Simposio CEA de Robótica, Bioingeniería, Visión por Computador y Automática Marina 2026

Localización global mediante descriptores basados en un diccionario semántico-visual

Flores, M.[a,*], Heredia-Aguado, E.[a], Cabrera, J.J.[a], Gil, A.[a], Valiente, D.[a], Payá, L.[a]

[a] *Instituto de Investigación en Ingeniería de Elche (I3E), Universidad Miguel Hernández de Elche, Avenida de la Universidad, s/n, 03202, Elche, España.*

Resumen

La localización de un robot móvil se puede abordar como un problema de recuperación de imagen. En este enfoque, el entorno se representa mediante un mapa previo compuesto por descriptores de imagen. La posición actual del robot móvil se estima buscando la correspondencia más cercana entre el descriptor de la imagen actual y los almacenados en el mapa. Este trabajo se centra en la parte correspondiente a la generación del descriptor de imagen, proponiendo la creación de un diccionario semántico-visual. Para la creación del diccionario, se extraen los descriptores locales y se les asigna una categoría semántica según su posición en el mapa de segmentación semántica. A partir de esto, se crea un diccionario en dos niveles: semántico y visual. Por un lado, se tiene un conjunto de palabras semánticas que son las categorías asociadas a los descriptores locales. Por otro lado, cada palabra semántica tiene un conjunto de palabras visuales obtenidas agrupando los descriptores locales dentro de esa categoría.

Palabras clave: Robots móviles, Navegación de robots, Construcción de mapas, Localización, Procesamiento de imágenes

Global localization through descriptors based on a semantic-visual dictionary

Abstract

Mobile robot localization can be addressed as an image retrieval problem. In this approach, the environment is represented by a pre-existing map composed of image descriptors. The current position of the mobile robot is estimated by searching for the closest match between the current image descriptor and those stored in the map. This work focuses on the part corresponding to the generation of the image descriptor, proposing the creation of a semantic-visual dictionary. To create the dictionary, local descriptors are extracted and assigned a semantic category based on their position on the semantic segmentation map. From this, a dictionary is created on two levels: semantic and visual. On the one hand, there is a set of semantic words that are the categories associated with the local descriptors. On the other hand, each semantic word has a set of visual words obtained by clustering the local descriptors within that category.

Keywords: Mobile robots, Robot Navigation, Map building, Localization, Image processing

1. Introducción

La localización es una parte fundamental en la robótica móvil. Un robot móvil necesita conocer con precisión cuál es su pose para poder resolver otros desafíos de la navegación autónoma, como la planificación de rutas o la interacción de forma segura del robot móvil con el entorno.

Para abordar el desafío de la localización, se definen dos enfoques complementarios en la literatura. Por un lado, la localización local se centra en estimar la pose (posición y orientación) del robot móvil de forma incremental, a partir de una pose inicial conocida. Este es el enfoque que se utiliza para obtener una estimación de la pose en tiempo real del robot mientras sigue una trayectoria. La herramienta empleada para ello es la odometría. Agostinho et al. (2022) exponen una revisión profunda sobre los métodos de odometría visual y de nubes de puntos. Por otro lado, la localización global es el enfoque empleado cuando el objetivo es que el robot móvil averigüe dónde se encuentra (pose absoluta) dentro de un mapa sin tener información de su pose inicial. En este trabajo, nos centraremos en este enfoque, concretamente en reconocimiento de lugares (en inglés, Place Recognition).

Tanto en el ámbito de la robótica como en el de visión artificial, Place Recognition está recibiendo una gran atención. Esta técnica se centra en identificar la posición en la que una imagen ha sido capturada, encontrando imágenes visualmente similares

*Autor para correspondencia: m.flores@umh.es

dentro de una base de datos previamente geoetiquetada (Zeng et al., 2018). Place recognition formula la localización como un problema de recuperación de imágenes.

