of this thesis are focused on the challenges related to perception and knowledge representation and do not address the exploitation of this knowledge for high-level semantic reasoning. Additionally, it explores how the information stored on these semantic maps can be leveraged to solve the problem of cross-detector visual localization, a novel challenge introduced and addressed for the first time in the literature through this work. By tackling both semantic mapping and localization, the thesis provides a comprehensive approach to improving the robustness and applicability of mobile robotic systems in dynamic environments.

Concretely, the contributions of this thesis can be divided into four main topics. The following sections outline the common challenges in each topic and highlight the specific contributions this thesis makes to address them.

Place categorization

Place categorization algorithms attempt to determine the category of a place from either a single image or a sequence of images [1, 9–12]. When applied to robotics, where images often capture empty walls or regions, the performance of these algorithms degrades considerably as a consequence. In this context, a common concern with these approaches when applied to robotics is the lack of consideration from which point-of-view images should be captured. The latter is of great importance to maximize the information contained in these images about the place category (*e.g.* prioritizing to observe objects that are typically only found in specific places, such as toilets in bathrooms), in order to ensure the correct categorization.

In our work [13], we introduce an attention mechanism based on active vision that aims to enhance the efficiency and accuracy of place categorization algorithms. This mechanism dynamically selects the camera's line-of-sight using a pan-only unit to maximize the information gain at each moment. Specifically, it is formalized as a next-best-view problem within a Markov Decision Process (MDP) model. Our approach is evaluated with algorithms from both main paradigms of place categorization (*i.e.* object-based and image-based methods) and demonstrates its effectiveness by outperforming standard camera configurations used in robotics.

3D structural reconstruction

The generation of structural maps, also known as 3D floor plan reconstruction, involves collecting data from onboard sensors, such as RGB-D cameras and range sensors, while the robot explores the environment. Despite the promising results achieved in recent years, most existing approaches face a series of limitations that hinder their application in real-world scenarios. First, a major limitation is the need for a dense point cloud of the entire environment as input, which requires storing and handling large volumes of data. Second, these methods often overlook the inherent uncertainty in robot localization, which can lead to misaligned point clouds during registration. This concern is usually ignored by assuming the point clouds are already aligned. Third, some approaches rely on the unrealistic assumption of a Manhattan world, where all structural elements are orthogonal. These limitations reduce the practicality of these methods in real-world scenarios, where extensive pre-processing is not feasible, and more efficient solutions are needed.

In [14] we propose *Sigma-FP*, a method for the reconstruction of 3D floor plans from RGB-D sequences. The main advantage of *Sigma-FP* lies in its online operation, which mitigates the need for storing extensive data volumes by processing information on the fly and retaining only a compact

representation of structural elements (*i.e.* planes). *Sigma-FP* employs a probabilistic approach to account for the inherent uncertainties in both robot localization and plane estimation, therefore enhancing the reconstruction robustness. Moreover, *Sigma-FP* relaxes the room geometry assumption to the Atlanta world model, where the restriction is only that structural elements must be vertical. Finally, *Sigma-FP* not only reconstructs the walls, but also openings such as doors and windows, providing a more accurate and detailed reconstruction. Our approach is evaluated in environments of diverse nature, outperforming state-of-the-art methods that struggle when facing real-world conditions such as localization uncertainty.

Object-oriented semantic mapping

Object detection is a critical task for populating semantic maps with objects in a robot's workspace. While recent advances in deep learning have significantly improved object detection performance, these methods often perform poorly in robotic applications. This degradation is largely due to the particular characteristics of images captured by the sensors mounted on the robots, which typically have low resolution and a narrow field of view. Additionally, robot motion can result in blurred images, causing the detection outputs to vary considerably between consecutive frames, leading to a lack of consistency. Another challenge arises when robots capture images during exploration, where object detection must often handle difficult conditions such as occlusions, partial views, and changing lighting. These factors hinder the construction of reliable semantic maps, often resulting in multiple instances of the same physical object or misclassified objects, which can lead to wrong robot operation.

The work presented in [15] proposes a method to enhance object detection in image sequences captured by robots. By leveraging knowledge of the camera motion between consecutive frames, the approach propagates detections across frames and integrates them using a recursive Bayesian filter. This method ensures spatio-temporal consistency in the network's outputs for robotic applications. The results demonstrate its effectiveness, with only a minimal increase in processing time.

In [16], we present *LTC-Mapping*, a method for building object-oriented semantic maps using 3D bounding boxes as geometric primitives, with the main objective of ensuring long-term reliability for mobile robot operation. This method addresses two critical challenges that affect the long-term reliability of these maps: instance duplication and dynamic environments. To mitigate instance duplication, *LTC-Mapping* labels the vertices of each bounding box with visibility flags, indicating whether they have been observed. This information is crucial to determine if an object has only been partially observed due to partial views or occlusions, enabling us to identify multiple instances coming from partial views of the same physical object, and

to integrate these into a single, complete instance. In addition, to handle dynamic environments, we introduce the concept of non-detection, which facilitates the removal of objects from the map when they are no longer observed, ensuring that the map is kept up-to-date by taking into account both new observations and missing detections. These techniques have been proven to improve the reliability and robustness of semantic maps over time.

Lastly, in [17], we introduce *Voxeland*, an alternative for object-oriented semantic mapping that uses voxels as geometric primitives. This approach addresses common challenges in neural network predictions, such as overconfident incorrect predictions with out-of-distribution objects, and the generation of inaccurate masks. To overcome these issues, we propose a probabilistic framework inspired by the Theory of Evidence, where each network prediction is treated as a subjective opinion at both geometric and semantic levels. These opinions are aggregated over time into evidence using a probabilistic model, allowing for the quantification of uncertainty at both levels. The latter enables us to identify objects requiring reobservation or reclassification. In this work, we propose a strategy to leverage this information by implementing semantic disambiguation through the integration of a Large Vision-Language Model (LVLM). This model generates a more complete opinion for objects that exhibit high semantic uncertainty. The evaluation highlights the importance of incorporating and exploiting uncertainty to enhance the robustness and reliability of the resulting map.

Cross-detector visual localization for long-term interoperability

A prerequisite for building semantic maps is to be able to know, at each time moment, the robot localization, which is essential to place the observed scene elements with respect to a reference map. An ideal aspect would be for the localization method to be compatible with multiple robots operating in the same environment, allowing interoperability between them such as collaborating in the construction and maintenance of the semantic map. In addition, this localization approach should also be applicable in the long-term, in order to keep the map updated with changes that occur in the real-world (*e.g.* relocation of an object, or the removal of an element that is no longer in the environment). One solution to this problem is Visual Localization (VL), which uses on-board visual sensors such as cameras, to extract features from a query image to subsequently establish correspondences with features available in an existing 3D map. However, this approach has an important limitation: it works under the assumption that the feature detection algorithm used for both the query and the map are the same.

In our work [18], we extend the problem of VL to be applicable to heterogeneous detectors, which we coin as *Cross-Detector Visual*

Localization. The major challenge of this problem lies in the inherent spatial discrepancy of heterogeneous keypoints representing different, yet close, physical entities, which hinders the process of establishing correct correspondences. To tackle this, we leverage information about the structural elements available in the semantic map to impose coplanarity constraints during the matching process. The results demonstrate that, despite a small computational overhead, the consideration of these coplanarity constraints effectively guides the matching process, allowing to establish correct correspondences between heterogeneous keypoints.

1.4 Publications

This thesis is presented under the modality of *thesis supported by a compendium of relevant published papers*. Concretely, it encompasses the following publications:

Journals

- *Jose-Luis Matez-Bandera, Pepe Ojeda, Javier Monroy, Javier Gonzalez-Jimenez and Jose-Raul Ruiz-Sarmiento.* **Voxeland: Probabilistic Instance-Aware Semantic Mapping with Evidence-based Uncertainty Quantification.** Submitted and under review, (2024).

- *Jose-Luis Matez-Bandera, Alberto Jaenal, Clara Gomez, Alejandra C. Hernandez, Javier Monroy, José Araújo and Javier Gonzalez-Jimenez.* **Cross-Detector Visual Localization with Coplanarity Constraints for Indoor Environments.** Submitted and under review, (2024).

- *Jose-Luis Matez-Bandera, Javier Monroy and Javier Gonzalez-Jimenez.* **Sigma-FP: Robot Mapping of 3D Floor Plans with an RGB-D Camera Under Uncertainty.** In IEEE Robotics and Automation Letters, vol. 7, no. 4, pp. 12539-12546, (2022).
  DOI: 10.1109/LRA.2022.3220156

- *Jose-Luis Matez-Bandera, David Fernandez-Chaves, Jose-Raul Ruiz-Sarmiento, Javier Monroy, Nicolai Petkov and Javier Gonzalez-Jimenez.* **LTC-Mapping, Enhancing Long-Term Consistency of Object-Oriented Semantic Maps in Robotics.** In MDPI Sensors, vol. 22, no. 14, (2022).
  DOI: 10.3390/s22145308

- *Jose-Luis Matez-Bandera, Javier Monroy and Javier Gonzalez-Jimenez.* **Efficient Semantic Place Categorization by a Robot through Active Line-of-Sight Selection** In Knowledge-Based Systems, vol. 240, pp. 108022-108034, (2022). DOI: 10.1016/j.knosys.2021.108022

Conference proceedings

- *David Fernandez-Chaves, Jose-Luis Matez-Bandera, Jose-Raul Ruiz-Sarmiento, Javier Monroy, Nicolai Petkov and Javier Gonzalez-Jimenez.* **Exploiting Spatio-Temporal Coherence for Video Object Detection in Robotics.** In International Conference on Computer Analysis of Images and Patterns, (2021). DOI: 10.1007/978-3-030-89131-2_17

1.5 Thesis framework

This thesis is the result of four years of work as a researcher in the Machine Perception and Intelligent Robotics (MAPIR[2]) research group, part of the Department of Systems Engineering and Automation of the University of Malaga. The author received the FPU (*Formación de Profesorado Universitario*) grant (FPU19/00704) by the Spanish Ministry of Science, Innovation and Universities, which was the main funding source of this research. This research has also been supported by the group's research projects, in particular, the three national projects WISER (DPI2017-84827-R), ARPEGGIO (PID2020-117057GB-I00) and MINDMAPS (PID2023-148191NB-I00), and the regional project VOXELAND (JA.B1-09).

During this period, the author successfully completed the PhD program in Mechatronics Engineering, which is coordinated by the Department of Systems Engineering and Automation of the University of Malaga. Within this multidisciplinary program, the author gained extensive experience across the different domains that form the foundation of mechatronics, including mechanics, electronics, programming and automation. In addition, the author complemented his academic education with technical courses of great relevance for his research such as the *Artificial Intelligence* by Samsung Innovation Campus (September-December 2020), *Getting Started with AI on Jetson Nano* (February 2021) and *Fundamentals of Deep Learning for Multi-GPUs* (June 2021), both by NVIDIA.

[2] mapir.isa.uma.es

Throughout this thesis, the author actively engaged with the scientific community of the two main areas of his thesis, robotics and computer vision, attending and presenting his work in multiple national and international conferences such as the *International Conference on Computer Analysis of Images and Patterns* (Cyprus, remotely, 2021), the *International Conference on Robotics and Automation* (London, 2023), and the *Jornadas de Automática* (Malaga, 2024). Besides, the author participated as a volunteer in the organization of the *European Robotics Forum* in 2020. It is also worth to mention his active participation as a reviewer –something extremely necessary in the scientific research area– for prestigious conferences and journals, including the IEEE International Conference on Robotics and Automation (ICRA), the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), and the IEEE Transactions on Robotics (T-RO), among others.

