CEA Comité Español de Automática
XLVI JORNADAS AUTOMÁTICA CARTAGENA 2025
UNIVERSIDADE DA CORUÑA

Jornadas de Automática

Localización visual mediante imágenes omnidireccionales y técnicas de fusión temprana

Alfaro, Marcos[a,*], Cabrera, Juan José[a], Reinoso, Oscar[a,b], Gil, Arturo[a], Payá, Luis[a,b]

[a] *Instituto Universitario de Investigación en Ingeniería de Elche (I3E), Universidad Miguel Hernández (UMH), Av/ de la Universidad s/n, 03202, Elche, España.*
[b] *Valencian Graduate School and Research Network of Artificial Intelligence (ValgrAI), Camí de Vera, Edificio 3Q, 46020, Valencia, España.*

To cite this article: Alfaro, Marcos, Cabrera, Juan José, Reinoso, Oscar, Gil, Arturo, Payá, Luis. 2025. Visual place recognition with omnidirectional images and early fusion techniques. Jornadas de Automática, 46. https://doi.org/10.17979/ja-cea.2025.46.12239

Resumen

Las cámaras omnidireccionales son una opción recomendable para la localización de robots móviles, debido a su capacidad de extraer información abundante y contextual de la escena con un campo de visión elevado. No obstante, la información visual es inherentemente sensible a los cambios de apariencia del entorno, lo que puede afectar a la robustez del sistema. Para abordar esta limitación, en este trabajo se propone combinar imágenes omnidireccionales con características intrínsecas derivadas de ellas, como la intensidad promedio o la magnitud del gradiente, mediante técnicas de fusión temprana. Posteriormente, la información fusionada es procesada por una red neuronal convolucional, previamente entrenada con extensas bases de datos para la tarea de localización visual. Los resultados obtenidos demuestran que enriquecer la información visual con estas características mejora significativamente la robustez del sistema, permitiendo una localización precisa y fiable tanto en entornos interiores como exteriores, incluso bajo condiciones de iluminación muy variadas. El código utilizado está disponible a través del siguiente enlace: `https://github.com/MarcosAlfaro/LocalizacionVisualFusionTemprana/` .

Palabras clave: Robótica móvil, Localización visual, Cámaras omnidireccionales, Fusión sensorial, Aprendizaje profundo.

Visual place recognition with omnidirectional images and early fusion techniques.

Abstract

Omnidirectional cameras are a highly suitable option for mobile robot localization, given their ability to capture abundant and contextual scene information with a wide field of view. However, pure visual data is inherently sensitive to environmental appearance changes, which can impact system robustness. To address this limitation, this paper proposes combining omnidirectional images with intrinsic features derived from them, such as average intensity or gradient magnitude, using early fusion techniques. Subsequently, the fused information is processed by a convolutional neural network, pre-trained on extensive datasets for visual place recognition. The obtained results demonstrate that enriching visual information with these features significantly enhances system robustness, enabling precise and reliable localization in both indoor and outdoor environments, even under highly varied lighting conditions. The code used is available via the following link: `https://github.com/MarcosAlfaro/LocalizacionVisualFusionTemprana/` .

Keywords: Mobile robotics, Visual localization, Omnidirectional cameras, Sensory fusion, Deep learning.

*Autor para correspondencia: malfaro@umh.es

1. Introducción

La inteligencia artificial (IA) y la visión por computador han experimentado un crecimiento significativo en los últimos años, revolucionando el campo de la navegación autónoma. Esto ha impulsado el desarrollo de metodologías avanzadas para la localización, la creación de mapas (SLAM) y diversas tareas de comprensión de la escena, tales como la detección de objetos y la estimación de transitabilidad (Flores et al., 2021; Santo et al., 2023), entre otras. Dentro de este contexto, el reconocimiento de lugares (*place recognition*) emerge como una tarea fundamental, la cual consiste en identificar la posición de un vehículo móvil dentro de un entorno conocido a partir de datos sensoriales previamente adquiridos (modelo visual). Su rol es crucial para la posterior localización precisa y la navegación segura, lo que ha motivado una extensa investigación en el diseño de nuevos modelos y técnicas para abordar este desafío (Masone and Caputo, 2021; Yin et al., 2025).