En la recuperación de imágenes, los descriptores tienen un papel fundamental. Estos se pueden definir como una firma compacta en la que la información de los píxeles se transforma en una representación de menor dimensionalidad. Dicho formato codifica las características más significativas permitiendo la identificación y comparación eficiente frente a grandes bases de datos. Según el método empleado para la extracción de información visual, se pueden dividir en: globales y locales. En el primero de ellos, una imagen es representada por un único vector. Los métodos tradicionales más utilizados son: el histograma de gradientes orientados (Dalal and Triggs, 2005) (HOG, siglas del inglés Histogram of Oriented Gradients) y Gist (Oliva and Torralba, 2006). Por ejemplo, (Román et al., 2022) analizan el desempeño de diversos descriptores globales en el contexto de una tarea de localización. En cuanto a los descriptores locales, estos se caracterizan por detectar los puntos de interés en la imagen (como esquinas, bordes, blobs) y extraer un descriptor visual para cada uno de ellos. Los métodos tradicionales más empleados para este objetivo son SIFT (Lowe, 2004), SURF (Bay et al., 2006) u ORB (Rublee et al., 2011). Dentro de esta categoría, se encuentra el modelo de bolsa de palabras visuales (BoVW, siglas del inglés Bag of Visual Words). En este método, los descriptores locales son agrupados mediante algoritmos de clustering (como K-means) formando un vocabulario visual. Así pues, cada imagen se representa mediante un histograma de frecuencias de estas ”palabras visuales”. Zeng et al. (2018) enumeran, en una sección de su trabajo, los descriptores de imagen, tanto basados en características locales como globales, que se utilizan para el reconocimiento de lugares.

En este ámbito, también existen trabajos que plantean combinar tanto información visual como semántica. Ouni et al. (2022) proponen tres enfoques para la generación de descriptores de imagen: dos de ellos fusionan ambas modalidades, mientras que el tercero emplea únicamente la segunda. Más tarde y siguiendo esta misma línea, Ouni et al. (2023) introdujeron dos nuevas formas de representación: por un lado, un descriptor que integra simultáneamente las proporciones semánticas de los objetos y sus posiciones espaciales y, por otro, uno que se basa en una bolsa de frases visuales.

La investigación que se presenta en este trabajo se centra en la generación de un descriptor de imagen tomando como referencia el trabajo de Ouni et al. (2022), concretamente, en la variante que combina información semántica y características visuales locales. Tal y como explican en la Sección 3.1 de su trabajo, la imagen viene representada por una matriz de dimensiones $N \times M$, vinculando el número de categorías semánticas (N) con la longitud del descriptor visual (M). Para su generación, dada una imagen, se extraen puntos locales, así como sus correspondientes descriptores visuales, y un mapa de segmentación semántica. Al proyectar los puntos sobre el mapa, a cada uno de los descriptores locales se le asigna una clase, lo que permite agrupar los descriptores en clusters según su clase semántica. El centroide de cada uno de estos clusters constituye una fila. En este trabajo, proponemos representar cada clase semántica mediante una bolsa de palabras visuales. Además, el objetivo del presente trabajo es analizar el desempeño de estos descriptores de imagen basados en información semántica y visual para la realización de la tarea de reconocimiento de lugares.

Las contribuciones principales de este estudio se detallan a continuación:

- La principal propuesta consiste en construir un diccionario semántico-visual formado por palabras semánticas (las clases identificadas en las imágenes del mapa de referencia), vinculando a cada una de ellas una bolsa de palabras visuales. El descriptor de imagen será el resultado de concatenar K_S vectores, siendo K_S igual al número de palabras semánticas. Cada vector se construirá a partir de la bolsa de palabras visuales asociada a esa palabra semántica.

- Siguiendo con la propuesta anterior, se plantean dos variantes para estos vectores: cada celda codifica (a) cuántas veces ha aparecido esa palabra visual (i.e. es un histograma de frecuencia tradicional) y (b) la suma de la distancia coseno cada vez que se ha identificado dicha palabra.

- Se evalúan las dos propuestas y se comparan con el descriptor de imagen propuesto por (Ouni et al., 2022). Esta evaluación se ha centrado en analizar el desempeño de estos descriptores para la localización global mediante un enfoque de recuperación de imágenes. Cabe destacar que únicamente se ha implementado y, por tanto, evaluado la representación de imagen descrita por los autores en la Sección 3.1 de su artículo y que identifican mediante BoSW (no su marco global completo para CBIR).

- El desempeño de estos descriptores de imagen se analizan con imágenes que presentan cierta distorsión, pues las imágenes utilizadas son equirectangulares, y capturadas en un entorno exterior.