The author also had the opportunity to complete a four months research internship at Ericsson AB in the offices located in Stockholm (Sweden), within the Sensing and Perception team, under the supervision of Dr. José Araújo. During this internship, the author acquired knowledge on topics tangential to the core of his thesis, mainly focusing on heterogeneous visual localization. This experience allowed him to incorporate new knowledge into his research while contributing to the team's progress. Not less important, this internship provided the author with the opportunity to establish close collaborations with other researchers in the team.

The FPU grant also offered the opportunity to collaborate as a teaching assistant at the Department of System Engineering and Automation of the University of Malaga. In particular, the author taught in the School of Computer Engineering the subjects of *Biomedical System Modelling* (2021/2022, 2022/2023 and 2023/2024), *Robot Programming* (2021/2022) and *Automatic Control* (2023/2024).

From the author's collaboration with members of the MAPIR research group and international researchers, a number of additional publications have resulted:

Conference proceedings

- *Jesús Moncada Ramírez, Jose-Raul Ruiz-Sarmiento, Jose-Luis Matez-Bandera, Javier Gonzalez-Jimenez.* **Modelos a gran escala para mapeo semántico en robótica móvil.** In XLV Jornadas de Automática, (2024).
  DOI: 10.17979/ja-cea.2024.45.10940

- *Gregorio Ambrosio-Cestero, Jose-Luis Matez-Bandera, Jose-Raul Ruiz-Sarmiento, Javier Gonzalez-Jimenez.* **Entorno basado en contenedores Linux para el desarrollo de aplicaciones**

robóticas. In XLV Jornadas de Automática, (2024).
DOI: 10.17979/ja-cea.2024.45.10943

Workshops

- *Matteo Luperto, Jose-Luis Matez-Bandera, Tomasz Piotr Kucner, Michele Antonazzi, Gabriele Somaschini, Javier Monroy, Javier Gonzalez-Jimenez, Nicola Basilico.* **Ensuring the Consistency of Heterogeneous World Representations Using Structural Features.** In ICRA 2023 Workshop on Unconventional spatial representations. Opportunities for robotics, (2023).

Patents

- *Jose-Luis Matez-Bandera, Alberto Jaenal, Clara Gomez, Alejandra C. Hernandez, and José Araújo.* **Determining Location of a Device within an Environment Comprising Planar Surfaces.** Under review.

1.6 Thesis outline

The subsequent chapters of this thesis are organized as follows:

Chapter 2: Theoretical background introduces the fundamental concepts supporting this thesis, focusing on the three main pillars: mathematics, computer vision, and scene understanding. Each section covers the basics of these areas to facilitate its understanding and explains their specific relevance and application within the context of this work.

Chapter 3: Place categorization presents an attention mechanism for mobile robots to improve the efficiency of place categorization methods by continuously selecting the most informative point-of-views from where to capture images.

Chapter 4: 3D reconstruction of structural elements introduces a method for the automatic reconstruction of 3D floor plans of multi-room environments from a sequence of RGB-D images, while accounting for the probabilistic nature of robot localization and plane estimation.

Chapter 5: Object-oriented semantic mapping initially proposes an approach that leverages knowledge of camera motion between frames to provide spatio-temporal coherence in video object detection for robotics. The chapter then describes two object-oriented semantic mapping methods. The first method uses 3D bounding boxes as geometric primitives to represent objects, placing importance on efficiency. The second method improves reconstruction accuracy by employing voxels as geometric primitives. Both methods support online operation and are classified as *instance-aware*, meaning that objects are represented as individual entities, which is crucial for effective interaction with scene elements.

Chapter 6: Cross-detector visual localization for long-term interoperability introduces and formalizes cross-detector visual localization, a challenging problem aiming to ensure the interoperability of cameras with heterogeneous detectors within a single map representation. Subsequently, it presents a method which leverages information from structural planes to tackle this problem.

Chapter 7: Conclusions and future work concludes this PhD dissertation, summarizing the main results of this work and suggesting possible lines of future research.

It is important to note that there is no general *related work* section. Instead, each section in the following chapters provides a comprehensive review of relevant literature, specific to the topics addressed.

2 Theoretical background

This section aims to introduce the fundamentals on which this thesis is based, with the purpose of providing the reader with the necessary background to facilitate the understanding of the thesis. In particular, this chapter is organized in three sections, each one covering one of the main pillars: i) mathematics, ii) computer vision and iii) scene understanding. In each section, we present and describe the essentials of each topic that has been used in this work, along with their specific applications within this thesis.

2.1 Mathematical foundations

2.1.1 Markov decision processes

A Markov Decision Process (MDP) [19] is defined as a stochastic process which provides a mathematical approach to formalize problems that involve making a sequence of decisions over time, aiming to optimize a certain objective, typically the cumulative reward. Concretely, in this thesis, we particularized an MDP to model an information maximization problem aiming at improving the efficiency of place categorization methods, as described in Section 3.A. By definition, an MDP is represented by a 5-tuple (S, A, P, R, γ):

- **States (S).** The finite set of all possible states, which form the state-space.
- **Actions (A).** The finite set of actions that can be considered by the decision-maker, confirming the action-space.
- **Transition Probability (P**: $\mathrm{S} \times \mathrm{A} \to \mathrm{S}$**).** The probability of reaching the state s' after taking the action a from state s, and is denoted as $\mathrm{P}(s'|s,a)$.
- **Reward Function (R**: $\mathrm{S} \times \mathrm{A} \to \mathbb{R}$**).** The expected reward to receive after transitioning from state s to state s' by taking action a, and is represented as $\mathrm{R}(s,a,s')$.
- **Time-Horizon Factor (f_γ).** A factor $f_\gamma \in [0,1]$ that adjusts the influence of rewards according to the time-horizon γ, placing greater value on rewards of immediate actions compared to those from prospective actions.

The objective of an MDP is to find the optimal policy $\hat{\pi}$, that is, a function $a = \pi(s)$ which yields the action a that maximizes the expected cumulative reward starting from state s. In this sense, the value of considering a policy π in state s is quantified as follows:

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\gamma} f_\gamma^t \mathrm{R}(s_t, a_t, s_{t+1})\right]. \tag{2.1}$$

The value function $V^{\pi}(s)$ is decomposed in two terms, deriving the Bellman's equation, in which the first term states for the reward of the current state while the second term refers to the reward of next states:

$$V^{\pi}(s) = R(s,a,s') + \sum_{s' \in \mathrm{S}} f_\gamma \mathrm{P}(s'|s,a) V^{\pi}(s') \tag{2.2}$$

The optimal policy $\hat{\pi}(s)$ is derived by searching the optimal value function $V^{\pi}(s)$ that provides the maximum value achievable from state s:

$$\hat{\pi}(s) = \underset{a}{\operatorname{argmax}} \left\{ R(s,a,s') + \sum_{s' \in \mathrm{S}} f_\gamma \mathrm{P}(s'|s,a) V^{\pi}(s') \right\} \tag{2.3}$$

In practice, solving an MDP often involves iterative algorithms such as value iteration and policy iteration, among others.

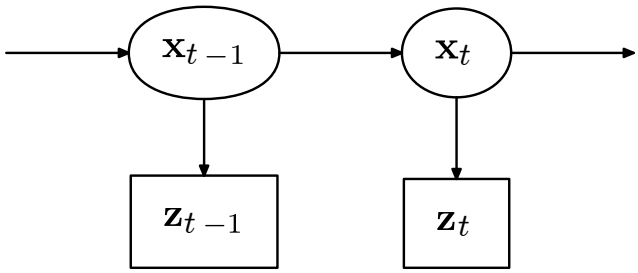


Figure 2.1: Bayesian network of a hidden Markov model (HMM).

2.1.2 Recursive Bayesian filter

The recursive Bayesian filter [20] is a probabilistic framework for estimating the state of a dynamic system over time, especially under conditions of uncertainty and noise. In particular, this framework involves two steps that are recursively performed: i) prediction, in which the probable future is estimated based on the previous observations, and ii) update, where the current state is estimated given previous and current observations. Particularly, in this thesis, we applied the recursive Bayesian filter to integrate observations over time, thereby providing consistency to the classification in place categorization and object detection tasks.

These classification tasks over time can be expressed as a hidden Markov model (HMM), in which we assume relevant variables to be normally distributed. Concretely, the latent or hidden variable $\mathbf{x}_t = \{\mathrm{x}^{(1)}, ..., \mathrm{x}^{(N)}\}$, represents the state of the system at time t to be estimated (*i.e.* the place or object category), while $\mathbf{z}_t$ depicts the observation gathered at time t, which is known and dependent on the unobservable state $\mathbf{x}_t$. Figure 2.1 shows the Bayesian network representing such HMM.

The process begins with a prior distribution $\mathrm{P}(\mathbf{x}_0)$ over the system's state, which is typically modeled with a uniform distribution under a fully uncertain scenario or with a particular distribution if any prior knowledge is available. The evolution of this state over time is captured by the transition model $\mathrm{P}(\mathbf{x}_1|\mathbf{x}_0)$, which describes how the state changes from $\mathbf{x}_0$ to $\mathbf{x}_1$. Using this model, the prior distribution is propagated forward to obtain the predicted distribution $\mathrm{P}(\mathbf{x}_1)$. After initializing with the predicted distribution, the first observation z_1 is incorporated through the update step using Bayes' theorem:

$$\mathrm{P}(\mathbf{x}_1|\mathrm{z}_1) = \frac{\mathrm{P}(\mathrm{z}_1|\mathbf{x}_1)\mathrm{P}(\mathbf{x}_1)}{\mathrm{P}(\mathrm{z}_1)}, \tag{2.4}$$

where $\mathrm{P}(\mathbf{x}_1|\mathrm{z}_1)$ states for the posterior distribution updated with the first observation, $\mathrm{P}(\mathrm{z}_1|\mathbf{x}_1)$ is the likelihood function, which represents how likely

is z_1 given $\mathbf{x}_1$, and $P(z_1)$ acts as a normalization factor for the posterior to integrate up to 1.

From now on, the posterior distribution becomes the current belief about the system's state, and the update step at an arbitrary time t is formulated as:

$$P(\mathbf{x}_t|\mathbf{z}_{1:t}) = \frac{P(z_t|\mathbf{x}_t, \mathbf{z}_{1:t-1})P(\mathbf{x}_t|\mathbf{z}_{1:t-1})}{P(z_t|\mathbf{z}_{1:t-1})}, \tag{2.5}$$

which computes the state of the system at time t upon all observations from the past. By assuming they are conditional independent of each other, the formulation can be simplified to:

$$P(\mathbf{x}_t|\mathbf{z}_{1:t}) = \frac{P(z_t|\mathbf{x}_t)P(\mathbf{x}_t|\mathbf{z}_{1:t-1})}{P(z_t|\mathbf{z}_{1:t-1})}. \tag{2.6}$$

The formulation expressed in Eq. (2.6) is the general form of a Bayesian filter's update step. However, in robotics and computer vision, areas in which the present thesis is framed, the observations often comes from neural networks. These networks typically yield the posterior distribution $P(\mathbf{x}_t|z_t)$ instead of the desired likelihood function $P(z_t|\mathbf{x}_t)$. To adapt the formulation to such scenarios, we employ Bayes' theorem for the likelihood term, which results in the update step to be rewritten as:

$$P(\mathbf{x}_t|\mathbf{z}_{1:t}) = \frac{P(\mathbf{x}_t|z_t)P(z_t)P(\mathbf{x}_t|\mathbf{z}_{1:t-1})}{P(\mathbf{x}_t)P(z_t|\mathbf{z}_{1:t-1})}. \tag{2.7}$$

Finally, considering that all possible states for $\mathbf{x}_t$ are equally probable a priori, and the terms $P(z_t)$ and $P(z_t|\mathbf{z}_{1:t-1})$ are equal for any $\mathbf{x}_t$, we group these terms together as a normalization factor. Then, the formulation is reduced to:

$$P(\mathbf{x}_t|\mathbf{z}_{1:t}) \propto P(\mathbf{x}_t|z_t)P(\mathbf{x}_t|\mathbf{z}_{1:t-1}), \tag{2.8}$$

which states that the probability of the current state $\mathbf{x}_t$ given the set of all observations through time can be estimated from the probability estimation derived only from the most recent observation, and the previous estate estimation.

