Simposio CEA de Robótica, Bioingeniería, Visión Artificial y Automática Marina 2025

LVLMs aplicados al refinamiento de mapas semánticos en robótica móvil

Torremocha, A.*, Ojeda, P., Ruiz-Sarmiento, J. R., Gonzalez-Jimenez, J.

Grupo de Percepción Artificial y Robótica Inteligente (MAPIR), Dept. de Ingeniería de Sistemas y Automática, Instituto Universitario en Ingeniería Mecatrónica y Sistemas Ciberfísicos (IMECH.UMA), Universidad de Málaga, Blvr. Louis Pasteur, 35, 29071 Málaga, España

Resumen

Los mapas semánticos son representaciones del entorno que incluyen información sobre la geometría de la escena y la clasificación en categorías de los objetos presentes. En este trabajo se proponen técnicas para el refinamiento de mapas semánticos mediante la desambiguación de los objetos con alta incertidumbre en su clasificación semántica. Concretamente, nuestra propuesta consiste en la identificación sistemática de aquellas instancias que requieren un proceso adicional de desambiguación, y el empleo de un modelo de visión-lenguaje (*LVLM*) para llevar a cabo este proceso. La implementación de nuestra propuesta se basa en Voxeland, un marco de trabajo que construye dichos mapas siguiendo un enfoque probabilístico, permitiendo así cuantificar la incertidumbre en las clasificaciones de objetos. Las pruebas realizadas sobre el conjunto de datos de SceneNN validan la efectividad del método, mejorando la clasificación de los objetos y reduciendo la incertidumbre de los mapas semánticos.

Palabras clave: Robótica inteligente, Aprendizaje automático, Métodos Bayesianos, Robots móviles autónomos, Construcción de mapas

Application of Large Vision-Language Models for Semantic Map Refinement in Mobile Robotics

Abstract

Semantic maps are representations of the environment that combine geometric information with knowledge of the semantic classes of the objects in the scene. This work presents techniques for the refinement of semantic maps through the disambiguation of objects that present high uncertainty in their semantic classification. Specifically, our proposal consists of the systematic identification of instances that require an additional disambiguation process, and the utilization of a large vision-language model (*LVLM*) to carry out this process. The implementation of this proposal is based on Voxeland, a framework that constructs semantic maps in a probabilistic manner, thus allowing for the quantification of the uncertainty in the classification of the objects. Experimental results on the SceneNN dataset demonstrate the effectiveness of the proposed method, improving the classification of the selected objects and reducing the uncertainty of the semantic maps.

Keywords: Intelligent Robots, Machine Learning, Bayesian methods, Autonomous robotic systems, Map building

1. Introducción

Los robots móviles desplegados en entornos centrados en humanos son utilizados cada vez en más ámbitos, entre otros, la asistencia al hogar, la industria, la sanidad o la agricultura (Bac et al., 2014). Un requisito fundamental para su despliegue efectivo en tareas de alto nivel es que posean capacidades cognitivas avanzadas para interpretar su entorno y razonar sobre él.

Un enfoque comúnmente aplicado para alcanzar dicha comprensión es la construcción de mapas semánticos (Han et al., 2021), es decir, modelos del entorno de trabajo donde, además, de la habitual reconstrucción geométrica de los elementos que lo componen, también se integra información semántica sobre los mismos (propiedades, funcionalidades, relaciones, etc). Por ejemplo, un robot de servicio en un hospital podría utilizar un mapa semántico para gestionar la distribución de medicamentos y suministros médicos. El mapa contendría información sobre la ubicación de cada medicamento, sus propiedades (fecha de caducidad, temperatura de almacenamiento) y sus relaciones (compatibilidad entre medicamentos, restricciones de acceso). Gracias a esta información, el robot podría optimizar su entrega,

*Autor para correspondencia: adriantorremocha@uma.es

garantizar el cumplimiento de protocolos de seguridad y emitir alertas en caso de uso incorrecto de un medicamento.