2. Metodología

El proceso de reconocimiento de lugares consta de al menos dos fases: la creación del mapa de referencia (offline) y la localización global (online). En la primera, el objetivo es crear un modelo visual del entorno por el que va a navegar el robot móvil. Para ello, se adquieren previamente, en ese mismo entorno, un conjunto de imágenes junto con las coordenadas en las que fueron capturadas. Posteriormente, se extraen los descriptores de imagen de cada captura y se almacenan junto con las posiciones, generando así el mapa de referencia. Durante la segunda fase (online), el robot captura una imagen en tiempo real mientras navega. Para poder determinar dónde se encuentra, compara el descriptor de la imagen actual con los almacenados en el mapa de referencia y recupera la coincidencia más alta.

Este sistema se divide principalmente en dos módulos: la generación de descriptor de imagen y la recuperación de imágenes. El primero tiene como objetivo extraer la información más relevante y codificarla en una representación matemática compacta, permitiendo que la comparación entre dos imágenes sea

más eficiente. El presente trabajo se enfoca en este módulo, proponiendo distintos tipos de descriptores de imagen basados en el enfoque de bolsa de palabras visuales.

La generación del descriptor de imagen inicia con las extracción de características. En este estudio se combinan dos tipos: semánticas y visuales, aspecto que se detallará en el apartado 2.1. Dado que las representaciones propuestas se basan en el enfoque de bolsa de palabras visuales, la integración de ambos tipos de información da lugar a la construcción diccionario jerárquico compuesto de dos niveles.

La Figura 1 muestra visualmente cómo se compone esta jerarquía. El nivel alto, basado en información semántica, consta de un conjunto de categorías identificadas en las imágenes que constituyen el mapa de referencia. Siguiendo la analogía del diccionario, a estas categorías se les denomina palabras semánticas. Por su parte, cada palabra semántica tiene a su vez su propio diccionario visual (nivel bajo).

El proceso de creación de este diccionario jerárquico se aborda en el apartado 2.2, mientras que el apartado 2.3 detalla cómo se construyen los descriptores de imagen que se proponen.

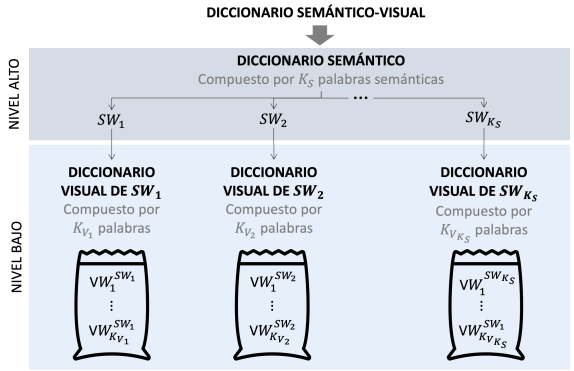


Figura 1: Diccionario semántico-visual en dos niveles: palabras semánticas (e.g. SW_i) y palabras visuales (e.g. $VW_j^{SW_i}$)

2.1. Extracción de características

La extracción de características se realiza en ambas fases (online y offline) y constituye la etapa fundamental para la generación del descriptor de imagen, así como para la creación del diccionario. Tal y como se ha señalado, en este trabajo, se combina información semántica y visual.

Para ello, se utiliza un algoritmo clásico de extracción de características locales. Este paso se divide en una primera etapa de detección (es decir, la búsqueda de puntos o zonas relevantes en la imagen, como esquinas o blobs) y en una segunda etapa de descripción (donde se analiza el entorno local y se genera un descriptor). Dada una imagen (I), se aplica ORB (Rublee et al., 2011) obteniendo así un conjunto de n puntos locales con sus descriptores visuales:

$$\mathrm{ORB}(I) = \{(\mathbf{p}_i, \mathbf{d}_i)\}_{i=1}^{n} \tag{1}$$

donde $\mathbf{p}_i \in \mathbb{R}^2$ son las coordenadas espaciales (x, y) en las que se encuentra el punto local y $\mathbf{d}_i \in \{0,1\}^{dim}$ es el descriptor binario correspondiente a $\mathbf{p}_i$. El siguiente paso es asignar una clase semántica a cada uno de esos descriptores ($\mathbf{d}_i$). Para lograrlo, se utiliza una técnica de segmentación semántica que asigna, a cada píxel de la imagen, una categoría semántica, tal y como se muestra a continuación:

$$S : \Omega \rightarrow C \tag{2}$$

donde $\Omega = \{(x, y) \in \mathbb{Z}^2 \mid 1 \leq x \leq W, 1 \leq y \leq H\}$ representa el dominio espacial de la imagen y C es el conjunto de las categorías semánticas que el modelo es capaz de identificar. El modelo de segmentación S asocia a cada coordenada $(x, y) \in \Omega$ de la imagen $I \in \mathbb{R}^{H \times W \times 3}$ una clase $c \in C$.

