(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 980 672**

(21) Número de solicitud: 202231063

(51) Int. CI.:

***G06V 20/50*** (2012.01)
*G06F 18/24* (2013.01)

(12) SOLICITUD DE PATENTE A1

(22) Fecha de presentación:

**13.12.2022**

(43) Fecha de publicación de la solicitud:

**02.10.2024**

(71) Solicitantes:

**UNIVERSIDAD DE LEÓN (50.0%)**
**Avda. de la Facultad, 25**
**24071 León (León) ES y**
**S.M.E. INSTITUTO NACIONAL DE CIBERSEGURIDAD DE ESPAÑA M.P., S.A. (50.0%)**

(72) Inventor/es:

**SAIKIA, Surajit;**
**FERNÁNDEZ ROBLES, Laura;**
**ALEGRE GUTIÉRREZ, Enrique y**
**FIDALGO FERNÁNDEZ, Eduardo**

(74) Agente/Representante:

**CARVAJAL Y URQUIJO, Isabel**

(54) Título: **MÉTODO Y SISTEMA DE CLASIFICACIÓN Y RECUPERACIÓN DE ESCENAS DE INTERIOR**

(57) Resumen:

Método y sistema de clasificación y recuperación de escenas de interior. El método de clasificación de escenas de interior (100) comprende detectar (120) y clasificar (130) objetos (122) en una imagen de entrada (102); generar (140) una descripción (142) de la imagen de entrada incluyendo etiquetas (132) de objetos detectados e información relativa a la frecuencia de aparición de las etiquetas (132) en la imagen de entrada (102); obtener (160), a partir de una representación vectorial (152) de la descripción (142), una primera predicción (162) de categorías de escena asignadas a la imagen de entrada (102) con sus probabilidades ($P_{OC}$) asociadas; obtener (170), a partir del contenido global de la imagen de entrada (102), una segunda predicción (172) de categorías de escena asignadas a la imagen de entrada (102) con sus probabilidades ($P_{SC}$); y combinar (180) la primera (162) y segunda (172) predicción de categorías de escena mediante una primera función de peso (350) para obtener una clasificación de categoría de escena (182) de la imagen de entrada (102).

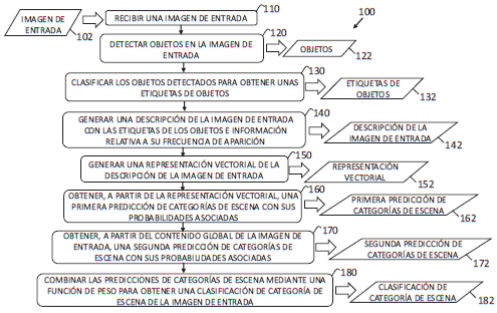


FIG. 1

ES 2 980 672 A1

## DESCRIPCIÓN

# MÉTODO Y SISTEMA DE CLASIFICACIÓN Y RECUPERACIÓN DE ESCENAS DE INTERIOR

## OBJETO DE LA INVENCIÓN

El objeto de la presente invención es un procedimiento y sistema automatizado para reconocer escenas de interior en función de las características de una imagen tomada en esa escena. La invención permite reconocer la escena de interior indicando la categoría de la escena contenida de la imagen. La presente invención también permite identificar las imágenes, de un conjunto, tomadas en la misma categoría de escena aunque en diferentes ángulos y condiciones que una imagen dada, y las devuelve ordenadas en función de similitud con respecto a la imagen dada.

## ANTECEDENTES DE LA INVENCIÓN

El reconocimiento de escenas consiste en anotar las imágenes en función de las categorías de las escenas en las que han sido adquiridas. Estas categorías se definen según los objetos y superficies que aparecen en ellas, su distribución en el espacio y el entorno en el que se encuentran.

En el campo de la visión por computador, el reconocimiento de escenas es una de las tareas que habitualmente se estudian. Una escena representa un entorno del mundo real que contiene múltiples objetos y superficies que están organizados de manera significativa. Los seres humanos pueden reconocer sin esfuerzo y rápidamente una imagen de escena observando el contenido sin tener que observar todos los detalles; por ejemplo, identificar un cuarto de baño sin fijarse en algunos objetos concretos, como una toalla, un fregadero o el jabón. Esta capacidad derivada del sistema visual del cerebro se conoce como reconocimiento de objetos y es fundamental para la interpretación del entorno. Aunque los enfoques modernos de aprendizaje automático y profundo pueden lograr este objetivo, el reconocimiento de escenas de interiores sigue siendo una tarea de investigación compleja.