La localización en un mapa a partir de información sensorial requiere el procesamiento y la compresión de dicha información en un descriptor representativo. Anteriormente, este procesamiento se realizaba mediante técnicas analíticas, tales como HOG o gist (Payá et al., 2016). Sin embargo, en la actualidad, los modelos de aprendizaje profundo, tales como las redes neuronales convolucionales (CNNs) (Arandjelovic et al., 2016) y los transformers (Dosovitskiy et al., 2020), han demostrado ser altamente efectivos para esta tarea.

La arquitectura de estos modelos varía significativamente en función de la modalidad sensorial empleada. Por un lado, existen arquitecturas especializadas en el procesamiento de nubes de puntos generadas por sensores LiDAR (Cabrera et al., 2024; Karypidis et al., 2024). Si bien estos sensores ofrecen precisión e invariancia ante cambios de iluminación, su coste es considerablemente alto. Por otro lado, se han desarrollado numerosos modelos basados en imágenes, las cuales proporcionan información rica en color, textura y forma a un coste inferior, convirtiéndolas en una opción atractiva para el reconocimiento de lugares (Ali-Bey et al., 2023; Izquierdo and Civera, 2024). No obstante, los datos visuales son susceptibles a variaciones lumínicas y al *visual aliasing*, fenómeno que ocurre cuando ubicaciones distintas en el entorno presentan apariencias muy similares.

Dadas las limitaciones inherentes a cada tipo de sensor, una estrategia común consiste en la fusión multisensorial, combinando LiDAR y cámaras para explotar las ventajas de cada modalidad y mitigar sus inconvenientes (Lai et al., 2022). Estas estrategias se clasifican generalmente en fusión temprana (*early fusion*), en la cual la información sensorial se combina antes de ser procesada por la red (Liu et al., 2022), y fusión tardía (*late fusion*), donde las distintas modalidades son procesadas de manera independiente y se fusionan sus salidas (descriptores) (Pan et al., 2024). Sin embargo, la integración efectiva de múltiples sensores introduce complejidad adicional y suele implicar un mayor coste computacional y de hardware.

Por este motivo, el presente trabajo propone enriquecer las imágenes capturadas por robots móviles mediante la extracción de características intrínsecas derivadas de las propias imágenes. Estas características se combinan con la información visual mediante técnicas de fusión temprana. De esta manera, se consigue abordar la localización de manera precisa y robusta utilizando cámaras omnidireccionales como única fuente de información sensorial, emergiendo como una solución efectiva y económica para la localización de un robot móvil.

2. Metodología

2.1. *Visión omnidireccional*

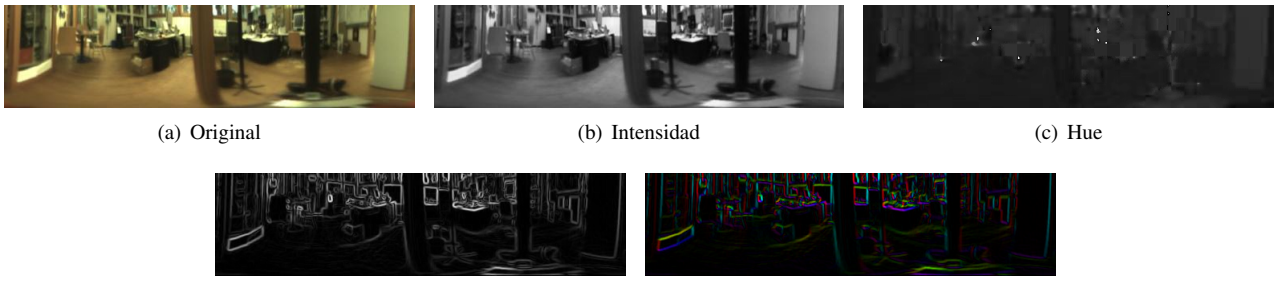
Dentro de los sensores visuales, las cámaras omnidireccionales destacan por su elevado campo de visión. Gracias a esta propiedad, es posible extraer información completa del entorno independientemente de la orientación del vehículo, lo que incrementa la robustez en la localización ante cambios de orientación. En este trabajo, se han empleado dos tipos de sistemas de visión omnidireccional para abordar la tarea de place recognition:

- Sistema catadióptrico: Está formado por una cámara estándar y un espejo hiperbólico. Su funcionamiento se basa en la incidencia de los haces de luz sobre el espejo y su reflexión en el foco de la hipérbola, donde se sitúa la cámara. En los experimentos realizados, las imágenes resultantes han sido convertidas a formato panorámico (véase la Figura 1, parte superior).
- Cámara 360°: Este sensor es capaz de obtener imágenes con un campo de visión de 360° en todos los ejes. Este tipo de cámaras sobresale por su capacidad de generar imágenes con una resolución muy elevada. A partir de la información visual capturada, se realiza una proyección equirrectangular, tal y como se muestra en la Figura 1, parte inferior.

Figura 1: Sistemas de visión omnidireccional empleados en este trabajo: sistema catadióptrico (parte superior) y cámara 360° (parte inferior).

2.2. *Características visuales*

Aunque la información visual es de gran utilidad para la comprensión de la escena que rodea a un robot móvil, a menudo se ve afectada por los cambios de apariencia de la escena o el visual aliasing. Por este motivo, en este trabajo se ha enriquecido esta información con una serie de características



(a) Original (b) Intensidad (c) Hue

(d) Magnitud del gradiente (e) Orientación del gradiente

Figura 2: (a) Ejemplo de una imagen panorámica de la base de datos COLD (Pronobis and Caputo, 2009) y mapas de características obtenidos a partir de la imagen: (b) intensidad, (c) tono, (d) magnitud del gradiente y (e) orientación del gradiente.

intrínsecas obtenidas a partir de las propias imágenes, las cuales se describen a continuación. Las Figuras 2 y 3 muestran ejemplos de cada una de las características extraídas a partir de las imágenes originales.

- Intensidad: Calculada como el valor promedio de los canales rojo (R), verde (G) y azul (B).

$$I = \frac{R + G + B}{3} \tag{1}$$

- Tono (*Hue*): Se define como el color puro predominante en el píxel, y viene dado por:

$$H = \cos^{-1}\left[\frac{(R-G)+(R-B)}{2\sqrt{(R-G)^2+(R-G)(G-B)}}\right] \tag{2}$$

- Gradiente: Representa la variación de intensidad en el entorno de un píxel, obtenida mediante operadores de Sobel de tamaño 3x3:

$$\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0^* & 0 \\ 1 & 2 & 1 \end{bmatrix}; \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0^* & 2 \\ -1 & 0 & 1 \end{bmatrix} \tag{3}$$

  - Intensidad: A partir del resultado de aplicar los filtros anteriores, la magnitud del gradiente se calcula mediante la suma del valor absoluto de las variaciones de intensidad en ambos ejes:

$$Mag = |\Delta x| + |\Delta y| \tag{4}$$

  - Orientación: A partir de los filtros de Sobel, la orientación del gradiente se define como la dirección de máximo cambio de la intensidad, y viene dada por:

$$\theta = \tan^{-1}\left(\frac{\Delta y}{\Delta x}\right) \tag{5}$$

2.3. *Técnica de fusión utilizada*

En este trabajo, se ha optado por emplear una estrategia de fusión temprana (*early fusion*) para combinar la información visual con el resto de características. Esta técnica consiste en concatenar una imagen a color de dimensiones $C \times H \times W$, donde C es el número de canales (en este caso, $C = 3$), H es el número de filas y W es el número de columnas, con el mapa de características resultante. Esto genera una imagen de entrada de dimensiones $(C + n) \times H \times W$, siendo n el número de características añadidas.