2.1.3 Dirichlet distribution

The Dirichlet distribution [21] is a multivariate probability distribution that models the probabilities of a set of mutually exclusive events, making it particularly effective for representing distributions over categorical data. In the context of this thesis, Dirichlet distributions are employed to probabilistically model two critical aspects of scene understanding: i) the probability that a specific discrete volume within an environment belongs to a particular object, and ii) the probability that an object belongs to each of

the possible object categories. This probabilistic modelling plays a key role, as it enables uncertainty quantification, which is crucial for real-world applications that are inherently subject to various sources of error and noise.

Formally, a Dirichlet distribution, denoted as $\mathrm{Dir}(\boldsymbol{\alpha})$, is parameterized by a concentration vector $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \ldots, \alpha_K\}$, where each $\alpha_i > 0$ and K represents the number of possible categories. The probability density function (PDF) of the Dirichlet distribution is given by:

$$\mathrm{Dir}(\boldsymbol{x}; \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}, \tag{2.9}$$

where $\boldsymbol{x} = \{x_1, x_2, \ldots, x_K\}$ represents the categories probabilities and belong to the space called $K-1$ simplex, which implies that $\sum_{i=1}^{K} x_i = 1$ and each $x_i \in [0, 1]$. $B(\boldsymbol{\alpha})$ is the multivariate Beta function, which acts as the normalizing term and is defined as follows:

$$B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}, \tag{2.10}$$

where $\Gamma(\cdot)$ is the Gamma function.

A key property of the Dirichlet distribution is its role as the conjugate prior for categorical and multinomial distributions. This property is of particular relevance to the scope of this thesis, as the common nature of robotic semantic mapping is incremental, requiring continuous updates to the map as the robot explores its environment.

To better understand its usefulness, let's illustrate it with a practical example. Consider an scenario where a robot aims to determine the category of a certain object in its workspace. Initially, the robot assumes a prior Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\alpha})$, with concentration parameters $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \ldots, \alpha_K\}$, where K denotes the number of possible object categories. The robot processes an image through a Convolutional Neural Network (CNN), which outputs a confidence score vector ($\boldsymbol{s} = \{s_1, s_2, \ldots, s_k\}$). This confidence score vector can be interpreted as an additional set of concentration parameters for another Dirichlet distribution. To update the prior distribution with the incoming observation, the concentration parameters of the Dirichlet distribution are updated as follows:

$$\alpha_i' \leftarrow \alpha_i + s_i, \tag{2.11}$$

where $\boldsymbol{\alpha'}$ are the concentration parameters of the posterior distribution, which due to the Dirichlet distribution's properties, it still remains following a Dirichlet distribution.

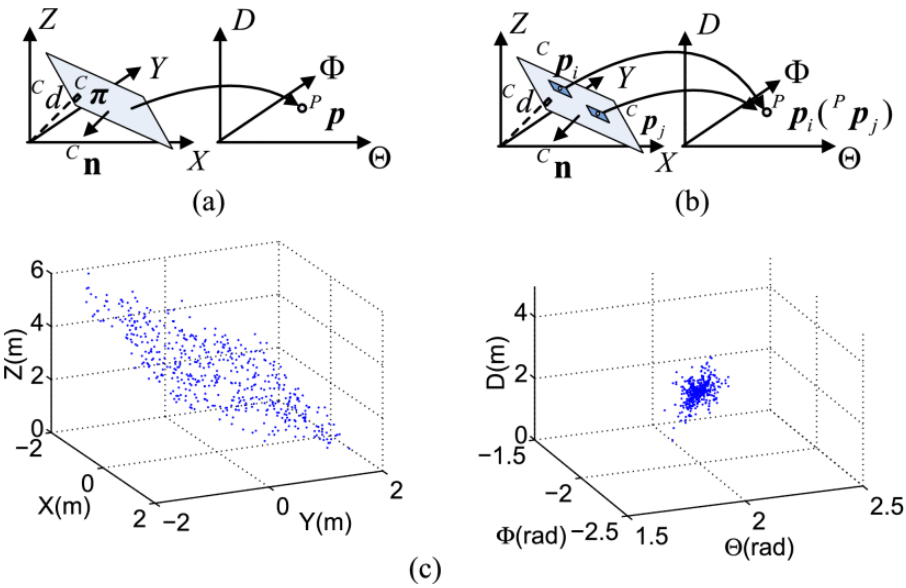


Figure 2.2: Mapping of planar surfaces from Cartesian space into PPS. (a) Illustrates the projection of a Cartesian plane π into the PPS. (b) Two planar patches sharing same support plane are projected into the same location in the PPS. (c) A planar surface from point cloud data projected into the PPS. Images obtained from [22].

2.1.4 Plane parameter space

The plane parameter space (PPS) [22] is a compact representation space for planes defined in the Cartesian space. Specifically, the PPS maps a Cartesian plane to a single point within this space, enhancing the efficiency and robustness of tasks such as plane extraction and association. This thesis leverages the PPS to segment planes from noisy point clouds and to perform plane matching, which are critical tasks for 3D floor plan reconstruction.

For a formal definition, let's consider a plane $\pi^C = [\mathbf{n}_\pi, d_\pi]$ defined in the Cartesian space by its unit normal vector $\mathbf{n}_\pi^C = [\mathrm{n}_x, \mathrm{n}_y, \mathrm{n}_z]$ and its distance-to-origin d_π. The projection of that plane in the PPS is a point $\mathbf{p}^{\mathrm{PPS}}$ defined by two angles, the elevation $\Theta_\mathbf{p}$ and the azimuth $\Phi_\mathbf{p}$ angles of the normal vector, respectively, and the distance-to-origin $d_\mathbf{p}$ of the plane (see Figure 2.2a). Mathematically, the projection from Cartesian space to PPS is given by:

$$\mathbf{p}^{\mathrm{PPS}} = \begin{bmatrix} \Theta_\mathbf{p} \\ \Phi_\mathbf{p} \\ d_\mathbf{p} \end{bmatrix} = \begin{bmatrix} \arccos(\mathrm{n}_z) \\ \mathrm{atan2}(\mathrm{n}_y, \mathrm{n}_x) \\ d_\pi \end{bmatrix}. \tag{2.12}$$

Likewise, from a point $\mathbf{p}^{\mathrm{PPS}}$ represented in the PPS, its representation π^C in the Cartesian space can be retrieved as follows:

$$\pi^C = \begin{bmatrix} \mathbf{n}_\pi \\ d_\pi \end{bmatrix} = \begin{bmatrix} \mathrm{n}_x \\ \mathrm{n}_y \\ \mathrm{n}_z \\ d_\pi \end{bmatrix} = \begin{bmatrix} \sin(\Theta_\mathbf{p})\cos(\Phi_\mathbf{p}) \\ \sin(\Theta_\mathbf{p})\sin(\Phi_\mathbf{p}) \\ \cos(\Theta_\mathbf{p}) \\ d_\mathbf{p} \end{bmatrix}. \tag{2.13}$$

As mentioned above, representing planes in the PPS offers a particular advantage: planes are represented as points. This means that two planar patches sharing the same support plane are represented by the same point in the PPS (see Figure 2.2b), which simplifies plane association. Moreover, in robotics, scene geometry is often acquired through range sensors that produce point cloud data. Identifying high-level geometric elements, such as planes, is generally a challenging task. However, by leveraging the PPS definition, one can search for local planar regions, that is, points that are locally planar with their neighbors. Therefore, projecting the local planar patch associated to each 3D point into the PPS, as shown in Figure 2.2c, facilitates the plane segmentation even in the presence of noise.

2.2 3D Computer vision basics

2.2.1 Planar transformations: homography

In the context of 3D computer vision, a homography describes the mapping between two different views of the same planar surface in the 3D world, as illustrated in Figure 2.3. Mathematically, a homography $\mathbf{H}$ is a 3×3 matrix that defines the transformation of a point $\mathbf{p_1} = [x_1, y_1]$ in the first image to its corresponding point $\mathbf{p_2} = [x_2, y_2]$ in the second image as follows:

$$\lambda \begin{bmatrix} x_2 \\ y_2 \\ 1 \end{bmatrix} = \underbrace{\begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}}_{\mathbf{H}} \begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix}, \tag{2.14}$$

where λ is a scalar that accounts for the scale introduced by the projective transformation and h_{ij} represent the elements of the homography matrix that encapsulate the rotational, translational, and projective components of the transformation. The points $\mathbf{p_1}$ and $\mathbf{p_2}$ are expressed in homogeneous coordinates to facilitate the representation of the transformation in matrix form.

To compute the homography matrix $\mathbf{H}$, correspondences between at least four pairs of points in the two images are required [23]. These

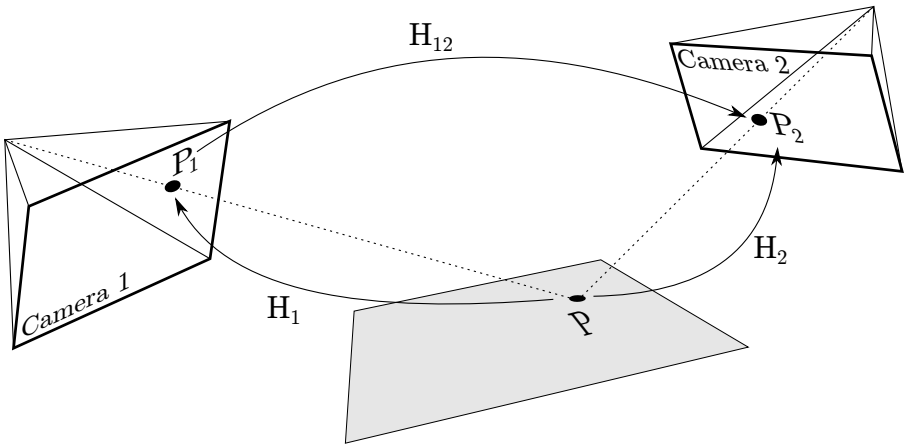


Figure 2.3: Illustration of the existing homographies between a 3D plane and its projection in two different cameras. $\mathbf{H_1}$ and $\mathbf{H_2}$ denote the homographies between the 3D plane to its respective projections in the image planes of the first and second cameras, while $\mathbf{H_{12}}$ represents the homography relating the projections of the 3D plane between the two cameras.

correspondences can be established by first detecting keypoints using feature detection algorithms such as SIFT [24] or ORB [25], to later match the features from both images. Once the correspondences are identified, the homography matrix can be estimated by solving a system of linear equations, typically using methods such as Direct Linear Transformation (DLT) or RANSAC.

In this thesis, homographies have been used for opening detection in 3D structural reconstruction and to impose coplanarity constraints on cross-detector visual localization.

2.2.2 Image formation through the pinhole model: from 3D to 2D

In computer vision, the process of projecting 3D points in the world onto a 2D image plane is known as image formation. The pinhole camera model is the simplest yet fundamental representation of how cameras capture images, and is the most widely used model for this task. As illustrated in Figure 2.4, the pinhole model assumes that the camera is represented by a single point (the camera center), and all light rays from the 3D world pass through this point to form an image on a 2D plane located at a fixed distance from the camera center, known as the focal length f.