Al integrar estos dos pasos, la posición espacial de cada descriptor local permite acceder al mapa de segmentación semántica y extraer su categoría. De esta manera, al finalizar esta etapa, se tiene un conjunto de descriptores locales ($\{\mathbf{d}_i\}_{i=1}^{n}$) junto a sus respectivas etiquetas semánticas ($\{c_i\}_{i=1}^{n}$):

$$\Phi : I \in \mathbb{R}^{H \times W \times 3} \rightarrow F \tag{3}$$

donde Φ representa el proceso de extracción de características descrito en este apartado y $F = \{(\mathbf{d}_i, c_i)\}_{i=1}^{n}$ es el conjunto de n características visuales ($\mathbf{d}_i$) y semánticas (c_i).

2.2. Creación del diccionario jerárquico

La creación del diccionario se ejecuta exclusivamente durante la fase offline. Una vez extraídas las características de todas las M imágenes del mapa de referencia (tal como se describe en el apartado 2.1), se obtiene un conjunto total de características ($\mathcal{F}$):

$$\mathcal{F} = \bigcup_{j=1}^{M} F_j \tag{4}$$

donde $F_j = \left\{(\mathbf{d}_{i,j}, c_{i,j}) \mid i = 1, \ldots, n_j\right\}$ representa los n_j elementos extraídos de la imagen j. Después de esto, se procede a crear el diccionario semántico (D_S), que se compondrá de todas las categorías únicas identificadas en todo el conjunto de referencia:

$$D_S = unique(\bigcup_{j=1}^{M} \{c_{i,j}\}_{i=1}^{n_j}) \tag{5}$$

De este modo, el tamaño del diccionario semántico (K_S) vendrá dado por el tamaño del conjunto D_S. Puesto que cada descriptor local tiene asociada una categoría, estos se agrupan en función de ella.

La construcción del diccionario visual para cada palabra semántica consiste en aplicar el algoritmo de K-means sobre sus descriptores asociados, agrupándolos en K clusters según su proximidad en el espacio de características. En consecuencia, el diccionario asociado a esa palabra semántica constará de K_V palabras visuales que son los centroides de los K clusters. Considerando que cada palabra semántica tendrá un número distinto de descriptores, el tamaño de cada diccionario visual (K_V) será el mínimo entre la mitad del número de descriptores y un valor fijo (K_{max}) especificado previamente. En este trabajo, este parámetro se ha configurado en 10, por lo que ningún diccionario visual excederá esta cantidad de palabras, adaptándose a la baja si hay un menor número de descriptores.

2.3. *Generación del descriptor de imagen*

La generación del descriptor de imagen se lleva a cabo en ambas fases (online y offline), requiriendo previamente la extracción de características (apartado 2.1) y la disponibilidad del diccionario jerárquico (apartado 2.2).

De modo que, dada una imagen (I_x), tras el proceso descrito en el apartado 2.1, se obtiene un conjunto de características ($F_x = \{(\mathbf{d}_i, c_i)\}_{i=1}^{n_x}$). A continuación, mediante el algoritmo de K Nearest Neighbors (KNN), se asigna a cada descriptor $\mathbf{d}_i$ la palabra visual más cercana dentro del diccionario de la palabra semántica c_i.

Ambas representaciones propuestas consisten en concatenar K_S vectores (uno por palabra semántica). La diferencia entre estos dos enfoques reside en la información que codifica cada celda. El primer método corresponde al modelo tradicional de bolsa de palabras, generando un histograma de frecuencias (al sumar +1 en la posición de la palabra visual asignada). El segundo método utiliza la distancia coseno obtenida durante el algoritmo KNN, acumulando el valor de dicha distancia en la celda correspondiente en lugar de registrar únicamente la frecuencia de aparición.