Posteriormente, esta imagen es introducida a la red, lo que permite una interacción entre la información visual y el resto de características desde la primera capa convolucional. La salida de la red es un único descriptor global que se utilizará posteriormente para la tarea de place recognition.

2.4. *Selección y adaptación del modelo de red*

Para comprimir las imágenes en descriptores visuales, se ha empleado CosPlace (Berton et al., 2022), un modelo de CNN entrenado con millones de imágenes para la tarea de place recognition. Dentro de los modelos disponibles, se ha seleccionado la arquitectura VGG16 como extractor de características, con un tamaño de descriptor de 512. Se ha aplicado la técnica de *transfer learning*, es decir, se ha partido de los pesos pre-entrenados del modelo en todas las capas de la red.

A continuación, este modelo ha sido adaptado para procesar las imágenes junto con el resto de características. Concretamente, se ha modificado la capa de entrada de tal forma que el número de canales de entrada sea $3 + n$. En cuanto a la inicialización de los pesos de la primera capa, se ha realizado una evaluación de distintas técnicas, la cual se describe con detalle en la Sección 3.4.

3. Experimentos

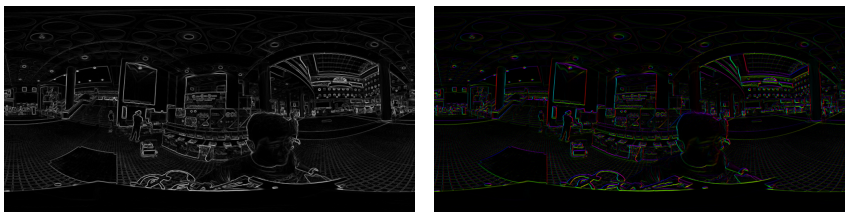
3.1. *Bases de datos utilizadas*

El objetivo principal de este trabajo es evaluar la influencia de las características visuales en el rendimiento de los modelos entrenados para la tarea de place recognition, así como su capacidad de generalización a otros escenarios. Para ello, hemos utilizado dos bases de datos: una con imágenes obtenidas en entornos interiores (COLD) y otra capturada en exteriores (360Loc).

3.1.1. *COLD*

La base de datos COLD (Pronobis and Caputo, 2009) contiene imágenes capturadas con un sistema catadióptrico montado en un robot móvil. Las imágenes corresponden a varios entornos interiores y fueron obtenidas bajo tres condiciones de iluminación distintas: nublado, noche y soleado. Para adaptar las imágenes a las convoluciones rectangulares de la red utilizada, las imágenes omnidireccionales originales, con una

(a) Original (b) Intensidad (c) Hue



(d) Magnitud del gradiente (e) Orientación del gradiente

Figura 3: (a) Ejemplo de una imagen panorámica de la base de datos 360Loc (Huang et al., 2024) y los mapas de características obtenidos a partir de la imagen: (b) intensidad, (c) tono, (d) magnitud del gradiente y (e) orientación del gradiente.

dimensión de 640 × 480 píxeles, se han convertido a formato panorámico, resultando en imágenes de 128 × 512 píxeles. En los experimentos, se ha empleado un conjunto reducido de imágenes capturadas bajo condiciones nubladas en el entorno *Friburgo A* para el entrenamiento. La evaluación se ha realizado en cuatro entornos distintos, *Friburgo A (FR-A)* y *B (FR-B)*, y *Saarbrücken A (SA-A)* y B *(SA-B)*, bajo las tres condiciones de iluminación disponibles. La Tabla 1 detalla el número de imágenes que componen las secuencias de entrenamiento y test.

Tabla 1: Conjuntos de imágenes utilizados para el entrenamiento y test de los modelos en la base de datos COLD. *Secuencia de entrenamiento.

Entorno	Entrenamiento / Modelo Visual	Test Nublado	Test Noche	Test Soleado
FR-A	556*	2595	2707	2114
FR-B	560	2008	-	1797
SA-A	586	2774	2267	-
SA-B	321	836	870	872

Tabla 2: Conjuntos de imágenes utilizados para el entrenamiento y test de los modelos en la base de datos 360Loc. *Secuencia de entrenamiento.