To project a 3D point $\mathbf{P} = [X, Y, Z]$ in the camera coordinate system onto the 2D image plane at coordinates $[x, y]$, the pinhole model use the principle

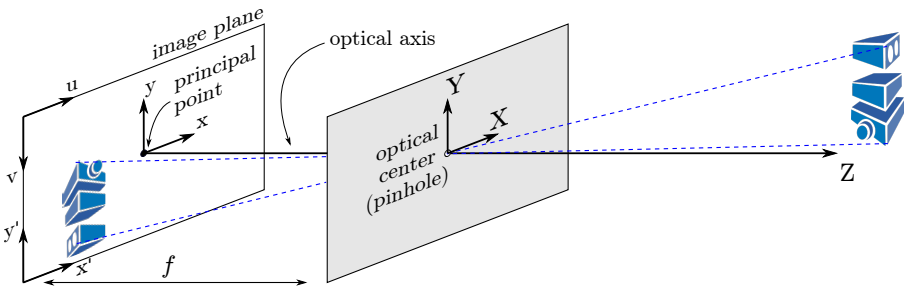


Figure 2.4: Pinhole model for image formation and the associated coordinate systems. Only one light ray per 3D scene point passes through the pinhole and projects onto the image plane.

of perspective projection. That is, the coordinates of the projected point are computed based on the ratios of distances between the point, the camera center, and the image plane. Mathematically, the projection equations are formulated as follows:

$$x = f \cdot \frac{X}{Z}, \tag{2.15}$$

$$y = f \cdot \frac{Y}{Z}, \tag{2.16}$$

where Z represents the distance of the 3D point from the camera along the z-axis, which is also known as the depth. This scaling effect, *i.e.* the perspective projection, explains why objects that are farther from the camera appear smaller in the image, while closer objects appear larger.

In practice, images are typically processed using pixel coordinates, while these equations project points onto 2D but in the original coordinate system units, which are typically meters. The conversion to pixels is performed by introducing the camera's intrinsic parameters into the formulation. These parameters include the focal lengths f_x and f_y in pixel units, as well as the coordinates (c_x, c_y) of the principal point, where the optical axis intersects the image plane. Commonly, these parameters are given through the camera intrinsic matrix $\mathbf{K}$:

$$\mathbf{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{2.17}$$

Knowing the intrinsic parameters, a 3D point $\mathbf{P}$ is projected onto a 2D point $\mathbf{p} = [x', y']$ in the image plane, in pixel units, as follows:

$$x' = f_x \cdot \frac{X}{Z} + c_x, \tag{2.18}$$

$$y' = f_y \cdot \frac{Y}{Z} + c_y, \tag{2.19}$$

To express these coordinates to the left-top corner of the image another tomography, defined by a rotation of $-90°$ and a translation, is required. This results in the following expressions:

$$u = M - y', \tag{2.20}$$

$$v = x', \tag{2.21}$$

where M is the number of rows of the sensor image (*i.e.* image height).

A more compact and mathematically convenient approach to express this projection is through homogeneous coordinates, which represent both 2D and 3D points as 3D and 4D vectors, respectively. In homogeneous coordinates, a 3D point $\mathbf{P} = [X, Y, Z]$ is expressed as $\widetilde{\mathbf{P}} = [X, Y, Z, 1]$, while a 2D pixel with coordinates $[u, v]$ becomes $[u, v, 1]$. This additional dimension enables matrix multiplications that encapsulate both the projection and scaling operations. Then, the projection from 3D to 2D using homogeneous coordinates can be expressed as the following matrix multiplication:

$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & -1 & M \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\mathbf{H}} \underbrace{\begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}}_{\mathbf{K}} \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}}_{\mathbf{P_0}} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}, \tag{2.22}$$

where $\mathbf{P_0}$ is the normalized perspective projection matrix. Here, λ is a scalar factor used to account for the projective nature of homogeneous coordinates.

In this thesis, image formation is used for occlusion estimation as well as to identify elements that were previously mapped but are no longer present in the scene.

2.2.3 Scene reconstruction through the pinhole model: from 2D to 3D

The inverse problem of image formation involves retrieving a 3D point in the world from its coordinates in the image. Applying the pinhole model, this transformation is under-constrained, that is, a single point in the 2D image plane corresponds to a ray in the 3D space. Thus, to recover the original 3D coordinates, additional information such as the depth or prior knowledge about the scene geometry is required.

Let's formulate mathematically this problem. Considering a point in the image plane with pixel coordinates $[u, v]$, its corresponding 3D point $\mathbf{P} = [X, Y, Z]$ in the 3D space can be computed by reversing the perspective

projection presented in Eqs. (2.18) and (2.19) and considering the transformation presented in Eqs. (2.20) and (2.21) as follows:

$$X = Z \cdot \frac{v - c_x}{f_x}, \tag{2.23}$$

$$Y = Z \cdot \frac{M - u - c_y}{f_y}, \tag{2.24}$$

$$Z = Z, \tag{2.25}$$

where c_x, c_y, f_x and f_y are the intrinsic parameters of the camera. In matrix form, it can be expressed as:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = Z \cdot \underbrace{\begin{bmatrix} \frac{1}{f_x} & 0 & -\frac{c_x}{f_x} \\ 0 & \frac{1}{f_y} & -\frac{c_y}{f_y} \\ 0 & 0 & 1 \end{bmatrix}}_{\mathbf{K^{-1}}} \underbrace{\begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & M \\ 0 & 0 & 1 \end{bmatrix}}_{\mathbf{H^{-1}}} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}. \tag{2.26}$$

In Eq. (2.26), the value of Z is unknown, resulting in a direction vector in 3D space instead of a 3D point. To resolve the actual 3D coordinates, the depth Z must be obtained or estimated. This depth can be acquired through direct measurements from depth sensors, such as RGB-D cameras, or estimated using techniques such as stereo vision, structure from motion (SfM), or a learning-based method for monocular depth estimation, among other approaches.

Concretely, in this thesis, scene reconstruction is crucial, as it is the basis to create 3D representations of the environment from visual cues, such as representing objects, structural elements, etc. Additionally, it is used for the creation of the 3D feature maps used for Cross-Detector Visual Localization.

2.3 Scene understanding concepts

2.3.1 Identifying objects from images

The identification of objects from images is a fundamental aspect of computer vision, essential for enabling robots to effectively perceive and interact with their surroundings. The excellent performance of deep learning models in this area has established neural network-based approaches as the de facto solution, providing superior accuracy and robustness compared to traditional methods. Depending on the desired outcome, this task can be addressed from four different points of view (see Figure 2.5):

- **Object detection [27].** This approach involves identifying and localizing objects within an image by predicting bounding boxes and

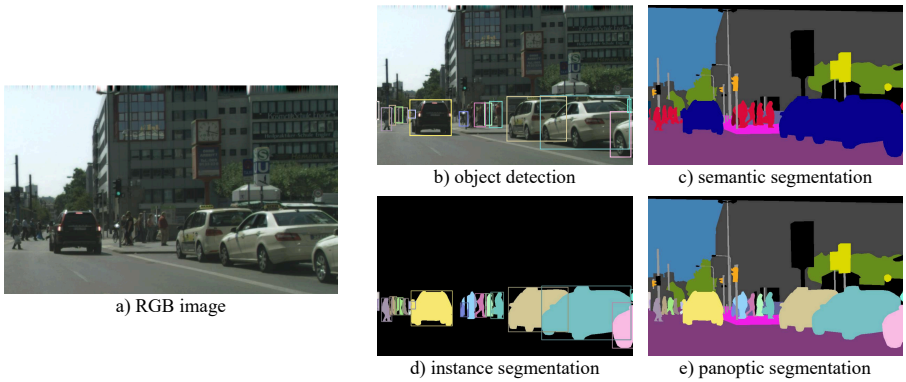


Figure 2.5: Given an input RGB image, we illustrate the four different approaches for object identification in images: b) object detection (per-object bounding boxes and class labels), c) semantic segmentation (per-pixel class labels), d) instance segmentation (per-object masks and class labels) and e) panoptic segmentation (per-pixel class and instance labels). Images obtained from [26].

assigning corresponding class labels to each detected object (see Figure 2.5b). While efficient and effective for tasks such as object tracking and real-time applications, object detection does not capture detailed information about the shape or precise boundaries of objects, as it relies on bounding boxes rather than per-pixel accuracy.

- **Semantic segmentation [28].** Methods within this approach typically assigns a class label to each pixel in an image, as shown in Figure 2.5c. However, semantic segmentation does not differentiate between individual instances of objects within the same category, *i.e.* all objects of the same class are uniformly labeled without distinguishing between separate occurrences. This is the case, for example, of the four cars present in the image, which are all labeled just as `car` in blue color. This limitation narrows its applicability to tasks that need instance-level information, requiring post-processing to segment individual instances within the same class.

- **Instance segmentation [29].** Building on top of the capabilities of object detection, instance segmentation approaches not only identify and localize objects, but also provide pixel-level masks for each individual object instance, as can be seen in Figure 2.5d. Unlike semantic segmentation, which uniformly labels all pixels of a given object category, instance segmentation distinguishes and segments each detected object separately. However, only the pixels corresponding to recognized objects are segmented, leaving non-object areas unlabeled.

- **Panoptic segmentation [26].** Combining semantic and instance segmentation, this approach provides a comprehensive understanding of the scene by assigning a class label to every pixel while also distinguishing between individual object instances of the same category, as shown in Figure 2.5e. It provides a unified framework that labels pixels according to object instances and background elements, ensuring a complete scene representation. This approach effectively mitigates the limitations of semantic segmentation and instance segmentation by offering instance-level information and overall scene context at once, making it appropriate for applications that require complete scene understanding.

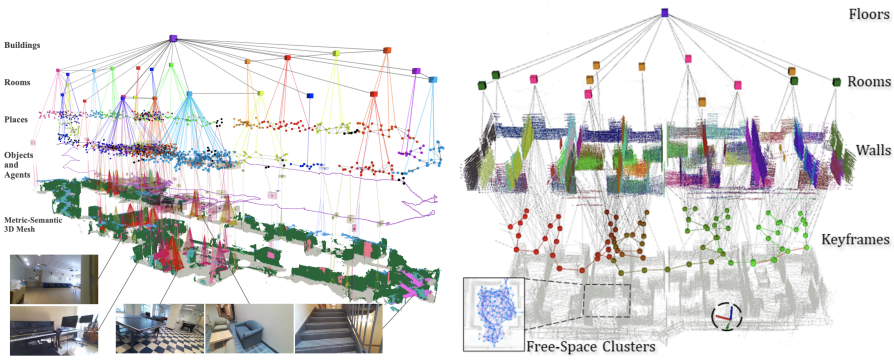
In this thesis, the three first approaches are employed. Specifically, the knowledge of the camera motion is leveraged to improve object detection methods applied to video sequences in robotics, thereby incorporating spatio-temporal coherence into their outputs. Furthermore, instance semantic segmentation networks are used to build object-oriented semantic maps, while panoptic segmentation networks are considered for 3D floor plan reconstruction.

2.3.2 Semantic maps

In robotics, semantic maps are internal representations created and maintained by robots that encapsulate meaningful information about their workspace. These maps extend the geometric and topological information with semantic knowledge, which allows robots to understand the environment at a higher level. The latter includes identifying objects, recognizing places and inferring relationships between entities within the environment, among others.

Incorporating such semantic understanding, robots are provided with the capability to interpret not only the spatial aspects of their surroundings, but also to understand the meaning of the entities (objects, structural elements, rooms, etc.) present in the environment and their possible interactions with humans or even other entities (functionalities, behaviors, contextual relations, etc.).