Sin embargo, la construcción de estos mapas semánticos está inherentemente afectada por incertidumbre, especialmente la proveniente de las técnicas de percepción. Los métodos actuales de Visión por Computador, basados en Aprendizaje Profundo, si bien potentes, no garantizan la corrección en la identificación de objetos y pueden generar predicciones erróneas (detecciones incorrectas, confianza alta en categorías fuera del vocabulario, máscaras imprecisas, etc.) (Chaves et al., 2019; Matez-Bandera et al., 2024). Esta incertidumbre perceptual, si se ignora, se acumula a medida que el robot explora y mapea el entorno secuencialmente a partir de múltiples imágenes. El resultado es un mapa semántico final con una fiabilidad mermada, lo que puede comprometer la seguridad y eficacia del robot, llevándolo a comportamientos erráticos (Matez-Bandera et al., 2022). Por tanto, es crucial gestionar explícitamente la incertidumbre durante la construcción y uso de mapas semánticos. Este contexto motiva el desarrollo de marcos de trabajo como Voxeland (Matez-Bandera et al., 2024), que construye mapas semánticos probabilísticos capaces de cuantificar explícitamente la incertidumbre tanto geométrica como semántica (ver Figura 1). Voxeland representa la información semántica como "opiniones" probabilísticas sobre las categorías de los objetos.

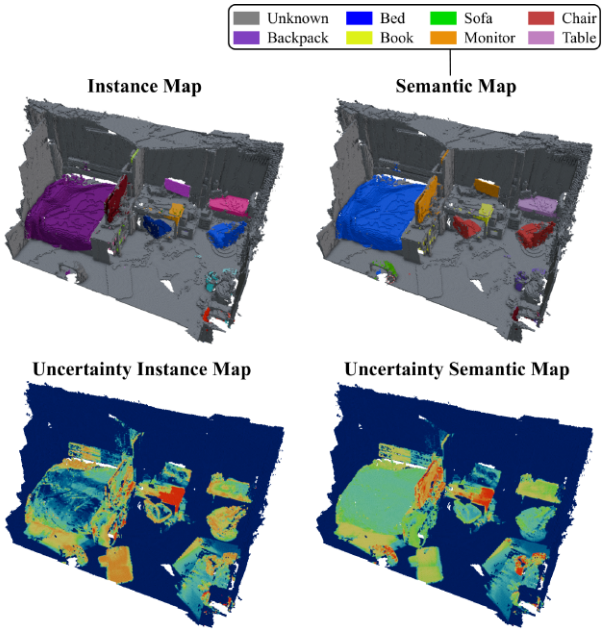

Figura 1: Representación de los cuatro tipos de mapas generados por Voxeland, extraída del artículo original (Matez-Bandera et al., 2024). Este trabajo tiene como propósito reducir la ambigüedad del mapa de incertidumbre semántica

Partiendo de esta base, este trabajo presenta un método para analizar y reducir activamente la incertidumbre semántica en mapas construidos con Voxeland, con el objetivo de producir representaciones del entorno más confiables. El método propuesto incluye:

- **Identificación de incertidumbre**, mediante técnicas como el cálculo de la entropía sobre las opiniones semánticas para localizar las áreas o instancias de objetos más ambiguas en el mapa
- **Desambiguación**, con el uso de herramientas externas, como Modelos de Lenguaje y Visión a Gran Escala (LVLMs), para obtener una "segunda opinión" sobre las instancias identificadas como inciertas.
- **Integración** de la información obtenida de la desambiguación como nuevas opiniones ponderadas dentro del marco probabilístico de Voxeland, refinando así el mapa y reduciendo la incertidumbre local.

Para validar la propuesta, se presentan experimentos empleando un conjunto de datos del estado del arte ampliamente extendido, SceneNN (Hua et al., 2016). Para ello se analiza el mapa semántico que Voxeland crea a partir de una de las escenas de dicho repositorio, se identifican las instancias con mayor incertidumbre semántica y se aplica el proceso de desambiguación para cada una de ellas, obteniendo unas categorías finales válidas y con mayor certidumbre.