Entorno	Entrenamiento / Modelo Visual	Test Día	Test Noche
atrium	581*	875	1219
concourse	491	593	514
hall	540	1123	1061
piatrium	632	1008	697

3.1.2. 360Loc

La base de datos 360Loc (Huang et al., 2024) permite la localización cruzada entre imágenes en distintos formatos: estándar (*pinhole*), equirrectangular y ojo de pez (*fisheye*). Las imágenes fueron capturadas en un entorno semiabierto, caracterizado por cambios de apariencia más pronunciados. En este entorno, se han registrado diversas trayectorias tanto de día como de noche. En el presente trabajo, hemos utilizado únicamente las imágenes equirrectangulares, las cuales poseen una resolución de 6144x3072 píxeles. Para optimizar el procesamiento de la red, su tamaño ha sido reducido a 384x192 píxeles. Para el entrenamiento, se ha empleado la secuencia *database* del entorno *atrium*, y la evaluación se ha realizado en los cuatro entornos disponibles: *atrium*, *concourse*, *hall* y *piatrium*, tanto de día como de noche. La Tabla 2 muestra la distribución de las secuencias de entrenamiento y test.

3.2. Entrenamiento de la red

Para el entrenamiento de los modelos, se ha adoptado una estrategia de aprendizaje por contraste mediante una arquitectura tripleta. Es decir, la red ha sido entrenada con 100000 combinaciones de tres imágenes denominadas ancla (I_a), positiva (I_p) y negativa (I_n). Estas combinaciones se seleccionan de tal forma que se cumpla: $dist(I_a, I_p) <= r_p$, y $dist(I_a, I_n) > r_n$, siendo $r_p <= r_n$. De este modo, la red debe aprender a generar descriptores similares para imágenes capturadas en posiciones cercanas, y descriptores diferentes para imágenes que pertenecen a zonas distintas. Para el experimento con la base de datos COLD, r_p y r_n han tomado el valor de $0,4m$, mientras que para la base de datos 360Loc, r_p y r_n han tomado los valores de $2m$ y $5m$, respectivamente. Se ha empleado la función de pérdida Lazy Triplet (Uy and Lee, 2018), que devuelve el mayor error cometido en la predicción de todo el batch:

$$\mathcal{L} = \left[\max\left(\vec{D}_{a,p} - \vec{D}_{a,n} + m\right)\right]_+ \tag{6}$$

, donde $\vec{D}_{a,p}$ representa las distancias euclídeas entre los descriptores ancla y positivo, $\vec{D}_{a,n}$ son las distancias euclídeas entre los descriptores ancla y negativo, m es el margen, y $[\cdot]_+$ es la función ReLU.

En los experimentos, se ha empleado un margen $m = 0,5$ y un tamaño de batch $N = 4$. El algoritmo de optimización empleado ha sido SGD (Stochastic Gradient Descent), con una tasa de aprendizaje de 0,001. Todos los experimentos se han realizado con una GPU NVIDIA GeForce RTX 3090 de 24GB de memoria.

Tabla 3: Resultados de R@1 (Early Fusion) en la base de datos COLD, desglosados por entorno y condición de iluminación.