Following the proposal from Galindo *et al.* [30], semantic maps are formalized and represented as a twofold hierarchical representation: the **spatial hierarchy** that captures the geometry of the environment in its lowest level, and upon it builds inter- and intra-connected topological layers such as `Objects`, `Rooms` and `Buildings`, and the **conceptual hierarchy** (also known as *terminological hierarchy* [3]), which codifies the semantic knowledge through a taxonomy of concepts, defining potential entities and spaces that could be found in an environment of a certain type (in a home, some examples are a `Television is-a Object` or a `Bedroom is-a Room`), along with their attributes and relations.



(a) 3D scene graph from Hydra. (b) 3D scene graph from S-Graphs+.

Figure 2.6: Global caption for both images

2.3.2.1 The spatial hierarchy

Most contributions in semantic mapping predominantly focus on the automatic construction of one or more layers within the spatial hierarchy. For instance, place categorization methods [1, 13, 31] aim to classify and annotate different spaces within an environment (*e.g.* `place-1 is-a Kitchen`). Research on object-oriented semantic mapping [4, 5, 16] emphasizes the construction of the `Objects` layer, where objects are reconstructed using different representation models (*e.g.* 3D points, voxels or 3D bounding boxes) and annotated with their respective categories. Other research directions include the development of the `Structure` layer, which involves reconstructing the environment's floor plan [14, 32, 33], and the `Agents` layer, responsible for mapping movable entities such as humans, or vehicles in outdoor scenarios [6, 7]. More recently, works such as Hydra [7] and S-Graphs+ [34] have been exploring the use of 3D scene graphs to facilitate the construction of multiple layers or even the complete spatial hierarchy (see Figure 2.6).

Note that, while specific layer names such as `Objects`, `Structure`, and `Agents` are used in this section to refer to different layers of the spatial hierarchy, there is no standard definition, and identical layers may be referred to by different names in other works. In particular to this thesis, we have contributed different approaches for the automatic construction of the layers `Places`, `Objects` and `Structure`, separately.

2.3.2.2 The conceptual hierarchy

The grounding of spatial information to concepts gives their semantics, which is crucial for robots to achieve an effective understanding of the environment

beyond spatial properties. These concepts (*e.g.* `Couch is-a Object` or `Antonio is-a Agent`), along with their attributes and contextual relations, enable robots to interpret the purpose, interactions, and roles of entities within the workspace. This knowledge shapes the conceptual hierarchy, commonly defined as an ontology, which is constituted by a hierarchy of concepts that are organized according to a subsumption ordering [35]. Traditionally, this hierarchy is constructed through human elicitation, where experts design taxonomies by embedding domain-specific definitions of objects, spaces, and their relationships [36]. However, this process, while accurate, introduces challenges. Expert knowledge may be difficult to access, the process could become time-consuming, and the resulting taxonomy may be limited by the static nature of predefined entities. To overcome these limitations, approaches such as online search and web mining have been explored to automate the acquisition of semantic knowledge from external data sources [37]. However, they still require substantial human supervision to ensure accuracy and relevance. Recently, approaches leveraging Large Language Models (LLMs), such as ConceptGraphs [38], have shown promising results in this task, though there is still significant room for improvement.

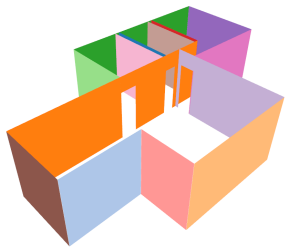
2.3.3 Geometric representation models

The choice of the geometric primitives to shape scene elements from the different layers of the spatial hierarchy (*e.g.* objects, structural elements, etc.) is a critical aspect, as it determines the effectiveness and applicability of the resulting maps. The following are some of the more widely adopted representation models:

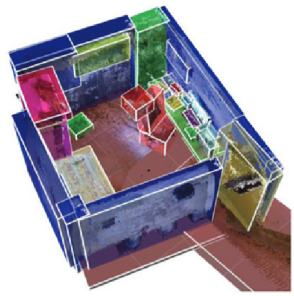
- **Point Clouds (Figure 2.7a)** are collections of 3D points, often annotated with additional attributes such as color or semantic labels. This representation is particularly convenient since it is commonly produced by sensors such as RGB-D cameras and 3D LiDARs. However, the lack of explicit connectivity between points restricts their applicability in tasks that require continuous surface information. In addition, the inherent nature of point clouds makes their storage and processing resource-intensive, particularly in large-scale environments, where their manipulation becomes challenging.

- **Planes (Figure 2.7b)** are geometric primitives that represent flat, continuous surfaces in 3D space. This primitive is of great relevance in scenes featuring planar areas, offering a highly compact representation that is well-suited for large-scale environments. For example, planes are particularly useful for modeling structural elements such as walls, floors and ceilings.

(a) Point Clouds. Image obtained from [39].

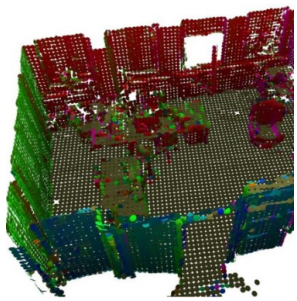


(b) Planes.

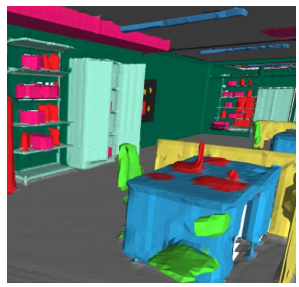


(c) 3D Bounding Boxes. Image obtained from [39].

(d) Voxels. Image obtained from [40].



(e) Surfels. Image obtained from [41].



(f) Meshes. Image obtained from [42].

Figure 2.7: Global caption for all images

- **3D Bounding Boxes (Figure 2.7c)** are a lightweight and efficient primitive to represent the extent, position and orientation of scene elements in the 3D space. This compact representation simplifies data association and integration is simplified, which makes it particularly useful for resource-constrained devices. However, the efficiency of 3D bounding boxes comes at a price: they lack to capture information about the shape of the objects, limiting their applicability in tasks like object manipulation.

- **Voxels (Figure 2.7d)** represent 3D space as a regular grid of cubic cells, each annotated with additional information such as occupancy, color or semantics. This representation model encodes volumetric data, which is of great value for tasks requiring detailed geometry information. For instance, in object-oriented semantic mapping, voxels facilitate the effective reconstruction of objects. There is a trade-off

between resolution and scene size: smaller voxels produce detailed reconstruction but increase computational and storage demands, while larger voxels reduce resource requirements but may compromise the accuracy of the reconstruction.

- **Surfels (Figure 2.7e)** represent 3D surfaces using a collection of oriented disk-like elements, each annotated with attributes such as color and surface normal. This model excels in detailed surface reconstruction by providing a compact and efficient representation of local surface features. However, surfels may face limitations in environments with complex geometries, such as curved surfaces, where high-density sampling is necessary to maintain accuracy.

- **Meshes (Figure 2.7f)** represent 3D surfaces through a network of interconnected vertices, edges and faces, which define the geometry of scene elements with significant accuracy. This representation captures detailed and continuous surface structures, which facilitates accurate spatial modeling and understanding. However, meshes can be computationally demanding, both in terms of processing and storage, posing a challenge for real-time or online applications.

In this thesis, four of the six aforementioned representation models are used: i) point clouds, commonly produced by RGB-D cameras and annotated with semantic information to later use as input for different applications, ii) planes, selected as the geometric primitive for 3D floor plan reconstruction, iii) 3D bounding boxes, chosen for an efficient approach to object-oriented semantic mapping, and iv) voxels, employed as an alternative to achieve higher reconstruction accuracy in object-oriented semantic mapping.

3 Place categorization

For the next chapters, let's build an analogy to conceptualize robotic semantic mapping as a house. In this analogy, the floor of the house represents place categorization, serving as the foundational layer upon which everything else is built. Just as the floor dictates the usability of the space, categorizing different areas –such as living rooms, kitchens, and bedrooms– provides the essential context for the robot's navigation and interaction. In this chapter, we introduce an attention mechanism to improve the performance of existing place categorization methods in robotics scenarios.

3.1 Introduction

Place categorization involves assigning a semantic label to distinct areas in the environment, such as rooms or user-defined spaces, indicating their category (*e.g.* kitchen, living room, etc.) [1]. This categorization is crucial for autonomous robots performing high-level tasks, as it enables them to interpret their workspace and extend their knowledge by connecting this understanding with other semantic information, such as the objects present in the scene, leading to the creation of contextual relations [3, 43]. The latter, for example, allows the robot to interpret human commands with implicit information. When instructed to *place this glass in the dishwasher*, the robot

can infer that dishwashers are typically found in kitchens, allowing it to act correctly without needing a fully explicit command.

One of the two main approaches to place categorization is object-based methods [9, 10, 31, 44], which infer the category of a certain place based on the set of detected objects located within that area. These methods leverage semantic relationships (*e.g.* a toilet is typically found in a bathroom, while a microwave is usually located in a kitchen) that are often encoded in an ontology [35]. The second approach is image-based [1, 11, 12, 45], where the category of a place is directly derived from a given image, primarily by applying deep learning models.

Regardless of the selected approach, a significant limitation comes from the non-optimal observation of the scene. These methods usually process images captured by the robot as it navigates the environment, but often do not take into account the point-of-view from which these images are taken. As a result, numerous images are from walls or empty spaces that provide few or no information about the place category, leading to poor performance. The scope of this chapter is to propose an attention mechanism that can be integrated with any of these methods to improve their efficiency by selecting, at each time moment, the line-of-sight of the camera that maximizes the information gain.

3.2 Contributions

This chapter contributes to improving the efficiency of place categorization algorithms in robotic applications. In particular, our work [13] introduces an attention mechanism based on active vision principles, which selects the line-of-sight of the camera at each time moment using a pan-only unit, aiming to maximize the information gain. This mechanism is formulated as a next-best-view problem, mathematically modelled through a Markov Decision Process (MDP). Our approach is applicable to both object-based and image-based place categorization methods. Experimental results in both scenarios demonstrate its effectiveness, outperforming standard camera configurations, increasing accuracy and reducing the time required for correct categorization.

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

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

Abstract

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

Published in
Knowledge-Based Systems
2022
DOI: 10.1016/j.knosys.2021.108022

4

3D reconstruction of structural elements

Building on our analogy, the structural skeleton of the house –including walls, doors and windows– defines the spatial boundaries and support of the environment. In semantic mapping, these structural elements provide the necessary context for understanding the physical layout and connectivity of spaces. In this chapter, we present a method for the automatic reconstruction of this structural model, commonly referred to in the literature as a 3D floor plan.

4.1 Introduction

Despite the fact that much of the focus in semantic mapping literature tends to concentrate on movable entities (*e.g.* objects) in the scene [2, 4, 5, 16], it is essential to highlight the importance of structural elements in semantic maps. Building a model of such elements (*e.g.* walls, doors, staircases, etc.) plays a crucial role in defining the spatial layout of the environment, which is of great value for effective robot navigation [46]. Furthermore, the representation of these elements yield the robot of a comprehensive understanding of the different spaces in the scene, providing context and boundaries that guide robot operation. Also, the relation between movable entities and structural elements can contribute to refine the semantic map, for example, by enhancing the alignment of objects in the map [47].

The construction of such structural models is commonly known as 3D floor plan reconstruction, and typically involves a robot equipped with a RGB-D camera capturing scene information while exploring the environment. However, most of existing methods face some of the following limitations that hinder their effectiveness in real-world scenarios:

- **Large amount of input data required.** It is common that a point cloud of the complete environment is required as input of the reconstruction process [48, 49]. Besides, this point cloud needs to be pre-processed, ensuring a correct alignment of the whole point cloud. However, mobile robots face several limitations in this regard. First, acquiring a detailed point cloud of the entire environment can be challenging due to the limited sensor range and accuracy inherent in mobile platforms. In addition, real-time processing power and memory constraints on mobile robots often limit the ability to efficiently handle large-scale point clouds. Furthermore, ensuring accurate alignment of point cloud data is complicated by sensor noise, drift in localization systems, and the presence of dynamic obstacles in the environment, all of which can lead to misalignment or incomplete reconstructions.
- **Neglect of localization uncertainty.** Most current approaches ignore the unavoidable uncertainties in robot localization, relying directly on a pre-processed point cloud [50, 51]. However, in real-world applications, handling this uncertainty is crucial for correct reconstruction.
- **Over-simplification of the problem.** Often, methods assume a Manhattan world, simplifying the environment by assuming all structural elements are orthogonal, which is not always realistic for real-world scenarios [52, 53]. Additionally, some works disregard wall thicknesses, leading to inaccuracies in the reconstruction.