2. Contexto y trabajos relacionados

En materia de mapas semánticos, existen numerosos artículos que se enfrentan a las dificultades en la compleja creación de este tipo de representaciones (Ruiz-Sarmiento et al., 2017; Kostavelis and Gasteratos, 2015). Quizás uno de los trabajos recientes más populares sea el presentado por Gu et al. (2023), donde se introduce el concepto de *ConceptGraphs*. Un ConceptGraph es una representación de una escena mediante un grafo en el cual los nodos representan los objetos y los arcos representan las relaciones geométricas entre ellos. Aunque destaca por su planteamiento claro y sencillo, en el articulo se obvian las implicaciones de la incertidumbre en la creación del grafo.

Como se ha mencionado con anterioridad, este trabajo se enmarca en el contexto del marco de trabajo probabilístico Voxeland, diseñado para construir, de manera incremental, mapas semánticos que tengan en cuenta las instancias de objetos. El enfoque probabilístico es dual, ya que se considera tanto para determinar la posición como la categoría semántica de los objetos. Principalmente, se basa en la Teoría de la Evidencia de Jsang (2018), Concretamente, a partir de una secuencia de imágenes del entorno, es capaz de generar 4 mapas: mapa de las instancias de objetos, mapa semántico con las categorías de estos objetos y los correspondientes mapas de incertidumbre (recuérdese la Figura 1).

Para la construcción del mapa semántico con Voxeland es necesario el análisis de imágenes RGB-D, concretamente es vital el uso de algunas técnicas de Visión por Computador, actualmente basadas en Aprendizaje Profundo, que permitan la detección y clasificación de los elementos del entorno (*p.ej.,* YOLO (Redmon et al., 2016) y Mask R-CNN (He et al., 2017)) a partir de dichas imágenes. Son el uso de estas técnicas las que fundamentan este trabajo, pues dado que el margen de tiempo operativo es muy reducido, a menudo, las detecciones pueden ser inexactas (producción de máscaras erróneas o sobredimensionadas) y las clasificaciones erróneas (categoría incorrecta), provocando incertidumbre en el mapa.

Para la desambiguación semántica se hace uso de LVLMs "multimodales" de propósito general. Algunos de los más utilizados en este contexto son LLaVA (Liu et al., 2023) y , mas recientemente, MiniCPM (Yao et al., 2024). Ambos ofrecen muy

buenos resultados, similares a otros modelos propietarios como ChatGPT4o.

3. Descripción del método

El flujo de trabajo del método propuesto puede observarse en la Figura 2. El punto de partida es la obtención de la información relacionada con la semántica proporcionada por Voxeland (ver columna derecha de la Figura 1). Esto incluye la información relativa a las instancias de objetos detectadas (distribución de probabilidad sobre las categorías de objeto y número total de observaciones), así como una lista por objeto de asociaciones categoría-imagen. Esta lista asocia a cada categoría el conjunto de imágenes en las que el objeto aparece y ha sido clasificado en dicha categoría.

Gracias a la información de las distribuciones de probabilidad, podemos cuantificar la incertidumbre semántica. Esto se realiza mediante el cálculo de la entropía y el establecimiento de un umbral (ver Sección 3.1). Una vez identificadas las instancias con mayor incertidumbre, se realiza una selección de categorías e imágenes relevantes para cada una de ellas (Sección 3.2).

Posteriormente, se procede con la desambiguación de las mismas instancias. Para ello, se emplea un Modelo de Lenguaje y Visión a Gran Escala (LVLM) junto con las imágenes anteriores, de tal manera que sea capaz de especificar claramente cuál de las categorías previamente seleccionadas es la correcta (ver Sección 3.3). Finalmente, este resultado se integra de vuelta en Voxeland en forma de nuevas "opiniones subjetivas", aportando una mayor certidumbre al mapa (ver Sección 3.4).