Características	FR-A			FR-B		SA-A		SA-B			Global
	Cloudy	Night	Sunny	Cloudy	Sunny	Cloudy	Night	Cloudy	Night	Sunny	
Color	92,91	95,01	83,35	85,86	85,92	**76,74**	64,31	88,35	78,51	83,94	83,59
Hue	93,41	95,16	85,24	**86,50**	89,93	74,85	62,51	87,20	78,51	**85,57**	83,90
Intensidad (*I*)	93,14	94,75	85,43	85,96	88,26	75,86	63,08	88,88	79,43	84,29	83,91
Gradiente (*Mag*)	92,83	95,09	84,91	85,91	90,04	76,43	63,74	90,31	80,00	83,49	84,28
Gradiente (θ)	90,79	91,50	73,70	84,06	88,87	71,46	45,48	79,43	64,02	79,93	76,92
Mag + Hue	**93,49**	95,35	83,21	85,81	86,53	75,73	64,34	89,23	80,92	84,75	83,99
I + Hue	93,29	**95,68**	84,91	85,91	**90,48**	74,90	66,74	87,56	82,07	83,26	84,47
I + Mag	93,22	95,27	85,86	85,66	89,48	75,99	67,45	88,23	**83,56**	84,63	**85,04**
I + Hue + Mag	92,22	95,05	**86,14**	86,11	87,76	74,30	**67,98**	**90,56**	82,30	85,55	84,85

3.3. Evaluación y métrica utilizada

Para realizar la evaluación, se ha seguido la estrategia común de place recognition: para cada imagen de test I_{test}, se obtiene su descriptor $\vec{d}_{test}$ a través del modelo entrenado. Este descriptor se compara, mediante la distancia euclídea, con los descriptores de las imágenes que componen el modelo visual $D^{VM} = \left[\vec{d}_1, ..., \vec{d}_n\right]$. El descriptor más cercano indica la imagen predicha como la más próxima $\left(I_{pred}\right)$, y la distancia geométrica entre la posición de esta imagen y la de test (I_{test}) determina el error cometido $e_{pred} = \left\|(x_{pred}, y_{pred}) - (x_{test}, y_{test})\right\|$.

La métrica utilizada en los experimentos ha sido el Recall@1 ($R@1$), que representa la proporción de imágenes localizadas correctamente dentro de un radio d. El Recall@1 se calcula mediante la siguiente fórmula:

$$R@1(\%) = \frac{\sum e_{pred} <= d}{\sum I_{test}} \tag{7}$$

, donde $d = 0,5m$ para todos los entornos de COLD, $d = 5m$ para el entorno *concourse* de 360Loc, y $d = 10m$ para el resto de entornos de 360Loc.

3.4. Análisis experimental

En primer lugar, se han comparado distintas formas de adaptar los pesos de la capa de entrada de la red. Se han evaluado cuatro métodos: la inicialización aleatoria de todos los pesos de la primera capa, la inicialización parcialmente aleatoria (manteniendo los pesos originales para los canales RGB e inicializando aleatoriamente los adicionales), asignar a los canales adicionales la media de los canales RGB, y copiar sucesivamente los pesos de los canales RGB en los canales adicionales. Para este experimento, se ha utilizado la característica de intensidad. La Tabla 4 muestra los resultados de $R@1$ obtenidos en el entorno FR-A con cada método.

Tabla 4: Recall@1 (%) obtenido con el método de early fusion para distintas técnicas de transfer learning en la primera capa.

Método	**Cloudy**	**Night**	**Sunny**	**Medio**
Totalmente aleatorio	84,97	93,24	32,92	70,38
Parcialmente aleatorio	90,79	92,13	36,33	73,08
Media canales RGB	**93,14**	94.75	**85,43**	91,11
Copia canales RGB	92,64	**95,86**	84,86	**91,12**

La Tabla 4 indica que los métodos que implican una inicialización total o parcialmente aleatoria de los pesos resultaron en un rendimiento muy bajo en la condición de soleado, la cual difiere considerablemente de la condición de entrenamiento (nublado). Esto sugiere que estos modelos experimentaron un mayor sobreajuste a la condición de entrenamiento. Por el contrario, los otros dos métodos mostraron un desempeño considerablemente mejor en todas las condiciones de iluminación. Aunque ambos métodos tuvieron resultados globalmente muy similares, se ha escogido el método "Copia de canales RGB" para los experimentos posteriores debido a su ligero mejor rendimiento global y en la condición de noche.

A continuación, se han evaluado los modelos entrenados con cada conjunto de características visuales. Las Tablas 3 y 5 muestran los resultados de R@1 obtenidos para los distintos entornos y condiciones de iluminación de las bases de datos COLD y 360Loc, respectivamente. El mejor resultado para cada condición se muestra en negrita, mientras que el segundo mejor resultado aparece subrayado.