En los primeros años, las representaciones de escenas se fundamentaban principalmente en descriptores globales que utilizaban representaciones de baja dimensionalidad para modelar la percepción humana [1]. Sin embargo, la capacidad de estos descriptores está limitada debido a la complejidad de las imágenes de la escena, que se caracterizan por la presencia

de patrones complejos, fondos desordenados y múltiples objetos [2]. Además, varios métodos implementaron modelos computacionales para el reconocimiento holístico de escenas utilizando una representación de baja dimensión, conocido como envolvente espacial [3].

Para mejorar el rendimiento, se han propuesto muchos métodos basados en descriptores locales para extraer características correspondientes a varios parches en las imágenes. Algunos de los descriptores notables incluyen los patrones binarios locales (del inglés, Local Binary Patterns - LBP) [4], SIFT [5], SURF [6], HOG [7] y las bolsas de palabras visuales (del inglés, Bag of Visual Words -BOVW) [8]. En particular, Wu y Rehg [9] utilizaron el kernel de intersección de histogramas para mejorar la generación de un libro de códigos visual, e introdujeron un algoritmo k-means del kernel de histograma para aumentar la precisión del reconocimiento. Posteriormente, Quin y Yun [10] incorporaron información contextual con palabras visuales para crear un diccionario más discriminativo.

La mayoría de las imágenes de escenas están definidas por algunas regiones discriminantes, que pueden ser parches de textura u objetos. Algunos métodos ([11], [12]) utilizan la detección de objetos para determinar dichas regiones discriminantes para clasificar las escenas. Asimismo, otros enfoques utilizan un gran número de parches de imagen para identificar regiones importantes ([13], [14], [15]). Lin et al. [16] propusieron un método que usa filtros de partes para obtener las respuestas de las regiones de objetos que constituyen una escena determinada. Otros enfoques ([17],[18]) explotaron la información proporcionada por la distribución de patrones de objetos con respecto a diferentes escenas.

Sin embargo, el reto común de estos algoritmos de reconocimiento de escenas es el control de las variaciones intra e interclase. Para superar este reto, Zuo et al. [19] introdujeron un método conocido como aprendizaje discriminativo y de características compatibles (del inglés Discriminative and Shareable Features Learning, -DSFL), que minimiza las distancias de las características aprendidas de las mismas clases y maximiza las distancias de las características aprendidas de las clases diferentes.

El aprendizaje profundo ha demostrado ser eficaz para el reconocimiento de escenas. Las capas inferiores de las redes neuronales convolucionales (del inglés, Convolutional Neural Networks – CNN) capturan características locales mientras que las capas superiores generan características más abstractas. Fundamentándose en esta propiedad, Xie et al. [20] y Tang et al. [21] fusionaron características correspondientes a capas convolucionales multiescala para

el reconocimiento de escenas. De manera similar, una arquitectura CNN multirresolución fue propuesta por Guo et al. [22], demostrando que se pueden capturar características correspondientes a diferentes capas para la comprensión de la escena. Dixit et al. [23] propusieron vectores semánticos de Fisher para fusionar características de las capas convolucionales y totalmente conectadas de las CNN. Por otro lado, Wang et al. [24] introdujeron una red llamada PatchNet, que agrega tanto el objeto como las características holísticas de la escena para desarrollar una representación visual efectiva de las escenas. Otro modelo que combina la información del objeto con la de la escena es el FOSNet, que consiste en una CNN de extremo a extremo propuesta por Soeng et al. [25]. Recientemente, Ma et al. [26] presentaron SceneNet, una arquitectura neuronal profunda que se utiliza para el reconocimiento de escenas en imágenes de teledetección.

Para el reconocimiento de escenas en interiores, Basu et al. [27] entrenaron una red neuronal de cápsulas, que obtuvo un mejor rendimiento en comparación con otros modelos basados en CNN. Las características extraídas de las CNN también pueden utilizarse con diferentes métodos de reconocimiento de escenas. Yoo et al. [28] utilizaron la activación de CNN multiescala con el modelo del kernel de Fisher y obtuvieron una ganancia de rendimiento significativa en el conjunto de datos MIT-67. Cimpoi et al. [29] mejoraron la generalización del reconocimiento de escenas extrayendo la información de textura de un banco de filtros CNN y un BOVW. Recientemente, López-Cifuentes et al. [30] propusieron una CNN multimodal que combina la información del contexto con la imagen de la escena utilizando un módulo de atención. Además, para mejorar la precisión del reconocimiento, Li et al. [31] introdujeron MAPnet, que fusiona la información de profundidad y de la imagen RGB. En particular, la tarea de reconocimiento de escenas de interior se basa comúnmente en la detección de objetos de interior, y por lo tanto la mayoría de los algoritmos emplean detectores de objetos para identificar los objetos inherentes. En lo que respecta a la percepción de escenas de interiores para robots móviles, Ran et al. [32] diseñaron una estructura CNN superficial y eficiente que alcanzó una mayor precisión en el reconocimiento de escenas utilizando imágenes de cámaras monoculares.