3.1. Identificación de instancias con alta incertidumbre

Voxeland provee la funcionalidad necesaria para exportar la información semántica recopilada en forma de archivo JSON. Esto permite obtener de manera estructurada y sencilla los datos asociados a cada instancia de objeto. Concretamente, para cada instancia se proporciona: el identificador único del objeto, sus dimensiones, el número de observaciones en las que ha aparecido, la distribución de probabilidad sobre las categorías a las que puede pertenecer, y la lista de asociaciones categoría-imagen. Un ejemplo de una instancia representada en el archivo JSON:

```
{
    "obj15": {
            "bbox": {...},
            "n_observations": 103,
            "results": {
                "bed": 38.433941751718521,
                "couch": 50.72217634320259,
                ...
            },
            appearances: {
                "bed": [123,125,245, ... ],
                "couch": [301,302,315, ... ],
                ...
            }
        },
    ...
}
```

Con esta información, es posible cuantificar la incertidumbre asociada a la distribución de probabilidad de las categorías mediante el cálculo del valor esperado de la entropía de Shannon. Siendo C_k el conjunto de categorías de la instancia k , β_k los parámetros de concentración de la distribución de la instancia k, y ψ la función digamma, la entropía se calcula de la siguiente manera:

$$H(C_k) = \psi\left(\sum_l \beta_l\right) - \frac{1}{\sum_l \beta_l} \sum_l \beta_l \psi(\beta_l). \quad (1)$$

Si el valor calculado es mayor que un umbral de 0.7 *nats* (unidad de medida de información basada en logaritmos neperianos, empíricamente establecido) se considera que la instancia presenta un alto grado de incertidumbre y que, por tanto, requiere desambiguación.

3.2. Selección de imágenes

La lista de *appearances* en el anterior ejemplo en formato JSON nos proporciona el conjunto completo de categorías en las que ha sido clasificada una cierta instancia de objeto e imágenes asociadas a ellas. Las imágenes correspondientes a cada categoría aparecen referenciadas por su índice en orden cronológico.

Teniendo en cuenta que el número de categorías presentes puede ser elevado, lo que añadiría cierta complejidad al método propuesto, se propone seleccionar aquellas con mayor probabilidad (3 en este trabajo). De estas categorías más probables, a su vez, debemos escoger un conjunto cerrado y reducido de imágenes, pues un objeto puede llegar a tener miles de apariciones y, consecuentemente, es necesario realizar un proceso de selección de aquellas más convenientes. Dos opciones para esto serían:

- **Selección basada en propiedades de la imagen**: seleccionar aquellas imágenes en las que el objeto ocupe un mayor tamaño en relación a la imagen completa, las que presenten menos ruido, etc. Este enfoque puede llegar a ser muy costoso, puesto que conllevaría analizar varios miles de imágenes para decenas de objetos.

- **Selección independiente de la imagen**: seleccionar imágenes que estén suficientemente separadas en el tiempo para una mayor probabilidad de obtener distintos puntos de vista. Se puede implementar mediante selección aleatoria o basada en división de conjuntos. Por el contrario, este enfoque es más sencillo y menos costoso.

Dado que la segunda opción es más simple, eficiente y potencialmente igual de efectiva, es la que finalmente se aplica para este trabajo.

Además, con el propósito de obtener las clasificaciones más precisas posibles en la siguiente etapa, se considera únicamente el contenido del recuadro que incluye al objeto que se quiere desambiguar. Esto se debe a que, si el objeto en cuestión es demasiado pequeño o hay otro objeto común en todas las imágenes, algo bastante frecuente en escenas que presentan una gran cantidad de elementos, es posible que el LVLM no sea capaz de reconocer claramente cuál es el objeto que se desea desambiguar y resulte en una respuesta incorrecta.