4.2 Contributions

To the problem of 3D reconstruction of structural elements, the topic of this chapter, we contribute with our work [14], in which we present *Sigma-FP*, a method for the automatic 3D reconstruction of floor plans from RGB-D sequences. Concretely, the reconstruction is performed online in a frame-by-frame manner, approximating structural elements (*i.e.* walls) with 3D planes –a compact representation–, thus removing the need to store large amounts of data. Operating online requires handling of inherent uncertainties coming from both robot localization and plane estimation. To address this, *Sigma-FP* is formulated probabilistically, allowing for the weighting of

incoming observations based on their reliability. Furthermore, *Sigma-FP* relaxes the common Manhattan world assumption, adopting the less restrictive Atlanta world model, which only requires structural elements to be vertical. Last but not least, our proposal provides a more realistic reconstruction, taking into account wall thicknesses and including openings such as doors and windows. Comparative evaluations against state-of-the-art approaches demonstrate the effectiveness of *Sigma-FP* in real-world scenarios, where existing methods struggle due to rigid geometric assumptions and inadequate handling of uncertainty.

4.A Sigma-FP: Robot mapping of 3D floor plans with an RGB-D camera under uncertainty

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

Abstract

This work presents Sigma-FP, a novel 3D reconstruction method to obtain the floor plan of a multi-room environment from a sequence of RGB-D images captured by a wheeled mobile robot. For each input image, the planar patches of visible walls are extracted and subsequently characterized by a multivariate Gaussian distribution in the convenient Plane Parameter Space. Then, accounting for the probabilistic nature of the robot localization, we transform and combine the planar patches from the camera frame into a 3D global model, where the planar patches include both the plane estimation uncertainty and the propagation of the robot pose uncertainty. Additionally, processing depth data, we detect openings (doors and windows) in the wall, which are also incorporated in the 3D global model to provide a more realistic representation. Experimental results, in both real-world and synthetic environments, demonstrate that our method outperforms state-of-the-art methods, both in time and accuracy, while just relying on Atlanta world assumption.

Published in
IEEE Robotics and Automation Letters (RA-L)
2022
DOI: 10.1109/LRA.2022.3220156

5 Object-oriented semantic mapping

*Finally, the inner part of our house analogy symbolizes the movable entities (*i.e. *objects) that provide the space with functionality. These objects play a crucial role in a robot's interaction with its environment, enabling it to perform tasks and provide effective assistance to users. In this chapter, we propose a method to improve video object detection in robotics, and later, two approaches for the building of such object-oriented semantic maps.*

5.1 Introduction

For robots to effectively perform high-level tasks in human-centered environments, they must have an accurate understanding of the scene. Part of this understanding comes from recognizing the different objects in their workspace with which robots may interact. The recognition of these objects is usually achieved by using object detection methods, which allow the building of object-oriented semantic maps. In such maps, objects are not only reconstructed, but also linked to relevant information (*e.g.* their semantics), including categories, functionalities and relationships, among other details. An example of the latter could be that refrigerators are appliances, typically found in kitchens, used to store food and drinks at cool temperatures, and are controlled by an internal thermostat. Ideally, if an object-oriented semantic map is properly built, it could enable the robot to

recognize that if a user says *"what a scorching hot day!"*, it could respond by opening the refrigerator and offering a cold drink. However, it should be noted that the exploitation of such maps is out of the scope of this thesis.

Although object-oriented semantic mapping has achieved substantial advances in recent years [2, 4, 5], there are still some limitations that have yet to be addressed. An important concern lies in the object detection stage, which is typically performed through neural networks, the de facto solution for this task. While neural networks offer excellent accuracy in object detection, their performance degrades in robotic applications [54]. The latter is principally attributable to the low quality of images captured by the robot's on-board cameras, which typically are of lower resolution and with a limited field of view. In addition, these images are commonly captured while the robot is exploring the environment, thus the resulting images are prone to motion blur, occlusions and partial views, among other unfavorable conditions. All these factors reduce the performance of object detection approaches, leading to errors such as false negatives, misclassifications and over-segmentation [4, 54]. These errors are propagated to the semantic map, leading to incorrect or duplicated objects instances, which compromises the usage of the map. Another critical limitation of existing object-oriented semantic mapping methods is the assumption of a static environment [2], in which objects are added to the map but never removed. Tackling this limitation is critical to improve the long-term usefulness of these maps, since real-world environments are dynamic and changes in the position of objects must be reflected in the semantic map.

5.2 Contributions

This chapter first presents a method that focuses on improving video object detection performance in robotics. Next, two different approaches for building instance-aware semantic maps are introduced: the first employs 3D bounding boxes as primitives for efficiency, while the second leverages voxel-based representation to enhance reconstruction accuracy.

Our first contribution [15] proposes a method that exploits camera motion information to improve object detection in video sequences. Building on a two-stage detection framework with a Region Proposal Network (RPN) and an Object Classifier Network (OCN), our approach introduces a motion-guided propagation model to ensure temporal and spatial consistency. By leveraging planar homography from camera motion, object observations from previous frames are propagated and matched with current RPN proposals. Unmatched observations are reintroduced as new regions of interest. A recursive Bayesian filter further refines temporal coherence. The

experiments show that our method significantly boosts recall with minimal impact on precision and computation time.

Moving to the instance-aware semantic mapping contributions, the first work [16], namely *LTC-Mapping*, focus on addressing key challenges that limit long-term applicability of semantic maps, such as instance duplication and dynamic objects. The method uses 3D bounding boxes as geometric primitives and annotates their corners with visibility flags, allowing for the estimation of unseen object areas. The latter is of great value to mitigate instance duplication by identifying map instances that comes from partial views of the same physical object. On the other hand, to tackle the problem of dynamic objects, the method accounts for both object detections in the images and non-detections of previously mapped objects, enabling the identification and removal of these objects that have been relocated or removed from the scene. The effectiveness of the proposed mechanisms is validated through a set of experiments, demonstrating improvements over a state-of-the-art method.

The latest contribution in this field is *Voxeland* [17], a probabilistic framework for constructing instance-aware semantic maps that builds on the concepts of Dempster-Shafer Evidence Theory (DST). Specifically, each neural network prediction –both mask and category– is treated as a subjective opinion, which accumulates over time to create evidence. Our approach processes two types of evidence: i) geometric, where voxels are used as primitives and opinions are aggregated to update beliefs about the membership of object instances, and ii) semantic, where object instances are updated with incoming opinions about their category. Integrating these potentially conflicting evidences into a probabilistic framework allows quantifying uncertainty, which increases the robustness of semantic maps. By leveraging uncertainty at the geometric level, we identify voxels with an ambiguous instance belonging due to contradictory observations, while semantic uncertainty allows us to reclassify objects with insufficient or out-of-distribution data. The evaluation against state-of-the-art methods highlights the importance of accounting for uncertainty to improve reconstruction accuracy.

5.A Exploiting spatio-temporal coherence for video object detection in robotics

David Fernandez-Chaves, Jose-Luis Matez-Bandera, Jose-Raul Ruiz-Sarmiento, Javier Monroy, Nicolai Petkov and Javier Gonzalez-Jimenez

Abstract

This paper proposes a method to enhance video object detection for indoor environments in robotics. Concretely, it exploits knowledge about the camera motion between frames to propagate previously detected objects to successive frames. The proposal is rooted in the concepts of planar homography to propose regions of interest where to find objects, and recursive Bayesian filtering to integrate observations over time. The proposal is evaluated on six virtual, indoor environments, accounting for the detection of nine object classes over a total of ∼7k frames. Results show that our proposal improves the recall and the F1-score by a factor of 1.41 and 1.27, respectively, as well as it achieves a significant reduction of the object categorization entropy (58.8%) when compared to a two-stage video object detection method used as baseline, at the cost of small time overheads (120ms) and precision loss (0.92).

Published in the proceedings of the
Computer Analysis of Images and Patterns (CAIP 2021)
Nicosia, Cyprus (Virtual Event), 2021
DOI: 10.1007/978-3-030-89131-2_17

5.B LTC-Mapping, enhancing long-term consistency of object-oriented semantic maps in robotics

Jose-Luis Matez-Bandera, David Fernandez-Chaves, Jose-Raul Ruiz-Sarmiento, Javier Monroy, Nicolai Petkov and Javier Gonzalez-Jimenez

Abstract

This paper proposes LTC-Mapping, a method for building object-oriented semantic maps that remain consistent in the long-term operation of mobile robots. Among the different challenges that compromise this aim, LTC-Mapping focuses on two of the more relevant ones: preventing instance duplication of objects and handling dynamic scenes. The former refers to creating multiple instances of the same physical object in the map, usually as a consequence of partial views or occlusions. The latter deals with the typical assumption made by object-oriented mapping methods that the world is static, resulting in outdated representations when the objects change their positions. To face these issues, we model the detected objects with 3D bounding boxes, and analyze the visibility of their vertices to detect occlusions and partial views. Besides this geometric modeling, the boxes are augmented with semantic information regarding the categories of the objects they represent. Both the geometric entities (bounding boxes) and their semantic content are propagated over time through data association and a fusion technique. In addition, in order to keep the map curated, the non-detection of objects in the areas where they should appear is also considered, proposing a mechanism that removes them from the map once there is evidence that they have been moved (*i.e.* multiple non-detections occur). To validate our proposal, a number of experiments have been carried out using the Robot@VirtualHome ecosystem, comparing its performance with a state-of-the-art alternative. The results report a superior performance of LTC-Mapping when modeling both geometric and semantic information of objects, and also support its online execution.

Published in
MDPI Sensors
2022
DOI: 10.3390/s22145308

5.C Voxeland: Probabilistic instance-aware semantic mapping with evidence-based uncertainty quantification

Jose-Luis Matez-Bandera, Pepe Ojeda, Javier Monroy, Javier Gonzalez-Jimenez and Jose-Raul Ruiz-Sarmiento

Abstract

Robots in human-centered environments require accurate scene understanding to perform high-level tasks effectively. This understanding can be achieved through instance-aware semantic mapping, which involves reconstructing elements at the level of individual instances. Neural networks, the de facto solution for scene understanding, still face limitations such as overconfident incorrect predictions with out-of-distribution objects or generating inaccurate masks. Placing excessive reliance on these predictions makes the reconstruction susceptible to errors, reducing the robustness of the resulting maps and hampering robot operation. In this work, we propose Voxeland, a probabilistic framework for incrementally building instance-aware semantic maps. Inspired by the Theory of Evidence, Voxeland treats neural network predictions as *subjective opinions* regarding map instances at both geometric and semantic levels. These opinions are aggregated over time to form evidences, which are formalized through a probabilistic model. This enables us to quantify uncertainty in the reconstruction process, facilitating the identification of map areas requiring improvement (e.g. reobservation or reclassification). As one strategy to exploit this, we incorporate a Large Vision-Language Model (LVLM) to perform semantic level disambiguation for instances with high uncertainty. Results from the standard benchmarking on the publicly available SceneNN dataset demonstrate that Voxeland outperforms state-of-the-art methods, highlighting the benefits of incorporating and leveraging both instance- and semantic-level uncertainties to enhance reconstruction robustness. This is further validated through qualitative experiments conducted on the real-world ScanNet dataset.