Tal y como se describe en la literatura ([33],[34]), los métodos que se centran en aumentar las capas de la red de las CNN sólo conducen a una ganancia de rendimiento sub-lineal. Esto se debe a la incapacidad de las redes para abordar la diversidad en las similitudes dentro de las mismas clases [35].

Los métodos del estado del arte necesitan redes muy profundas y enormes conjuntos de entrenamiento para el reconocimiento y la recuperación de escenas en interiores. Se hace necesario disponer de un nuevo método de reconocimiento y recuperación de escenas en interiores que resuelva estos problemas.

*Referencias bibliográficas*

[1] Oliva, Aude. “Gist of the scene.” In Neurobiology of attention, pp. 251-256. Academic press, 2005.

[2] Xie, L., Lee, F., Liu, L., Kotani, K. and Chen, Q., 2020. Scene recognition: A comprehensive survey. Pattern Recognition, 102, p.107205.

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

[4] Ojala T, Pietikäinen M, Harwood D (1996) A comparative study of texture measures with classification based on featured distributions. Pattern Recognition 29(1):51–59.

[5] Lowe DG (2004) Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2):91–110.

[6] Bay H, Tuytelaars T, Van Gool L (2006) SURF: Speeded up robust features. In: European Conference on Computer Vision, Springer, pp 404–417.

[7] Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), IEEE, vol 1, pp. 886–893.

[8] Csurka G, Dance C, Fan L, Willamowski J, Bray C (2004) Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV, Prague, vol. 1, pp. 1–2.

[9] Wu y Rehg (Wu J, Rehg JM (2009) Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. In: 2009 IEEE 12th International Conference on Computer Vision, IEEE, pp 630–637.

[10] Qin J, Yung NH (2010) Scene categorization via contextual visual words. Pattern Recognition 43(5):1874–1888.

[11] Li LJ, Su H, Fei-Fei L, Xing EP (2010) Object bank: A high-level image representation for scene classification & semantic feature sparsification. In: Advances in Neural Information Processing Systems, pp 1378–1386.

[12] Pandey M, Lazebnik S (2011) Scene recognition and weakly supervised object localization with deformable part-based models. In: 2011 International Conference on Computer Vision,

IEEE, pp 1307– 1314.

[13] Singh S, Gupta A, Efros AA (2012) Unsupervised discovery of mid-level discriminative patches. In: European Conference on Computer Vision, Springer, pp 73–86.

[14] Juneja M, Vedaldi A, Jawahar C, Zisserman A (2013) Blocks that shout: Distinctive parts for scene classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 923–930.

[15] Yuan Y, Wan J, Wang Q (2016) Congested scene classification via efficient unsupervised feature learning and density estimation. Pattern Recognition 56:159–169.

[16] Lin D, Lu C, Liao R, Jia J (2014) Learning important spatial pooling regions for scene classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3726– 3733.

[17] Song X, Jiang S, Herranz L (2017) Multi-scale multi-feature context modeling for scene recognition in the semantic manifold. IEEE Transactions on Image Processing 26(6):2721–2735.

[18] Wu R, Wang B, Wang W, Yu Y (2015) Harvesting discriminative meta objects with deep CNN features for scene classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1287–1295.

[19] Zuo Z, Wang G, Shuai B, Zhao L, Yang Q, Jiang X (2014) Learning discriminative and shareable features for scene classification. In: European Conference on Computer Vision, Springer, pp 552–568.

[20] Xie GS, Zhang XY, Yan S, Liu CL (2015) Hybrid CNN and dictionary-based models for scene recognition and domain adaptation. IEEE Transactions on Circuits and Systems for Video Technology 27(6):1263–1274.

[21] Tang P, Wang H, Kwong S (2017) G-MS2F: GoogLeNet based multi-stage feature fusion of deep cnn for scene recognition. Neurocomputing 225:188–197.

[22] Guo S, Huang W, Wang L, Qiao Y (2016) Locally supervised deep hybrid model for scene recognition. IEEE Transactions on Image Processing 26(2):808– 820.

[23] Dixit M, Chen S, Gao D, Rasiwasia N, Vasconcelos N (2015) Scene classification with semantic fisher vectors. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp 2974– 2983.