3. Experimentos y resultados

3.1. *Base de datos*

Para este estudio, la base de datos seleccionada es KITTI-360 (Liao et al., 2023) debido a que las imágenes fueron capturadas por un sistema de visión de 360 grados. El sistema de visión empleado estaba formado por dos cámaras de ojo de pez (con un campo de visión de 180 grados cada una) montadas sobre un vehículo apuntando hacia lados distintos. Como etapa previa a la experimentación, se generaron proyecciones equirectangulares (véase la Figura 2) a partir de cada par de imágenes ojo de pez originales proporcionado por la propia base de datos.

Figura 2: Ejemplo de imagen equirectangular del dataset.

Cabe señalar que, en este trabajo, los mapas de segmentación semántica no se generan en la parte de extracción de características (apartado 2.1), sino que se han obtenido de manera offline (como etapa de preprocesamiento), evitando así la carga computacional durante la ejecución del algoritmo principal. El modelo utilizado es SegFormer (Xie et al., 2021), una arquitectura basada en transformers que proporciona la clasificación píxel a píxel necesaria para el sistema.

3.2. *Subconjuntos de imágenes*

Para la ejecución del proceso, se requieren dos subconjuntos de datos: uno para la fase offline (cuya finalidad es crear el mapa de referencia), y otro para la fase online (empleado para resolver la localización).

El primero de ellos, imágenes de referencia, se compone de un total de 791 imágenes, obtenidas mediante un muestreo espacial de la base de datos KITTI-360 (Liao et al., 2023). En concreto, se extrajeron de la secuencia original las capturas con un intervalo de 10 metros, empezando por la posición inicial de la trayectoria. Las 9723 imágenes restantes forman el segundo subconjunto, es decir, son las imágenes de consulta.

Tabla 1: Resumen de los subconjuntos: el nombre identificador, la fase del proceso, la función que se lleva a cabo y el número de imágenes que componen cada subconjunto.

Nombre	Fase	Función	Número
Imágenes de referencia	Offline	Modelo del entorno	791 imágenes
Imágenes de consulta	Online	Localización	9723 imágenes

3.3. *Medida de evaluación*

El proceso de reconocimiento de lugares consiste en obtener para cada imagen de consulta (fase online) su correspondencia más cercana en el mapa de referencia (fase offline). Para poder evaluar esta recuperación, se procede a calcular el error de localización cometido haciendo uso de la información de posición proporcionada por la base de datos (ground truth). Este error se define como la distancia euclídea ($dist_{t-r}$) entre la posición de la imagen de consulta (XYZ_t) y la posición de la imagen devuelta como la más similar (XYZ_r).

Para poder determinar la validez del algoritmo, se establecerá un umbral de distancia (d_{th}) y si la distancia obtenida es menor a este umbral, se considera que la localización se ha realizado con éxito. Atendiendo a este criterio, se asignará un valor de 1 al Recall@1 de la imagen de consulta (I_{t_i}) si la distancia ($dist_{t-r}$) es inferior a dicho umbral, y de 0 en caso contrario.

En la parte de evaluación (apartado 3.4), este se cuantificará mediante la media aritmética de los valores de $Recall@1$ obtenidos para el total de $n = 9723$ imágenes del conjunto de consulta, de acuerdo con la siguiente expresión:

$$\overline{Recall@1}(\,\%) = \frac{\sum_{i=1}^{n} R@1_{t_i}}{n} \cdot 100 \tag{6}$$

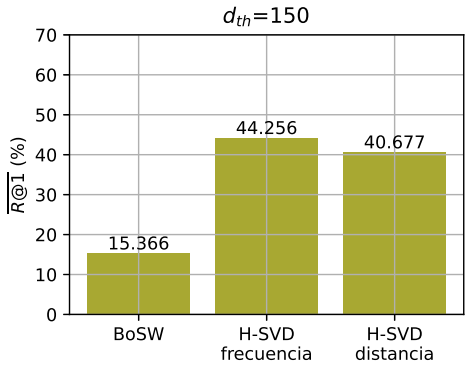
3.4. *Evaluación*

A continuación, se evalúa y compara la eficacia de los distintos descriptores de imagen para resolver el problema de localización global de un robot móvil. Este trabajo propone dos tipos de representaciones que han sido generadas a partir de un diccionario jerárquico semántico-visual (H-SVD, del inglés Hierarchical Semantic-Visual Dictionary).