Submitted to an international journal
and under review
2024

6

Cross-detector visual localization for long-term interoperability

*Stepping beyond the analogy of a house used to illustrate the construction of semantic maps, it is essential to acknowledge that creating such maps requires a robot capable of capturing scene information. Yet, not only must the robot acquire such information, but it also needs to determine its accurate pose (*i.e. *robot localization) within the environment to place correctly the observed scene elements. In this chapter, we introduce the challenge of Cross-Detector Visual Localization, which aims to facilitate interoperability among robots and enhance the long-term usability of constructed semantic maps. Finally, we propose a first approach to address this problem that, in fact, takes advantage of the information available in semantic maps.*

6.1 Introduction

Building semantic maps requires knowing the robot localization at each moment to place the different scene elements with respect to a geometric reference map. Ideally, this reference map should be common for the different robots operating in the same environment, enabling the interoperability and coexistence between them (*e.g.* collaborative mapping) [55]. Additionally, it is

essential that this reference map remains valid over the long-term, allowing to reflect the real-world changes into the semantic map, such as the relocation of objects or the removal of elements that are no longer present in the environment [16].

A common approach to determine the robot localization is by leveraging the on-board visual sensors (*i.e.* cameras) of the robot. The latter is known in the literature as Visual Localization (VL) [56], where the pose of a camera is estimated from a query image by establishing correspondences between features extracted in the query image and those available in an existing 3D feature map. This approach heavily relies on the assumption that the features in both, the query and the map, are of the same nature (*e.g.* blobs, corners, etc.) and extracted using the same algorithm.

In this context, the major problem arises when attempting to establish correspondences between features of differing nature (*e.g.* corners and blobs). These features represent different, but close, physical points, introducing a spatial discrepancy of keypoints. Such discrepancies hinder the matching process, especially when relying solely on the comparison of feature descriptors, therefore impeding the establishment of the required correct correspondences. This limitation do not guarantee the interoperability when trying to localize a camera that uses a different detection algorithm, which, for example, can be the case of the coexistence of commercial cameras using proprietary algorithms optimized for specific hardware. This challenge is coined as *Cross-Detector Visual Localization* and has not been addressed in the literature yet, but it is crucial as it could remove the need for creating and maintaining separate maps for each different algorithm.

6.2 Contributions

In this chapter, our contribution [18] introduces and provides an initial solution to the Cross-Detector Visual Localization problem. This problem focuses on enabling the localization of visual sensors using heterogeneous feature detectors within a map constructed using a specific type of feature. The main challenge stems from the spatial discrepancy between heterogeneous keypoints, which hinders traditional feature descriptor-based matching. We first formalize the problem mathematically and perform an analysis to provide insights into the underlying challenges. Then, we propose *CoplaMatch*, a novel approach that leverages the structural planes of the scene to impose coplanarity constraints to address this problem. This reduces the reliance on visual descriptors for matching while guiding correspondences through geometric relations. Although *CoplaMatch* represents a promising

initial solution, there is still significant room for improvement in this direction.

6.A Cross-detector visual localization with coplanarity constraints for indoor environments

Jose-Luis Matez-Bandera, Alberto Jaenal, Clara Gomez, Alejandra C. Hernandez, Javier Monroy, José Araújo and Javier Gonzalez-Jimenez

Abstract

Most Visual Localization (VL) methods are set on the premise that the keypoints in the query image are detected with the same algorithm as those stored in the reference map. This entails a serious limitation as new and better detectors may progressively appear, and we would like to ensure the interoperability and coexistence of cameras with heterogeneous detectors on a single map representation. While rebuilding the map with new detectors might seem a solution, it is often impractical as original images may be unavailable or restricted by data privacy constraints. In this paper, first, we introduce and formalize the problem of Cross-Detector VL, where the inherent spatial discrepancy of matching keypoints that represent different, though close, physical entities, hinders the process of establishing correct correspondences when relying strictly on the similarity of the descriptors for the matching. In our second contribution, we propose to address this problem by relaxing such descriptor similarity and complementing it with a coplanarity constraint, implemented with a 2D homography, between groups of observed and map points. This point correspondence process, called *CoplaMatch*, entails segmenting planar patches both, in the map, which is done offline just once, and in the query image, which adds an extra computational overhead to the VL process, which is demonstrated in our experiments that does not hinder the online applicability. Extensive experiments are conducted in indoor environments to i) analyze the suitability of our proposal to perform Cross-Detector VL and ii) compare its accuracy and computational overhead against two state-of-the-art feature matching approaches. All experiments are carried out on environments from real-world datasets.

Submitted to an international journal
and under review
2024

7 Conclusions and future work

This long but rewarding journey called PhD is approaching its end. This chapter summarizes the conclusions drawn, as well as proposes exciting lines of research open for future adventurers.

7.1 Conclusions

This thesis has focused its efforts on advancing indoor semantic mapping for mobile robotics, with a particular emphasis on facilitating its application to real-world scenarios. To summarize the conclusions of this work, let's continue drawing on the analogy presented previously in this thesis, which resembles our research as the construction of a house in three stages.

The first stage involved building the floor, which, in our context, was represented by the categorization of different places in the environment. Place categorization provides essential context for the robot, allowing it to understand the functionality of each area, such as recognizing that a kitchen is the appropriate place to heat food. While existing place categorization algorithms, both object-based and image-based, perform well under ideal conditions (where all objects are clearly detected and images are highly informative), applying these to robotic scenarios often poses difficulties, such as that the images captured by the robots are prone to contain occlusions, partial views of objects, or empty scenes. To address these limitations, in Chapter 3 we have proposed an attention mechanism to maximize the information available in the robot's images by selecting, at each time

moment, the most informative camera line-of-sight through a pan-only unit. In addition, we have also taken into account the short-term robot navigation path planning to further maximize the information gain along the robot's trajectory. The results obtained demonstrate the importance of selecting the appropriate point-of-view to maximize the accuracy of place categorization methods, while also reducing the required time for correct categorization.

Once the foundations of the house have been laid, it is time to proceed with the construction of the skeleton of the house (*i.e.* the structure formed by the walls, doors, windows, etc.), which in our work we refer to as 3D reconstruction of structural elements. In Chapter 4, we presented a method to address three key limitations encountered in real-world scenarios: i) minimizing the need for storing and handling large amount of data, ii) accounting for inherent uncertainties in the reconstruction process, and iii) adopting realistic assumptions rather than oversimplifying the problem. Our proposal, *Sigma-FP*, operates frame-by-frame, extracting a compact representation of the scene structure in the form of planes. Additionally, *Sigma-FP* was formulated in probabilistic terms, allowing for the incorporation of uncertainties from both the robot localization and plane estimation steps, while propagating these uncertainties to the model. The latter is crucial to weigh the different incoming observations and to determine which parts of the model are more reliable and which require further observation. Finally, we relaxed the conventional Manhattan world assumption to the more realistic Atlanta world, while we also incorporated details such as wall thickness and openings (*i.e.* doors and windows) into our model.

After the house is completely built, the next task is to furnish it so that it comes to life, that is, to place the relevant objects in each room. In our work, this stage corresponds to the object-oriented semantic mapping problem. This is addressed in Chapter 5, in which, first, we proposed a method to provide spatio-temporal consistency to the outputs of traditional two-stage object detection networks applied to videos captured by robots. Later, seeking the complete mapping of the scene objects, we proposed two approaches: *LTC-Mapping* and *Voxeland*. The former focused on solving the main limitations that prevents the usability of semantic maps in the long-term: i) the lack of map maintenance, which fails to capture dynamic changes in the real-world, and ii) the instance duplication problem, where multiple instances of the same physical object are mapped. Additionally, *LTC-Mapping* was designed to prioritize efficiency through the choice of a lightweight primitive such as 3D bounding boxes. In contrast, *Voxeland* was more oriented towards improving the reconstruction accuracy, while also aiming to account for the common deficiencies of object detection networks, which often produce overconfident yet erroneous detections that, when propagated, compromise the reliability of the map. The proposed approaches

have demonstrated significant progress towards enabling appropriate real-world deployment.

Last but not least, throughout this complete process, the worker constructing the house must constantly be aware of its position within the house to accurately place the next brick. In our analogy, this corresponds to robot localization, which is essential to correctly link the acquired knowledge to its corresponding location in the semantic map. Moreover, this localization must be with respect to a common reference frame for all robots, thus enabling their collaborative operation. It should also be applicable in the long-term, ensuring that the map can be kept updated. In Chapter 6, we introduced for the first time in the literature, the *Cross-Detector Visual Localization* problem, aimed to enable the interoperability between robots equipped with visual sensors. Particularly, the goal was to ensure that cameras using different feature detectors can be localized against a common map representation, which contains 3D features of a different nature. Looking to address this problem, as a first alternative we proposed *CoplaMatch*, an approach that leverages information from structural planes to guide the matching of heterogeneous keypoints. The results shown that in this scenario, relying heavily on feature descriptors for the feature matching is not recommended, as they lack the uniqueness typically exploited for standard visual localization. Thus, it becomes necessary to explore alternatives to complement and guide the matching process.

7.2 Future work

Although this thesis represents a successful breakthrough in our area –and in my personal life–, it does not come to an end here. There is still much exciting work ahead as we continue to explore new technologies, which are emerging more and more frequently. Here, we outline some of the open research lines that we find particularly interesting:

All-in-one modular semantic mapping framework

To facilitate the deployment of semantic mapping in real-world scenarios, it is crucial to develop an all-in-one, modular semantic mapping package that integrates multiple levels of information into a single, unified framework. Existing systems like Kimera and Hydra generate 3D scene graphs with essential semantic layers such as `Objects` and `Places`, but they lack full modularity and flexibility. A desirable solution, preferably implemented in ROS2 and/or any other robotic framework widely used by the community, would enable users to selectively map specific types of information based on

their application needs. The package should be fully customizable, allowing users to choose geometric primitives for reconstruction, select neural networks for different perception tasks, and adjust the system for various levels of computational resources. This flexibility would support a wide range of setups, from highly resource-constrained devices optimized for efficiency to advanced configurations utilizing edge computing for more accurate, resource-intensive applications.

Benchmarking standarization

The lack of standardization in the evaluation of semantic mapping approaches is a clear obstacle to progress in this field so far, leading to a wide variety of metrics, each of them adopted by different authors in their work. For example, different metrics are used for the reconstruction of objects according to the type of primitive used. Providing a standardized evaluation framework that transcends the specific types of primitives or data used would be extremely beneficial to better understand the advances in this field. In addition, a unified benchmarking framework could facilitate more meaningful comparisons between different methods, encouraging collaboration and promoting innovation in robotic semantic mapping.

Adoptation of generative AI techniques

Traditional semantic mapping techniques rely on object detection networks trained on specific datasets, limiting them to a closed set of categories (*closed vocabulary*). For example, models trained on COCO, which includes 80 object categories, can detect appliances such as microwaves and ovens, but fail to recognize objects like dishwashers. This limits their applicability in real-world environments, where many objects would remain undetectable. The advent of generative AI, in this case Large Vision-Language Models (LVLMs), offers a solution by enabling object detection in an *open vocabulary* manner, allowing for the detection of a wider range of objects. Similarly, Large Language Models (LLMs) can replace traditional semantic knowledge bases, which are typically built and maintained through *human elicitation*. This process involves the manual encoding of scene elements properties, functionalities, and relationships, and is time-consuming and results in a closed knowledge base that only reflects the expert's input. By leveraging LLMs, this stage can be enhanced, creating an open semantic knowledge base that reduces the need for continuous expert intervention and expands the semantic knowledge. These are just two potential applications, and the integration of generative AI into robotic semantic mapping unlocks countless possibilities for advancing the field.