A partir de las Tablas 3 y 5, se puede observar que, en líneas generales, la combinación de características que ha ofrecido los mejores resultados ha sido la intensidad junto con la magnitud del gradiente ($I + Mag$). Si se analizan los resultados por separado, se puede observar que, por un lado, en entornos interiores (COLD), se produce una mejora de los resultados al emplear cualquiera de las características por separado, a excepción de la orientación del gradiente (θ). Por otro lado, en entornos exteriores (360Loc), fusionar la información visual con una única característica no ha producido una mejora en los resultados globales en comparación con usar solo color. Esto se debe a que este entorno presenta cambios de apariencia mucho más pronunciados, lo que provoca una mayor diferencia entre las imágenes del modelo visual y las de test. No obstante, el empleo de las características adecuadas (intensidad y magnitud del gradiente) ha permitido aumentar la robustez ante estos cambios lumínicos.

4. Conclusiones

En este artículo, se ha llevado a cabo la tarea de reconocimiento de lugares (place recognition) utilizando imágenes capturadas por cámaras omnidireccionales. Estas imágenes han sido enriquecidas con características intrínsecas (como la intensidad media de la imagen o la magnitud del gra-

Tabla 5: Resultados de R@1 (Early Fusion) en la base de datos 360Loc, desglosados por entorno y condición de iluminación.

Características	atrium		concourse		hall		piatrium		Global
	Day	Night	Day	Night	Day	Night	Day	Night	
Color	93,00	**79,41**	91,95	**84,05**	92,05	**85,25**	83,70	43,76	79,14
Hue	86,41	67,26	91,05	64,79	88,87	51,48	77,27	42,75	71,23
Intensidad (I)	93,80	77,97	92,58	78,02	**92,40**	70,14	**85,25**	49,21	79,92
Gradiente (Mag)	88,23	68,66	91,77	77,32	91,22	84,76	84,07	**51,08**	77,20
Gradiente (θ)	**95,23**	60,60	87,99	75,68	89,96	42,99	83,14	29,27	70,57
Mag + Hue	84,33	69,22	**93,84**	63,81	87,79	58,36	80,58	39,74	72,21
I + Hue	85,47	76,84	88,64	63,62	88,22	58,48	78,09	41,32	72,58
I + Mag	93,98	79,12	92,71	81,13	91,84	73,54	83,24	46,77	**80,29**
I + Mag + Hue	92,76	79,22	89,59	81,13	89,86	62,06	84,30	49,49	78,45

diente) mediante técnicas de fusión temprana. El objetivo de esta aproximación era mejorar la robustez del sistema ante cambios de apariencia pronunciados y su capacidad de generalización a entornos no vistos. Los resultados demuestran que la combinación adecuada de imágenes omnidireccionales con las características extraídas mejora significativamente el rendimiento de los modelos de aprendizaje profundo, logrando resultados muy competitivos en entornos y condiciones de iluminación no utilizados durante el entrenamiento de la red.

En cuanto a trabajos futuros, se explorará el uso de otros tipos de características, tales como la profundidad o la información semántica. Además, se desarrollarán nuevas técnicas de fusión sensorial que incorporen mecanismos de atención, lo que podría permitir una integración más completa de las diferentes modalidades de información.

Agradecimientos

El Ministerio de Ciencia, Innovación e Universidades ha financiado este trabajo a través de las becas FPU23/00587 (M.A.) y FPU21/04969 (JJ. C.). Este trabajo se enmarca dentro del proyecto TED2021-130901B-I00, financiado por MICIU/AEI/10.13039/501100011033 y por la Unión Europea NextGenerationEU/PRTR. También forma parte del proyecto PID2023-149575OB-I00, financiado por MICIU/AEI/10.13039/501100011033 y por FEDER, UE.