[24] Wang Z, Wang L, Wang Y, Zhang B, Qiao Y (2017) Weakly supervised PatchNets: Describing and aggregating local patches for scene recognition. IEEE Transactions on Image Processing 26(4):2028–2041.

[25] Seong, H., Hyun, J. and Kim, E., 2020. FOSNet: An end-to-end trainable deep neural network for scene recognition. IEEE Access, 8, pp.82066-82077.

[26] Ma, A., Wan, Y., Zhong, Y., Wang, J. and Zhang, L., 2021. SceneNet: Remote sensing scene classification deep learning network using multi-objective neural evolution architecture search. ISPRS Journal of Photogrammetry and Remote Sensing, 172, pp.171-188.

[27] Basu A, Petropoulakis L, Di Caterina G, Soraghan J (2020) Indoor home scene recognition using capsule neural networks. Procedia Computer Science 167:440–448.

[28] Yoo, D., Park, S., Lee, J.Y. and Kweon, I.S., 2014. Fisher kernel for deep neural activations. arXiv preprint arXiv:1412.1628.

[29] Cimpoi, M., Maji, S. and Vedaldi, A., 2015. Deep filter banks for texture recognition and segmentation. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 3828-3836).

[30] López-Cifuentes A, Escudero-Viñolo M, Bescós J, García-Martín A (2020) Semantic-aware scene recognition. Pattern Recognition 102:107256.

[31] Li Y, Zhang Z, Cheng Y, Wang L, Tan T (2019) MAPNet: Multi-modal attentive pooling network for rgb-d indoor scene classification. Pattern Recognition 90:436–449.

[32] Ran, T., Yuan, L. and Zhang, J.B., 2021. Scene perception based visual navigation of mobile robot in indoor environment. ISA transactions, 109, pp.389-400.

[33] López-Cifuentes, A., Escudero-Viñolo, M., Bescós, J. and García-Martín, Á., 2020. Semantic-aware scene recognition. Pattern Recognition, 102, p.107256.

[34] Ran, T., Yuan, L. and Zhang, J.B., 2021. Scene perception based visual navigation of mobile robot in indoor environment. ISA transactions, 109, pp.389-400.

[35] Cheng, X., Lu, J., Feng, J., Yuan, B. and Zhou, J., 2018. Scene recognition with objectness. Pattern Recognition, 74, pp.474-487.

[36] Ren, S., He, K., Girshick, R. and Sun, J., 2016. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence, 39(6), pp.1137-1149.

[37] Bolei Zhou, Agata Lapedriza, Antonio Torralba, Aude Oliva; Places: An Image Database for Deep Scene Understanding. Journal of Vision, 2017; 17(10):296.

[38] Mikolov, T., Chen, K., Corrado, G. and Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

**DESCRIPCIÓN DE LA INVENCIÓN**

La presente invención se refiere a un método implementado por ordenador y un sistema para reconocer de manera automática escenas de interior combinando las características globales y locales de una imagen tomada en esa escena, y un método implementado por ordenador y un sistema de recuperación automática de escenas de interior.

En comparación con los métodos descritos anteriormente, el método de la presente invención utiliza las frecuencias de los objetos de una escena para generar una descripción de la imagen utilizando las etiquetas de las clases de los objetos detectados. A continuación, la descripción se codifica en un vector que alimenta una red neuronal no profunda para la predicción de la escena. Además, se tiene en cuenta el contenido global de la escena para la predicción de esta, y se agregan tanto las predicciones locales (basadas en los objetos) como las globales utilizando una función de peso.

En esencia, el procedimiento y sistema propuesto en la presente invención mejora el reconocimiento de escenas en interiores sin necesidad de enormes conjuntos de entrenamiento y no necesita depender de redes muy profundas. Como resultado, conduce a tiempos de entrenamiento e inferencia muy eficientes tanto para el reconocimiento como para la recuperación de escenas.

El procedimiento y sistema de visión artificial para el reconocimiento automático de escenas de interior está basado en la descripción de los objetos y de la escena. Dada una imagen de una escena interior, el procedimiento asigna una categoría de la escena, como "dormitorio", "salón", “cocina", “habitación de los niños”, etc., en función del contenido de la imagen. Dicho procedimiento hace uso de un algoritmo de aprendizaje profundo para extraer las características de las imágenes y detectar los objetos con el fin de comprender el contexto de las imágenes.

El procedimiento para la clasificación automática de escenas comprende tres módulos. En el primero, a partir de una imagen de consulta de interior se detectan, clasifican y etiquetan los objetos de interés en imágenes. En un segundo módulo, se crea una leyenda de la imagen de consulta usando las etiquetas de los objetos. En un tercer módulo, se generan etiquetas de las escenas y características globales.