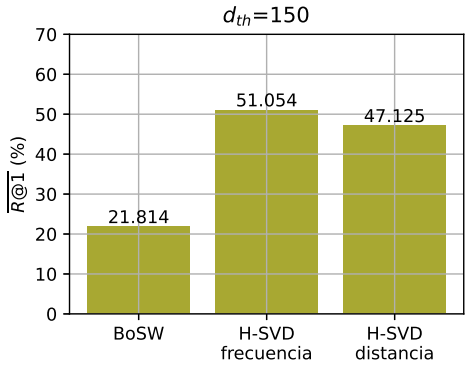
La diferencia entre ambas propuestas radica en el vector que identifica cada palabra semántica, tal y como se describe en el apartado 2.3. Además, estos dos enfoques también se compararán frente al que proponen Ouni et al. (2022), al cual nos referiremos como BoSW, en el apartado 3.1 de su trabajo (del cual

únicamente se ha implementado, en este trabajo, la generación de su representación de imagen).

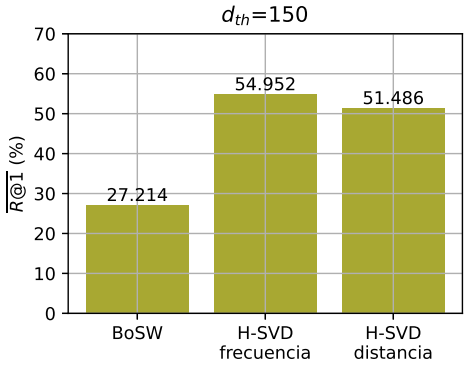
Para la evaluación, se han establecido tres valores para el umbral de distancia (d_{th}): 10, 50 y 100, de más a menos restrictivos. Las Figuras 3(a), 3(b) y 3(c) muestran respectivamente el valor medio de *Recall*@1 en tanto por cien ($\overline{Recall@1}$(%)) obtenido para cada umbral mediante un diagrama de barras. En cada gráfica, hay tres barras que corresponden a las tres representaciones de imagen evaluadas: la primera corresponde a la propuesta por Ouni et al. (2022) (BoSW) y las otras dos a los enfoques diseñados en este trabajo, donde la segunda representa al basado en frecuencia (H-SVD frecuencia), y la tercera, al basado en distancia (H-SVD distancia).



(a) d_{th} = 10 *metros*



(b) d_{th} = 50 *metros*



(c) d_{th} = 100 *metros*

Figura 3: Resultados del $\overline{Recall@1}$(%) obtenidos para los distintos umbrales de distancia.

En primer lugar, se analizan los resultados mostrados en la Figura 3(a), correspondientes al umbral más restrictivo (d_{th} = 10 *metros*). Se observa una clara superioridad al usar los descriptores propuestos en este trabajo frente a la representación de imagen desarrollada por Ouni et al. (2022) (BoSW). Mientras que el primero, BoSW, consigue un valor de $\overline{Recall@1}$ de 15,366 %, las variantes H-SVD basadas en frecuencia y distancia logran un 44,256 % y un 40,677 %, respectivamente. Aunque la diferencia entre nuestras dos propuestas no es tan notoria, el descriptor de imagen semántico-visual basado en frecuencia (H-SVD) es el que consigue el valor más alto. Si comparamos este mejor resultado con el peor (BoSW), se evidencia una mejora relativa del $\overline{Recall@1}$ del 188 %.

En segundo lugar, se analizan los resultados para un umbral d_{th} = 50 *metros*, presentados en la Figura 3(b). Al igual que ocurría con el caso anterior, se mantiene una diferencia notable entre el modelo de referencia (BoSW) (Ouni et al., 2022) y los enfoques propuestos en este trabajo. Al observar los valores de $\overline{Recall@1}$, BoSW alcanza un 21,814 %, mientras que las variantes H-SVD basadas en frecuencia y distancia logran un 51,054 % y un 47,125 %, respectivamente. Nuevamente, H-SVD basado en frecuencia es el que presenta el mejor rendimiento. Si comparamos el mejor resultado (H-SVD frecuencia) y el peor (BoSW), para d_{th} igual a 50 metros, se consigue una mejora relativa del $\overline{Recall@1}$ del 134 %.