Real-world semantic maps exploitation

The main goal of building semantic maps is to enable their effective use by mobile robots to perform high-level tasks in human-centered environments. Currently, most research is focused on the construction of such maps rather than their practical exploitation, with some works presenting only small use cases in controlled environments. To advance the field, it is essential to move beyond these limited applications and facilitate the comprehensive exploitation of semantic maps. In the future, robots should be able to use these maps to reason and perform complex decision-making, for example, using semantic maps as a digital twin of their workspace, to assist humans in a wider range of tasks. By encouraging a more holistic approach to semantic map exploitation, we can unlock the full potential of mobile robots to improve people's productivity in everyday situations.

Bibliography

[1] Niko Sünderhauf, Feras Dayoub, Sean McMahon, Ben Talbot, Ruth Schulz, Peter Corke, Gordon Wyeth, Ben Upcroft, and Michael Milford. Place categorization and semantic mapping on a mobile robot. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5729–5736, 2016.

[2] Niko Sünderhauf, Trung T. Pham, Yasir Latif, Michael Milford, and Ian Reid. Meaningful maps with object-oriented semantic mapping. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5079–5085, 2017.

[3] Jose-Raul Ruiz-Sarmiento, Cipriano Galindo, and Javier Gonzalez-Jimenez. Building multiversal semantic maps for mobile robot operation. *Knowledge-Based Systems*, 119:257–272, 2017.

[4] Margarita Grinvald, Fadri Furrer, Tonci Novkovic, Jen Jen Chung, Cesar Cadena, Roland Siegwart, and Juan Nieto. Volumetric instance-aware semantic mapping and 3D object discovery. *IEEE Robotics and Automation Letters*, 4(3):3037–3044, 2019.

[5] Ruben Mascaro, Lucas Teixeira, and Margarita Chli. Volumetric instance-level semantic mapping via multi-view 2D-to-3D label diffusion. *IEEE Robotics and Automation Letters*, 7(2):3531–3538, 2022.

[6] Antoni Rosinol, Andrew Violette, Marcus Abate, Nathan Hughes, Yun Chang, Jingnan Shi, Arjun Gupta, and Luca Carlone. Kimera: From SLAM to spatial perception with 3D dynamic scene graphs. *The International Journal of Robotics Research*, 40(12-14):1510–1546, 2021.

[7] N. Hughes, Y. Chang, and L. Carlone. Hydra: A real-time spatial perception system for 3D scene graph construction and optimization. *Robotics: Science and Systems (RSS)*, 2022.

[8] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[9] J.R. Ruiz-Sarmiento, C. Galindo, and J. Gonzalez-Jimenez. Joint categorization of objects and rooms for mobile robots. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2523–2528, 2015.

[10] Manuel Brucker, Maximilian Durner, Rareş Ambruş, Zoltán Csaba Márton, Axel Wendt, Patric Jensfelt, Kai O. Arras, and Rudolph Triebel. Semantic labeling of indoor environments from 3D RGB maps. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1871–1878, 2018.

[11] Peter Uršič, Rok Mandeljc, Aleš Leonardis, and Matej Kristan. Part-based room categorization for household service robots. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2287–2294, 2016.

[12] Anwesan Pal, Carlos Nieto-Granda, and Henrik I. Christensen. DEDUCE: Diverse scene detection methods in unseen challenging environments. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4198–4204, 2019.

[13] Jose Luis Matez-Bandera, Javier Monroy, and Javier Gonzalez-Jimenez. Efficient semantic place categorization by a robot through active line-of-sight selection. *Knowledge-Based Systems*, 240:108022, 2022.

[14] Jose-Luis Matez-Bandera, Javier Monroy, and Javier Gonzalez-Jimenez. Sigma-FP: Robot mapping of 3D floor plans with an RGB-D camera under uncertainty. *IEEE Robotics and Automation Letters*, 7(4):12539–12546, 2022.

[15] David Fernandez-Chaves, Jose Luis Matez-Bandera, Jose Raul Ruiz-Sarmiento, Javier Monroy, Nicolai Petkov, and Javier Gonzalez-Jimenez. Exploiting spatio-temporal coherence for video object detection in robotics. In *Computer Analysis of Images and Patterns*, pages 186–196, Cham, 2021. Springer International Publishing.

[16] Jose-Luis Matez-Bandera, David Fernandez-Chaves, Jose-Raul Ruiz-Sarmiento, Javier Monroy, Nicolai Petkov, and Javier Gonzalez-Jimenez. LTC-Mapping, enhancing long-term consistency of object-oriented semantic maps in robotics. *Sensors*, 22(14), 2022.

[17] Jose-Luis Matez-Bandera, Pepe Ojeda, Javier Monroy, Javier Gonzalez-Jimenez, and Jose-Raul Ruiz-Sarmiento. Voxeland: Probabilistic instance-aware semantic mapping with evidence-based uncertainty quantification, 2024.

[18] Jose-Luis Matez-Bandera, Alberto Jaenal, Clara Gomez, Alejandra C Hernandez, Javier Monroy, José Araújo, and Javier Gonzalez-Jimenez. Cross-detector visual localization with coplanarity constraints for indoor environments. *Submitted*, 2024.

[19] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition, 1994.

[20] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, 2005.

[21] Christopher M. Bishop. *Pattern Recognition and Machine Learning*, volume 4. Springer-Verlag, Berlin, Heidelberg, 2006.

[22] Qinxuan Sun, Jing Yuan, Xuebo Zhang, and Fengchi Sun. RGB-D SLAM in indoor environments with STING-based plane feature extraction. *IEEE/ASME Transactions on Mechatronics*, 23(3):1071–1082, 2018.

[23] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition, 2004.

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

[25] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *2011 International Conference on Computer Vision*, pages 2564–2571, 2011.

[26] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6392–6401, 2019.

[27] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[28] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5168–5177, 2017.

[29] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[30] C. Galindo, A. Saffiotti, S. Coradeschi, P. Buschka, J.A. Fernandez-Madrigal, and J. Gonzalez. Multi-hierarchical semantic maps for mobile robotics. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2278–2283, 2005.

[31] David Fernandez-Chaves, Jose-Raul Ruiz-Sarmiento, Nicolai Petkov, and Javier Gonzalez-Jimenez. From object detection to room categorization in robotics. In *Proceedings of the 3rd International Conference on Applications of Intelligent Systems*, APPIS 2020, New York, NY, USA, 2020. Association for Computing Machinery.

[32] Hriday Bavle, Jose Luis Sanchez-Lopez, Muhammad Shaheer, Javier Civera, and Holger Voos. Situational graphs for robot navigation in structured indoor environments. *IEEE Robotics and Automation Letters*, 7(4):9107–9114, 2022.

[33] Bolivar Solarte, Yueh-Cheng Liu, Chin-Hsuan Wu, Yi-Hsuan Tsai, and Min Sun. 360-DFPE: Leveraging monocular 360-layouts for direct floor plan estimation. *IEEE Robotics and Automation Letters*, 7(3):6503–6510, 2022.

[34] Hriday Bavle, Jose Luis Sanchez-Lopez, Muhammad Shaheer, Javier Civera, and Holger Voos. S-Graphs+: Real-time localization and mapping leveraging hierarchical representations. *IEEE Robotics and Automation Letters*, 8(8):4927–4934, 2023.

[35] Mike Uschold and Michael Gruninger. Ontologies: principles, methods and applications. *The Knowledge Engineering Review*, 11(2):93–136, 1996.

[36] José-Raúl Ruiz-Sarmiento, Cipriano Galindo, and Javier Gonzalez-Jimenez. Scene object recognition for mobile robots through semantic knowledge and probabilistic graphical models. *Expert Systems with Applications*, 42(22):8805–8816, 2015.

[37] Kai Zhou, Michael Zillich, Hendrik Zender, and Markus Vincze. Web mining driven object locality knowledge acquisition for efficient robot behavior. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3962–3969, 2012.

[38] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull.

ConceptGraphs: Open-vocabulary 3D scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5021–5028, 2024.

[39] J.R. Ruiz-Sarmiento, C. Galindo, and J. Gonzalez-Jimenez. Robot@Home, a robotic dataset for semantic mapping of home environments. *The International Journal of Robotics Research*, 36(2):131–141, 2017.

[40] Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 190–198, 2017.

[41] Jochen Klaess, Joerg Stueckler, and Sven Behnke. Efficient mobile robot navigation using 3D surfel grid maps. In *ROBOTIK 2012; 7th German Conference on Robotics*, pages 1–4, 2012.

[42] Antoni Rosinol, Marcus Abate, Yun Chang, and Luca Carlone. Kimera: an open-source library for real-time metric-semantic localization and mapping. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1689–1696, 2020.

[43] Santosh K. Divvala, Derek Hoiem, James H. Hays, Alexei A. Efros, and Martial Hebert. An empirical study of context in object detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1271–1278, 2009.

[44] Ren C. Luo and Michael Chiou. Hierarchical semantic mapping using convolutional neural networks for intelligent service robotics. *IEEE Access*, 6:61287–61294, 2018.

[45] Kamal M. Othman and Ahmad B. Rad. An indoor room classification system for social robots via integration of CNN and ECOC. *Applied Sciences*, 9(3), 2019.

[46] Yiming Wang, Stuart James, Elisavet Konstantina Stathopoulou, Carlos Beltrán-González, Yoshinori Konishi, and Alessio Del Bue. Autonomous 3-D reconstruction, mapping, and exploration of indoor environments with a robotic arm. *IEEE Robotics and Automation Letters*, 4(4):3340–3347, 2019.

[47] D. Fernandez-Chaves, J.R. Ruiz-Sarmiento, N. Petkov, and J. Gonzalez-Jimenez. ViMantic, a distributed robotic architecture for semantic mapping in indoor environments. *Knowledge-Based Systems*, 232:107440, 2021.

[48] Chen Liu, Jiaye Wu, and Yasutaka Furukawa. FloorNet: A unified framework for floorplan reconstruction from 3D scans. In *Computer Vision – ECCV 2018*, pages 203–219, Cham, 2018. Springer International Publishing.

[49] Ameya Phalak, Vijay Badrinarayanan, and Andrew Rabinovich. Scan2Plan: Efficient floorplan generation from 3D scans of indoor scenes, 2020.

[50] Srivathsan Murali, Pablo Speciale, Martin R. Oswald, and Marc Pollefeys. Indoor Scan2BIM: Building information models of house interiors. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6126–6133, 2017.

[51] Ameya Phalak, Zhao Chen, Darvin Yi, Khushi Gupta, Vijay Badrinarayanan, and Andrew Rabinovich. DeepPerimeter: Indoor boundary estimation from posed monocular sequences, 2019.

[52] Chen-Yu Lee, Vijay Badrinarayanan, Tomasz Malisiewicz, and Andrew Rabinovich. RoomNet: End-to-end room layout estimation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4875–4884, 2017.

[53] Chenggang Yan, Biyao Shao, Hao Zhao, Ruixin Ning, Yongdong Zhang, and Feng Xu. 3D room layout estimation from a single RGB image. *IEEE Transactions on Multimedia*, 22(11):3014–3024, 2020.

[54] David Morilla-Cabello, Lorenzo Mur-Labadia, Ruben Martinez-Cantin, and Eduardo Montijano. Robust fusion for bayesian semantic mapping. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 76–81, 2023.

[55] Mihai Dusmanu, Ondrej Miksik, Johannes L. Schönberger, and Marc Pollefeys. Cross-descriptor visual localization and mapping. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6038–6047, 2021.

[56] Nathan Piasco, Désiré Sidibé, Cédric Demonceaux, and Valérie Gouet-Brunet. A survey on visual-based localization: On the benefit of heterogeneous data. *Pattern Recognition*, 74:90–109, 2018.