Usando una representación vectorizada de las leyendas de las imágenes de consulta se obtienen una primera propuesta de categorías de escena con probabilidades. Utilizando las etiquetas de las escenas y características globales se genera una segunda propuesta de categorías de escena con probabilidades. Finalmente, la primera y segunda propuesta de categorías de escena se combinan mediante una función de peso para obtener una categoría de la escena resultado.

Además, también se define un procedimiento para la recuperación automática de escenas en diferentes ángulos y condiciones que una imagen dada de la misma escena, que utiliza los mismos módulos. El procedimiento comprende detectar los objetos de una imagen de consulta y extraer sus características locales, además de extraer las características globales de la imagen de consulta; extraer las características locales de los objetos y las características globales de un conjunto de imágenes de referencia; calcular las correspondencias entre la imagen de consulta y las imágenes del conjunto de imágenes usando una métrica de similitud; y recuperar las *k* imágenes del conjunto de imágenes más relevantes.

El procedimiento y sistema para el reconocimiento de escenas de interior de la presente invención permite clasificar la escena de interior indicando la categoría de la escena contenida en la imagen y ordenar automáticamente mediante tecnología digital (ordenador) grandes repositorios de imágenes tomadas también mediante tecnología digital (cámara digital) en escenarios interiores bajo diferentes condiciones de captura -ángulos, iluminación, oclusiones, etc.- en función de la similitud de las imágenes del repositorio con respecto a una imagen dada.

La clasificación automática de escenas y la ordenación de estas frente a la manual por un experto anula la subjetividad, los errores por falta de atención y cansancio, la disparidad de criterio entre expertos, y los costes asociados al tiempo requerido por el experto, disminuye el tiempo necesario para la clasificación y ordenación, y aumenta la fiabilidad del sistema. Por este motivo, este procedimiento puede ser implementado en herramientas utilizadas por las empresas y FFCCSE (Fuerzas y Cuerpos de Seguridad del Estado) para realizar una clasificación de las escenas de interior en diferentes categorías que puede ser aplicada en diferentes ámbitos como la comprensión de escenas, la recuperación de escenas, la interacción persona-ordenador y la navegación robótica. Por ejemplo, puede ayudar a los discapacitados visuales a reconocer el entorno de interior en el que se encuentran. Además, este procedimiento puede ser implementado en herramientas utilizadas por las empresas y FFCCSE para realizar una ordenación de las imágenes de repositorios tomadas en escenas de interior en función de sus similitudes con respecto a una imagen dada, por ejemplo, para el análisis de pruebas forenses para la investigación de casos mediante la identificación de escenas relacionadas con un delito.

En la presente descripción, se emplea de manera general el término “imagen” para hacer referencia tanto a imágenes fijas (o fotografías) como a cada una de las imágenes presentes en una secuencia (o vídeo), independientemente de la fuente de origen (descargada de Internet, suministrado a través de un dispositivo de almacenamiento externo, etc.). Se emplea el término “recuperación” para referirse a la ordenación, al menos parcial, de repositorios de imágenes en función de la similitud de las imágenes del repositorio con respecto a una imagen de consulta, y el término “reconocimiento” para referirse tanto a la clasificación como a la recuperación de las escenas de interior.

En un ejemplo de realización, el procedimiento para el reconocimiento automático de escenas de interior utilizando un enfoque basado en la descripción de los objetos y de la escena se compone de tres módulos principales: módulo centrado en objetos, módulo centrado en la escena y módulo objetos a escena. Los dos primeros módulos toman como entrada de datos la descripción de la imagen de consulta generada por una red neuronal base.

En un ejemplo de realización, el módulo centrado en objetos utiliza la descripción generada por la red neuronal base y alimenta una red detectora de objetos y una red clasificadora de objetos, ambos entrenados de extremo a extremo para generar las etiquetas de los objetos y sus correspondientes características. La salida del módulo centrado en objetos se conecta al módulo objetos a escena, que obtiene una primera propuesta de categorías de escena a partir de una leyenda generada usando las etiquetas de los objetos. El módulo centrado en la escena también utiliza la descripción generada por la red neuronal base que alimenta a una red neuronal con varias capas de neuronas totalmente conectadas, que permiten obtener una segunda propuesta de categorías de escena.

En un ejemplo de realización, para obtener la categoría de la escena resultado de una imagen de consulta dada, se combinan la primera y segunda propuestas de categorías de escena del módulo objetos a escena y del módulo centrado en la escena utilizando un cálculo de puntuación.

En un ejemplo de realización, para obtener una recuperación de una imagen de consulta dada, los resultados obtenidos del módulo centrado en objetos y del módulo centrado en la escena se combinan utilizando otro cálculo de puntuación diferente.