Referencias

Ali-Bey, A., Chaib-Draa, B., Giguere, P., 2023. MixVPR: Feature mixing for visual place recognition. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2998–3007.
DOI: 10.48550/arXiv.2303.02190

Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J., 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5297–5307.
DOI: https://arxiv.org/pdf/2303.02190

Berton, G., Masone, C., Caputo, B., 2022. Rethinking visual geo-localization for large-scale applications. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4878–4888.
DOI: 10.48550/arXiv.2204.02287

Cabrera, J. J., Santo, A., Gil, A., Viegas, C., Payá, L., 2024. MinkUNeXt: Point cloud-based large-scale place recognition using 3D sparse convolutions. arXiv preprint arXiv:2403.07593.
DOI: 10.48550/arXiv.2403.07593

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
DOI: 10.48550/arXiv.2010.11929

Flores, M., Valiente, D., Gil, A., Peidró, A., Reinoso, O., Payá, L., 2021. Evaluación de descriptores locales en localización visual con imágenes ojo de pez. In: XLII Jornadas de Automática. Universidade da Coruña, Servizo de Publicacións, pp. 507–514.
DOI: 10.17979/spudc.9788497498043.507

Huang, H., Liu, C., Zhu, Y., Cheng, H., Braud, T., Yeung, S.-K., June 2024. 360Loc: A dataset and benchmark for omnidirectional visual localization with cross-device queries. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22314–22324.
DOI: 10.48550/arXiv.2311.17389

Izquierdo, S., Civera, J., 2024. Optimal transport aggregation for visual place recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition. pp. 17658–17668.
DOI: 10.48550/arXiv.2311.15937

Karypidis, E., Kakogeorgiou, I., Gidaris, S., Komodakis, N., 2024. DINO-Foresight: Looking into the future with DINO. CoRR.
DOI: 10.48550/arXiv.2412.11673

Lai, H., Yin, P., Scherer, S., 2022. Adafusion: Visual-LiDAR fusion with adaptive weights for place recognition. IEEE Robotics and Automation Letters 7 (4), 12038–12045.
DOI: 10.1109/LRA.2022.3210880

Liu, W., Fei, J., Zhu, Z., 2022. MFF-PR: Point cloud and image multi-modal feature fusion for place recognition. In: 2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE, pp. 647–655.
DOI: 10.1109/ISMAR55827.2022.00082

Masone, C., Caputo, B., 2021. A survey on deep visual place recognition. IEEE Access 9, 19516–19547.
DOI: 10.1109/ACCESS.2021.3054937

Pan, Y., Xie, J., Wu, J., Zhou, B., 2024. Camera-LiDAR fusion with latent correlation for cross-scene place recognition. IEEE Transactions on Industrial Electronics.
DOI: 10.1007/978-3-031-72754-2_25

Payá, L., Reinoso, O., Berenguer, Y., Úbeda, D., 2016. Using omnidirectional vision to create a model of the environment: A comparative evaluation of global-appearance descriptors. Journal of Sensors 2016 (1), 1209507.
DOI: 10.1155/2016/1209507

Pronobis, A., Caputo, B., 2009. COLD: The CoSy localization database. The International Journal of Robotics Research 28 (5), 588–594.
DOI: 10.1177/0278364909103912

Santo, A., Gil, A., Valiente, D., Ballesta, M., Reinoso, O., 2023. Estimación de zonas transitables en nubes de puntos 3D con redes convolucionales dispersas. In: XLIV Jornadas de Automática. Universidade da Coruña. Servizo de Publicacións, pp. 737–737.
DOI: 10.17979/spudc.9788497498609.732

Uy, M. A., Lee, G. H., 2018. PointNetVLAD: Deep point cloud based retrieval for large-scale place recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4470–4479.
DOI: 10.48550/arXiv.1804.03492

Yin, P., Jiao, J., Zhao, S., Xu, L., Huang, G., Choset, H., Scherer, S., Han, J., 2025. General place recognition survey: Towards real-world autonomy. IEEE Transactions on Robotics.
DOI: 10.1109/TRO.2025.3550771