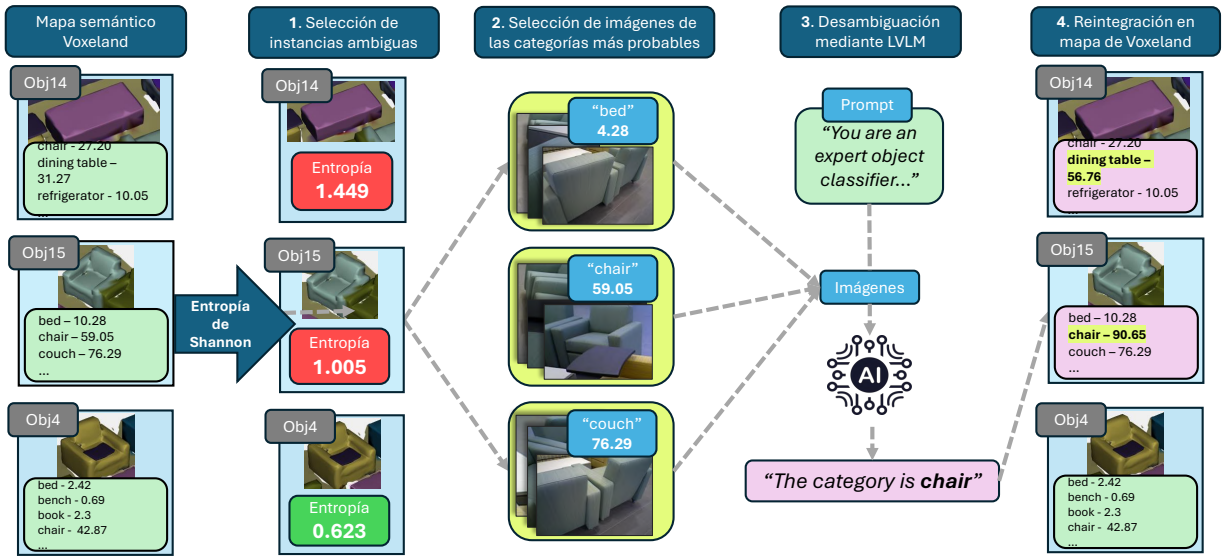

Figura 2: Descripción visual del flujo de trabajo del proceso de desambiguación semántica aplicado a la escena 61 de SceneNN. Cada uno de los recuadros del mapa semántico de Voxeland se corresponden a un objeto de la escena. Por último, el texto resaltado en color amarillo representa el nuevo parámetro de confianza de la categoría resultante por parte del LVLM. Este incremento se calcula a partir de la confianza del LVLM en dicha categoría.

3.3. Desambiguación mediante LVLM

Una vez se han procesado y seleccionado todas las imágenes y categorías de cada objeto, se procede con la desambiguación mediante el uso de un LVLM, el cual recibe de entrada un mensaje o *prompt* con el contexto, las posibles categorías, las imágenes del rectángulo que contiene al objeto, las instrucciones y el formato deseado de la respuesta:

```
You are an expert object classifier. I will provide you with several
images that contain an object seen from different perspectives.
The object belongs to one of the following categories:
[bed, chair, couch]

Your task is to analyze the object and its surrounding environment
across all images to determine the correct category.

Your response must include only one of the previously specified
categories and follow this exact format:
''The category is <category>''
```

En este contexto, el formato del *prompt* es vital, pues estructurar la información de manera correcta es determinante para obtener el mejor resultado posible Moncada-Ramirez et al. (2025). A menudo, no se obtiene el resultado esperado debido a que el *prompt* no está correctamente estructurado. Por lo tanto, tras un proceso de refinamiento y mejora del *prompt*, se propone seguir una estructura de "Contexto" - "Objetivo" - "Restricciones".

3.4. Integración en Voxeland

Con el fin de mantener un enfoque probabilístico acorde al sistema propuesto en Voxeland, el resultado proporcionado por el LVLM no se considera un hecho absoluto, sino una opinión más. Además, dado que la certeza se construye a medida que se incorporan nuevas observaciones, el mismo proceso de consulta al LVLM se repite en numerosas ocasiones para así disponer de una mayor variedad de opiniones.

Esta forma de proceder, además de proporcionar una mayor garantía de corrección, nos permite sacar conclusiones en función de las respuestas que proporcione el modelo. Por ejemplo:

1. Si el modelo responde una categoría con una frecuencia cercana al 100 % quiere decir que está muy convencido del resultado.
2. Si el modelo no responde con tanta certeza ninguna categoría concreta puede significar varias cosas:
   a) El objeto no se corresponde con ninguna de las categorías proporcionadas.
   b) Las categorías son ambiguas y, por tanto, el objeto podría pertenecer a varias.

Consecuentemente, los parámetros de concentración de cada categoría se incrementan de acuerdo con la frecuencia de aparición en las respuestas del modelo. Por ejemplo, si el modelo responde que la categoría es "chair" un 80 % de las ocasiones, se incrementará acordemente.

Por último, el resultado de la ejecución del método se serializa de nuevo en formato JSON, replicando la estructura empleada en Voxeland, de tal modo que pueda ser reintegrado en el marco de trabajo actualizando la información semántica del mapa.