De acuerdo a una realización, el procedimiento para la clasificación automática de escenas de interior utilizando un enfoque basado en la descripción de los objetos y de la escena comprende las siguientes etapas:

1. Obtención de la imagen de consulta. Esta obtención se puede realizar de un modo en línea, a través de un ordenador con conexión a Internet, o en un modo sin línea, obteniendo la imagen a través de un dispositivo de almacenamiento externo.

2. Alimentar una red neuronal base con la imagen de consulta para realizar la extracción de características. El contenido de la imagen de consulta se transforma en una serie de vectores incrustados, de modo que imágenes con escenas similares van a tener representaciones vectoriales similares, y viceversa, imágenes que contengan escenas diferentes van a tener representaciones vectoriales diferentes. En una realización preferente de la invención se una red neuronal convolucional VGG-16.

3. Detección de objetos. Las características extraídas de la red neuronal base se utilizan para entrenar una red neuronal del módulo centrado en objetos utilizada para la detección de objetos. En una realización, la red neuronal de detección de objetos utilizada es de extremo a extremo. En una realización, para la detección de objetos se utiliza una red de propuesta de regiones, RPN (del inglés “Region Proposal Network”) [36]. Utilizando dicha RPN se obtienen regiones de interés, ROI (del inglés “Region Of Interest”), y las características locales de cada objeto detectado dentro de la ROI.

4. Clasificación de objetos del módulo centrado en objetos. Las ROI y las características locales producidas por la red RPN se utilizan para entrenar una red neuronal del módulo centrado en objetos utilizada para la clasificación de objetos. En una realización preferente, la red neuronal de clasificación de objetos utilizada es de extremo a extremo. Utilizando dicha red diseñada, se obtienen las etiquetas de los objetos previamente detectados y sus probabilidades.

5. Creación de la leyenda de la imagen de consulta del módulo objetos a escena. Las etiquetas de los objetos extraídas del módulo centrado en objetos se tratan como palabras visuales que describen el contenido de la imagen. La idea se basa en el hecho de que escenas similares suelen compartir las mismas clases de objetos y las mismas o similares frecuencias de aparición, confiriendo a la futura leyenda una ventaja a la hora de

discriminar tipos de escenas. Por ejemplo, las imágenes de dormitorios contienen camas, y quizá una mesa y una silla. En cambio, aunque una escena que representa un laboratorio de informática también contiene algunos objetos de la misma clase (mesas y sillas), su frecuencia de aparición es mucho mayor. Utilizando dicha propuesta de leyendas, se obtiene una leyenda que incluye las etiquetas de los objetos extraídos e información de la frecuencia de aparición de los objetos dentro de la escena.

6. Representación vectorial de la leyenda del módulo objetos a escena. Las leyendas de las imágenes generadas son utilizadas para crear representaciones vectoriales de dichas leyendas.

7. Obtención de una primera propuesta de categorías de escena junto con sus probabilidades. A continuación, las representaciones vectorizadas de las leyendas se dan como entradas a una red neuronal, que obtendrá la primera propuesta de categorías de escenas.

8. Obtención de una segunda propuesta de categorías de escena del módulo centrado en la escena. Las características extraídas de la red neuronal base se utilizan para entrenar la red del módulo centrado en la escena utilizada para la obtención de categorías de escena. La red entrenada obtiene la segunda propuesta de categorías de escenas de interior y sus probabilidades. En una realización, la red neuronal da como resultado en su última capa características en forma de descriptores globales de la escena.

9. Cálculo de la función de peso WCOS. Para realizar el cálculo de la función de peso $WCOS_{SR}$ (del inglés “Weighted Combination of Objects and Scene”), se tienen en cuenta la primera y segunda propuestas de categorías de escena obtenidas por el módulo centrado en objetos, el módulo objetos a escena y el módulo centrado en la escena. En una realización, se tienen en cuenta las k categorías predichas con mayor probabilidad por el módulo objetos a escena y por el módulo centrado en la escena.

10. Decisión de la categoría de escena resultante. Una vez calculado el valor de peso $WCOS_{SR}$, se procede a la decisión de la categoría de la escena resultante. La selección se realiza asignando la categoría de escena resultante a aquella con una puntuación $WCOS_{SR}$ más alta.

De acuerdo a una realización, el procedimiento para la recuperación automática de escenas de interior utilizando un enfoque basado en la descripción de los objetos y de la escena comprende las siguientes etapas:

1. Obtención de la imagen de consulta. Esta obtención se puede realizar, por ejemplo, de un modo en línea, a través de un ordenador con conexión a internet, o en un modo sin línea, obteniendo la imagen a través de un dispositivo de almacenamiento externo.