En tercer lugar, se analizan los resultados mostrados en la Figura 3(c), correspondientes al umbral menos restrictivo (d_{th} = 100 *metros*). Al analizar la gráfica, se puede ver que el valor más bajo de $\overline{Recall@1}$ se ha obtenido con el descriptor de imagen propuesto por Ouni et al. (2022) (BoSW) y el valor más alto con el enfoque basado en frecuencia (H-SVD frecuencia). La diferencia entre estos dos tipos de descriptores es de 101,93 %, ya que BoSW ha obtenido un valor de 27,214 % y H-SVD frecuencia de 54,952 %. El otro enfoque propuesto presenta un valor de $\overline{Recall@1}$ igual a 51,486 % y, al igual que en los casos anteriores, es el segundo que mejor rendimiento ofrece.

En ultimo lugar, se van a analizar las tres gráficas de forma conjunta centrándonos en el progreso de cada tipo de descriptor de imagen al aumentar el valor del umbral de distancia (d_{th}). En el caso de BoSW, el valor de $\overline{Recall@1}$ aumenta un 41,96 % al ser menos restrictivos con el umbral y pasar de considerar que una imagen de consulta ha encontrado su imagen más cercana de forma satisfactoria cuando la distancia real entre ambas es de 10 metros a que sea de 50 metros. Al ser aún menos restrictivos y considerar que este umbral es de 100 metros, el valor de $\overline{Recall@1}$ se incrementa un 24,75 % con respecto al umbral de distancia de 50 metros. En el caso de H-SVD basado en frecuencia, el valor de $\overline{Recall@1}$ cuando el umbral de distancia es de 50 metros supera en un 15,36 % al valor cuando el umbral es de 10 metros. Siguiendo esta línea, al aumentar el umbral de distancia a 100 metros, el crecimiento del valor de $\overline{Recall@1}$ es del 7,64 %. Por último, en el caso de H-SVD basado en distancia, el valor de $\overline{Recall@1}$ se incrementa un 15,85 % al aumentar el umbral de distancia de 10 a 50 metros y un 9,25 % al extenderse hasta los 100 metros.

A continuación, se enumeran algunas de las conclusiones tras realizar los experimentos y analizar los resultados. El empleo del descriptor semántico-visual de imagen basado en fre-

cuencia (H-SVD frecuencia) es el más eficaz para recuperar la imagen de referencia más similar, aún aumentando el valor de umbral de distancia que se utiliza para determinar que el Recall de esa imagen de consulta es uno. Todos los descriptores de imagen mejoran el valor de $\overline{Recall@1}$ al aumentar el umbral de distancia (d_{th}), pero este crecimiento es mayor para el descriptor de imagen propuesto por Ouni et al. (2022) (BoSW). El mejor de los descriptores de imagen propuestos (H-SVD frecuencia) presenta una amplia diferencia con respecto al descriptor de imagen propuesto por Ouni et al. (2022) (BoSW), consiguiéndose una mejora de más del 100 % en los tres casos.

4. Conclusiones

El principal objetivo de este trabajo es resolver el problema de localización de un robot móvil como un enfoque de recuperación de imágenes. Para ello, se proponen dos descriptores de imágenes basados en un diccionario semántico-visual.

Este diccionario se divide en dos niveles, uno semántico y otro visual. Dado que integra ambas modalidades de información, el primer paso para su creación es extraer los descriptores locales de todas las imágenes que componen el mapa. A estos descriptores se le asigna una categoría semántica, la cuál se extrae del mapa de segmentación semántica accediendo a la posición píxel del descriptor local. Este conjunto de características (descriptores y categorías semánticas) es el utilizado para la creación del diccionario. Por un lado, el nivel alto se compone de un diccionario compuesto por K_S palabras semánticas que son las categorías del conjunto de características identificadas en las imágenes de referencia. Por otro lado, el nivel bajo está constituido por K_S diccionarios visuales, cada uno de ellos con un máximo de K_V palabras. Cada diccionario visual se obtiene mediante el agrupamiento (clustering) de los descriptores locales dentro de la misma categoría semántica.

En la parte de experimentos de este trabajo, se han evaluado estas dos representaciones de imagen junto con la propuesta por Ouni et al. (2022) (BoSW). Para la evaluación y comparación de los tres enfoques, se ha utilizado el valor medio de la métrica Recall@1.

Tras evaluar los tres descriptores de imagen, los resultados determinan que uno de los propuestos, concretamente el basado en frecuencia (H-SVD frecuencia), es el más eficaz para la recuperación de imágenes, consiguiendo una mejora superior al 100 % con respecto al descriptor de imagen propuesto por Ouni et al. (2022) (BoSW).