4. Detalles de implementación

La implementación de Voxeland está encapsulada en nodos ROS 2 (Macenski et al., 2022) que se comunican entre sí mediante *topics* y *services*. Concretamente, el nodo principal, "voxeland server" es el encargado de recibir toda la información geométrica y semántica, y, a su vez, integrarla y combinarla a lo largo del proceso de construcción. Este nodo, además, proporciona un *service* para exportar el mapa semántico a formato JSON, lo cual constituye la información de entrada del método de desambiguación propuesto.

Como principal aportación al sistema, se implementa un nuevo nodo encargado exclusivamente del proceso de desambiguación semántica, "voxeland disambiguation node". Principalmente, se encarga de recuperar la información del mapa semántico y de realizar las conversiones a entidades manejables, la identificación de instancias con incertidumbre mediante la entropía, la selección de imágenes y el envío del *prompt* al LVLM.

Para hacer uso de un LVLM, se ha implementado un nuevo nodo ROS 2 que ofrece una interfaz mediante servicios ROS para cargar, descargar y usar distintos modelos.

5. Validación

Para validar la propuesta presentada en este trabajo, se han realizado una serie de pruebas con escenas del popular conjunto de datos SceneNN (Hua et al., 2016). En esta sección se discuten los resultados obtenidos en una de ellas, ya que se considera de especial interés, aunque las pruebas iniciales han sido positivas en las escenas evaluadas. En concreto, se trata de la escena 61 (véase la Figura 3), una escena categorizada como "Lounge" la cual muestra una habitación con un sofá grande y unos sillones alrededor de una mesa central.

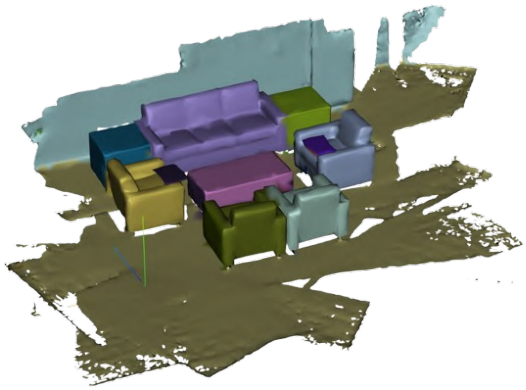

Figura 3: Escena 61 del conjunto de datos de SceneNN

Cabe destacar que todo el procedimiento que se realiza para esta escena, se encuentra visualmente detallado en la Figura 2.

Antes de comenzar con el método propuesto, se debe proceder con la construcción y exportación del mapa semántico mediante Voxeland, el cual serializa la información del mapa en formato JSON:

```
{
    "obj48": {...},
    "obj15": {
        "bbox": {...},
        "n_observations": 115,
        "results": {
            "bed": 10.28901931643486,
            "bench": 1.0200036764144897,
            "chair": 59.00535076856613,
            "couch": 76.29873323440552,
            "suitcase": 0.6537466645240784,
            "tv": 1.0624537765979767
        },
        "appearances: {
            "bed: [19,38,1062,...],
            "chair": [964,1056,1721,...],
            "couch": [110,112,1038,...],
            ...
        }
    },
    ...
}
```

Una vez hemos obtenido el archivo JSON, el nodo encargado de la desambiguación semántica analiza la información de las instancias de objetos para identificar aquellas con un mayor grado de incertidumbre. Para ello se calcula la entropía de la lista de "results" de cada objeto (sección 1 de la Figura 2).

Aplicando este proceso al ejemplo anterior, se determina que el "obj15" (véase la Figura 2) se corresponde con uno de los objetos con una alta entropía, concretamente de 1,005$nats$. Si analizamos su lista de "results", que representa los parámetros de concentración para cada categoría en la que ha sido clasificado dicho objeto, vemos que existe un gran número de categorías, y que los valores correspondientes a varias de ellas son de una magnitud similar, justificando así la incertidumbre en la clasificación.