2. Obtención de la categoría de la escena de la imagen de consulta (utilizando el procedimiento para la clasificación automática de escenas de interior previamente descrito) y comprobación de la categoría de escena de la imagen de consulta. Este es un paso preliminar opcional utilizado para comprobar si la categoría de escena de la imagen de la consulta coincide con alguna de las categorías de escena de unas imágenes de referencia (almacenadas por ejemplo en una base de datos), etiquetadas por dicho procedimiento. En caso de no haber imágenes etiquetadas de la misma categoría que la imagen de la consulta, el procedimiento para la recuperación se detiene y no devuelve ningún resultado. En caso de sí existir imágenes etiquetadas con la misma categoría, se continúa con el procedimiento,

3. Detección y clasificación de objetos. Se realiza la detección y clasificación de objetos usando el módulo centrado en objetos del método de clasificación automática de escenas de interior. Como resultado, se obtienen las etiquetas y las características locales de los objetos contenidos en la imagen de consulta.

4. Obtención de propuesta de categoría de escena. Se obtiene la propuesta de categoría de escena a través del módulo centrado en la escena previamente descrito en el método de clasificación automática de escenas de interior. Como resultado, se obtienen características globales de la imagen de consulta.

5. Creación de diccionario. Se crea un diccionario de la imagen de consulta, agrupando las características locales de los objetos, sus etiquetas y las características globales de la imagen.

6. Cálculo de la función de peso WCOS_Ret. Para realizar el cálculo de la función de peso $WCOS_{SRet}$ se tienen en cuenta que tiene en cuenta la similitud entre las características de

los objetos del diccionario de la imagen de consulta y de los diccionarios de las imágenes de referencia (almacenados preferentemente en una base de datos), y la similitud entre las características globales del diccionario de la imagen de consulta y de los diccionarios de las imágenes de referencia.

7. Recuperación de imágenes. Una vez calculado el valor de peso $WCOS_{SRet}$, se procede a la recuperación de imágenes. En una realización, se realiza la recuperación de las k imágenes de la base de datos que tienen una puntuación $WCOS_{SRet}$ más alta.

En una realización, el sistema de clasificación y de recuperación de escenas de interior comprende un ordenador. El sistema también puede comprender unos medios de almacenamiento de datos donde se almacenan las imágenes de escenas a analizar y los resultados de los cálculos intermedios.

Por último, la presente invención también se refiere a un producto de programa que comprende medios de instrucciones de programa para llevar a la práctica cualquiera de los procedimientos anteriormente descritos cuando el programa se ejecuta en un procesador. El producto de programa está preferentemente almacenado en un medio de soporte de programas. Los medios de instrucciones de programa pueden tener la forma de código fuente, código objeto, una fuente intermedia de código y código objeto, por ejemplo, como en forma parcialmente compilada, o en cualquier otra forma adecuada para uso en la puesta en práctica de los procesos según la invención.

El medio de soporte de programas puede ser cualquier entidad o dispositivo capaz de soportar el programa. Por ejemplo, el soporte podría incluir un medio de almacenamiento, como una memoria ROM, una memoria CD ROM o una memoria ROM de semiconductor, una memoria flash, un soporte de grabación magnética, por ejemplo, un disco duro o una memoria de estado sólido (SSD, del inglés solid-state drive). Además, los medios de instrucciones de programa almacenados en el soporte de programa pueden ser, por ejemplo, mediante una señal eléctrica u óptica que podría transportarse a través de cable eléctrico u óptico, por radio o por cualquier otro medio.

Cuando el producto de programa va incorporado en una señal que puede ser transportada directamente por un cable u otro dispositivo o medio, el soporte de programa puede estar constituido por dicho cable u otro dispositivo o medio.

Como variante, el soporte de programa puede ser un circuito integrado en el que va incluido el producto de programa, estando el circuito integrado adaptado para ejecutar, o para ser utilizado en la ejecución de los procesos correspondientes.

**BREVE DESCRIPCIÓN DE LOS DIBUJOS**

A continuación, se describen de manera muy breve una serie de figuras que ayudan a comprender mejor la invención y que se relacionan expresamente con una realización de dicha invención que se presenta como un ejemplo no limitativo de ésta.

La Figura 1 ilustra un método de clasificación de escenas de interior de acuerdo a una realización.

Las Figuras 2A y 2B representan diferentes realizaciones de un sistema de clasificación de escenas de interior de acuerdo a la presente invención.