Como trabajo futuro, se propone extender este estudio utilizando otros métodos de extracción y descripción de características locales. Finalmente, otro trabajo futuro que se planea es cuantificar la importancia de cada categoría semántica y utilizar esto para ponderar el diccionario semántico.

Agradecimientos

Este trabajo es parte del proyecto CIPROM/2024/8, financiado por la Generalitat Valenciana, Conselleria de Educación, Cultura, Universidades y Empleo (programa PROMETEO). También forma parte del proyecto PID2023-149575OB-I00, financiado por MICIU/AEI/10.13039/501100011033 y por FEDER,UE

Referencias

Agostinho, L. R., Ricardo, N. M., Pereira, M. I., Hiolle, A., Pinto, A. M., 2022. A practical survey on visual odometry for autonomous driving in challenging scenarios and conditions. IEEE Access 10, 72182–72205.
DOI: 10.1109/ACCESS.2022.3188990

Bay, H., Tuytelaars, T., Van Gool, L., 2006. SURF: Speeded Up Robust Features. In: Leonardis, A., Bischof, H., Pinz, A. (Eds.), Computer Vision – ECCV 2006. Springer, Berlin, Heidelberg, pp. 404–417.
DOI: 10.1007/11744023_32

Dalal, N., Triggs, B., Jun. 2005. Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). Vol. 1. pp. 886–893 vol. 1, iSSN: 1063-6919.
URL: https://ieeexplore.ieee.org/document/1467360
DOI: 10.1109/CVPR.2005.177

Liao, Y., Xie, J., Geiger, A., Mar. 2023. KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D. IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (3), 3292–3310.
URL: https://ieeexplore.ieee.org/document/9786676/
DOI: 10.1109/TPAMI.2022.3179507

Lowe, D. G., Nov. 2004. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60 (2), 91–110.
URL: https://doi.org/10.1023/B:VISI.0000029664.99615.94
DOI: 10.1023/B:VISI.0000029664.99615.94

Oliva, A., Torralba, A., 2006. Chapter 2 building the gist of a scene: the role of global image features in recognition. In: Martinez-Conde, S., Macknik, S., Martinez, L., Alonso, J.-M., Tse, P. (Eds.), Visual Perception. Vol. 155 of Progress in Brain Research. Elsevier, pp. 23–36.
URL: https://www.sciencedirect.com/science/article/pii/S0079612306550022
DOI: https://doi.org/10.1016/S0079-6123(06)55002-2

Ouni, A., Chateau, T., Royer, E., Chevaldonné, M., Dhome, M., Mar. 2023. An efficient ir approach based semantic segmentation. Multimedia Tools and Applications 82 (7), 10145–10163.
URL: https://link.springer.com/10.1007/s11042-022-14297-7
DOI: 10.1007/s11042-022-14297-7

Ouni, A., Royer, E., Chevaldonné, M., Dhome, M., Dec. 2022. Leveraging semantic segmentation for hybrid image retrieval methods. Neural Computing and Applications 34 (24), 21519–21537.
URL: https://link.springer.com/10.1007/s00521-021-06087-3
DOI: 10.1007/s00521-021-06087-3

Román, V., Payá, L., Cebollada, S., Peidró, A., Reinoso, O., 2022. Evaluating the Robustness of New Holistic Description Methods in Position Estimation of Mobile Robots. In: Gusikhin, O., Madani, K., Zaytoon, J. (Eds.), Informatics in Control, Automation and Robotics. Springer International Publishing, Cham, pp. 207–225.
DOI: 10.1007/978-3-030-92442-3_12

Rublee, E., Rabaud, V., Konolige, K., Bradski, G., Nov. 2011. ORB: An efficient alternative to SIFT or SURF. In: 2011 International Conference on Computer Vision. pp. 2564–2571, iSSN: 2380-7504.
URL: https://ieeexplore.ieee.org/document/6126544
DOI: 10.1109/ICCV.2011.6126544

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., Luo, P., 2021. Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems 34, 12077–12090.

Zeng, Z., Zhang, J., Wang, X., Chen, Y., Zhu, C., Nov. 2018. Place Recognition: An Overview of Vision Perspective. Applied Sciences 8 (11), 2257.
URL: https://www.mdpi.com/2076-3417/8/11/2257
DOI: 10.3390/app8112257