A continuación, se seleccionan las imágenes a partir de la lista de "appearances" de las 3 categorías con mayor concentración. En este caso, *bed*, *chair* y *couch* son las más probables de representar la categoría correcta del objeto. Siguiendo el procedimiento independiente de la imagen expuesto en la Sección 3.2, se seleccionan, para cada una de las 3 categorías, varias imágenes espaciadas en el tiempo, con el fin de obtener distintos puntos de vista del objeto. Si bien se obtiene la imagen completa, sólo almacenaremos el recuadro aproximado en el que se encuentra el objeto, pues no es conveniente sobrecargar al LVLM con información que no es relevante (sección 2 de la Figura 2).

Con las imágenes del objeto y el *prompt* que incluye las categorías y las instrucciones, podemos hacer uso del nodo LVLM para que decida, finalmente, cuál es la categoría correcta del objeto (sección 3 de la Figura 2). Las respuestas generadas por el LVLM son de la siguiente forma:

```
The category is chair.
```

La respuesta del LVLM se considera válida, pues cumple con las instrucciones que hemos indicado (la respuesta debe incluir sólo una de las 3 categorías que hemos mencionado) y el formato es el esperado (*The category is ...*). Además de ser válida, es correcta, pues según el valor de referencia (del inglés, *ground truth*) la categoría es "chair". Con el fin de obtener una respuesta lo más definitiva posible, se repite este proceso de decisión varias veces, empleando el mismo prompt y las mismas imágenes. Por último, por cada resultado válido devuelto por el nodo LVLM, se actualiza el parámetro de concentración de la categoría devuelta (en este último caso "chair"), incrementando su valor (sección 4 de la Figura 2).

Este procedimiento que se ha aplicado sobre el "obj15" se ha realizado sobre todos los objetos de la escena. Los resultados de las instancias de objetos que presentaban incertidumbre se pueden ver en la tabla 1. Cabe destacar los buenos resultados generalizados, especialmente el caso de los objetos 15, 4 y 9, ya que la categoría más probable (aquella con su parámetro de concentración más alto) no correspondía con el ground truth antes del proceso de desambiguación. Por otro lado, el objeto 14 no ha obtenido buenos resultados, siendo "dining table" y "bench" las categorías más respondidas por el nodo LVLM, ambas con una distribución cercana al 50 %, esto se debe a que la forma

Objeto	Entropía	Nº Categorías	Ground Truth	Predicción más probable	Resultado	% Confianza
Objeto 1	1.318	7	couch	couch	couch	100 %
Objeto 14	1.449	9	dining table	dining table	bench	52 %
Objeto 15	1.005	6	chair	couch	chair	100 %
Objeto 4	1.570	7	chair	couch	chair	96 %
Objeto 9	1.386	7	chair	couch	chair	99 %

Tabla 1: Resultados de clasificación de los objetos detectados como ambiguos en el mapa semántico de Voxeland sobre la escena 61 de SceneNN. *Objeto* hace referencia a los identificadores de los objetos de la escena. *Entropía* y *Nº Categorías* hacen referencia a la entropía calculada y al número de categorías en las que ha sido clasificiado el objeto, respectivamente. *Ground Truth* hace referencia a la categoría real proporcionada por SceneNN. *Predicción más probable* referencia aquella categoría con el parámetro de confianza más alto previo a la desambiguación. Por último, *Resultado* y *% Confianza* hacen referencia a aquella categoría más respondida por el LVLM y el porcentaje de dichas respuestas con respecto al total, respectivamente. Resultados obtenidos tras realizar 100 experimentos sobre cada objeto.

del objeto es muy similar a ambas categorías, correspondiendo con el caso 2a) descrito en la enumeración del punto 3.4.

6. Conclusiones y trabajos

En este trabajo se propone un método para el refinamiento de mapas semánticos construidos por robots móviles que incrementa la certidumbre sobre los mismos. Dicho método trabaja en consonancia con el marco de trabajo probabilístico Voxeland. Para ello se propone el cálculo de la entropía sobre la información semántica para la identificación de instancias de objetos ambiguas, y el uso de de Modelos de Lenguaje y Visión a Gran Escala para la desambiguación semántica. El método ha sido validado con escenas del conjunto de datos de SceneNN, un repositorio del estado del arte muy utilizado en este contexto, y ha obtenido buenos resultados. Además, el método es perfectamente aplicable a cualquier mapa semántico que proporcionen la información semántica de manera similar a Voxeland, ya sea a partir de escenas reales o de cualquier otro conjunto de datos.

Dada la mejora de los mapas semánticos presentada en este trabajo, en el futuro, se estudia la posibilidad de aprovechar la semántica proporcionada por los mapas para que el robot móvil sea capaz de inferir y actuar sobre el entorno a partir de una serie de instrucciones.

Agradecimientos

Este trabajo ha sido desarrollado en el contexto de los proyectos MINDMAPS (PID2023-148191NB-I00) y Voxeland (JA.B1-09), financiados por el Ministerio de Ciencia e Innovación y la Universidad de Málaga, respectivamente.

Referencias

Bac, C. W., van Henten, E. J., Hemming, J., Edan, Y., 2014. Harvesting robots for high-value crops: State-of-the-art review and challenges ahead. Journal of Field Robotics 31 (6), 888–911.
URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21525
DOI: https://doi.org/10.1002/rob.21525

Chaves, D., Ruiz-Sarmiento, J.-R., Petkov, N., Gonzalez-Jimenez, J., 2019. Integration of cnn into a robotic architecture to build semantic maps of indoor environments. In: Advances in Computational Intelligence: 15th International Work-Conference on Artificial Neural Networks, IWANN 2019, Gran Canaria, Spain, June 12-14, 2019, Proceedings, Part II 15. Springer, pp. 313–324.

Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K. M., Sen, B., Agarwal, A., Rivera, C., Paul, W., Ellis, K., Chellappa, R., Gan, C., de Melo, C. M., Tenenbaum, J. B., Torralba, A., Shkurti, F., Paull, L., 2023. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning.
URL: https://arxiv.org/abs/2309.16650

Han, X., Li, S., Wang, X., Zhou, W., 2021. Semantic mapping for mobile robots in indoor scenes: A survey. Information 12 (2).
URL: https://www.mdpi.com/2078-2489/12/2/92
DOI: 10.3390/info12020092

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969.

Hua, B.-S., Pham, Q.-H., Nguyen, D. T., Tran, M.-K., Yu, L.-F., Yeung, S.-K., 2016. Scenenn: A scene meshes dataset with annotations. In: 2016 fourth international conference on 3D vision (3DV). Ieee, pp. 92–101.

Jsang, A., 2018. Subjective Logic: A formalism for reasoning under uncertainty. Springer Publishing Company, Incorporated.

Kostavelis, I., Gasteratos, A., 2015. Semantic mapping for mobile robotics tasks: A survey. Robotics and Autonomous Systems 66, 86–103.

Liu, H., Li, C., Wu, Q., Lee, Y. J., 2023. Visual instruction tuning. arXiv preprint arXiv:2304.08485.

Macenski, S., Foote, T., Gerkey, B., Lalancette, C., Woodall, W., 2022. Robot operating system 2: Design, architecture, and uses in the wild. Science robotics 7 (66), eabm6074.

Matez-Bandera, J.-L., Fernandez-Chaves, D., Ruiz-Sarmiento, J.-R., Monroy, J., Petkov, N., Gonzalez-Jimenez, J., 2022. Ltc-mapping, enhancing long-term consistency of object-oriented semantic maps in robotics. Sensors 22 (14), 5308.

Matez-Bandera, J.-L., Ojeda, P., Monroy, J., Gonzalez-Jimenez, J., Ruiz-Sarmiento, J.-R., 2024. Voxeland: Probabilistic instance-aware semantic mapping with evidence-based uncertainty quantification.
URL: https://arxiv.org/abs/2411.08727

Moncada-Ramirez, J., Matez-Bandera, J.-L., Gonzalez-Jimenez, J., Ruiz-Sarmiento, J.-R., 2025. Agentic workflows for improving large language model reasoning in robotic object-centered planning. Robotics 14 (3), 24.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788.

Ruiz-Sarmiento, J.-R., Galindo, C., Gonzalez-Jimenez, J., 2017. Building multiversal semantic maps for mobile robot operation. Knowledge-Based Systems 119, 257–272.

Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al., 2024. Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800.