La Figura 3 representa un ejemplo de realización del método de clasificación de escenas de interior aplicado a una imagen de entrada concreta.

La Figura 4 muestra un ejemplo de implementación de los diferentes módulos utilizados en el método de clasificación de escenas de interior.

La Figura 5 representa, de acuerdo a una realización, los módulos empleados en el método de clasificación y en el método de recuperación de escenas de interior.

La Figura 6 representa las etapas de un método de recuperación de escenas de interior de acuerdo a una realización.

La Figura 7 representa, de acuerdo a una posible realización, un sistema de recuperación de escenas de interior.

La Figura 8 muestra un ejemplo de realización del método de recuperación de escenas de interior aplicado a una imagen de entrada.

La Figura 9 ilustra un ejemplo de las medidas de similitud realizadas para la obtención de la

puntuación de similitud entre la imagen de entrada y una imagen de referencia.

**REALIZACIÓN PREFERENTE DE LA INVENCIÓN**

La **Figura 1** ilustra un método de clasificación de escenas de interior 100 de acuerdo a una realización. El método comprende las siguientes etapas:

- Recibir 110 una imagen de entrada 102.
- Detectar 120, mediante una red neuronal detectora de objetos, unos objetos 122 presentes en la imagen de entrada 102.
- Clasificar 130, mediante una red neuronal clasificadora de objetos, los objetos detectados, obteniendo unas etiquetas de objetos 132.
- Generar 140 una descripción 142 de la imagen de entrada que incluye las etiquetas de objetos 132 e información relativa a la frecuencia de aparición de las etiquetas de objetos en la imagen de entrada 102.
- Generar 150 una representación vectorial 152 de la descripción 142 de la imagen de entrada.
- Obtener 160, a partir de dicha representación vectorial 152, una primera predicción 162 de categorías de escena asignadas a la imagen de entrada 102 con sus probabilidades asociadas utilizando una primera red neuronal predictora de escenas.
- Obtener 170, a partir del contenido global de la imagen de entrada 102, una segunda predicción 172 de categorías de escena asignadas a la imagen de entrada 102 con sus probabilidades asociadas utilizando una segunda red neuronal predictora de escenas.
- Combinar 180 la primera 162 y segunda 172 predicción de categorías de escena mediante una primera función de peso para obtener una clasificación de categoría de escena 182 de la imagen de entrada 102.

En la **Figura 2A** se representa, de acuerdo a una posible realización, un sistema de clasificación de escenas de interior 200. El sistema comprende una unidad de procesamiento de datos 202 (e.g. un procesador, un ordenador, etc.) configurada para ejecutar el método de clasificación de escenas de interior 100, obteniendo una clasificación de categoría de escena 182 a partir de una imagen de entrada 102 recibida. La unidad de procesamiento de datos

202 puede estar distribuida en diferentes elementos hardware que, en conjunto y combinadas entre sí, lleven a cabo las diferentes etapas del método de clasificación.

El sistema de clasificación de escenas de interior 200 puede comprender medios de almacenamiento de datos (e.g. al menos una memoria 204 o un disco duro) configurados para almacenar datos utilizados durante el proceso de clasificación, incluyendo datos de entrada del proceso (i.e. la imagen de entrada 102) y/o datos de salida del proceso (i.e. la clasificación de categoría de escena 182). El sistema puede comprender un dispositivo de visualización 206 (e.g. una pantalla o monitor) configurado para mostrar al usuario la clasificación de categoría de escena 182 obtenida por la unidad de procesamiento de datos 202.

El sistema de clasificación de escenas de interior 200 de la presente invención se puede implementar en múltiples aplicaciones. Por ejemplo, tal y como se ilustra en la **Figura 2B**, el sistema puede implementarse en unas gafas inteligentes para invidentes que comprende una cámara 210 configurada para capturar una imagen de entrada 102, un procesador 212 configurado para ejecutar el método de clasificación de escenas de interior 100, que identifica la categoría de escena correspondiente a la imagen de entrada 102 e informa al usuario a través de unos medios de señalización, como por ejemplo un altavoz 220 incorporados en las propias gafas inteligentes, o a través del envío de una señal inalámbrica a otro dispositivo (e.g. señal Bluetooth enviada a un teléfono inteligente del usuario) para que notifique al usuario el tipo de categoría de escena clasificada por el sistema. De esta forma, cada vez que el usuario invidente entra en una estancia diferente de un edificio, es informado acerca del tipo de estancia de la que se trata (dormitorio, salón, despacho, sala de conferencias, etc.).

En la **Figura 3** se ilustra un ejemplo de realización del método de clasificación de escenas de interior 100 aplicado a una imagen de entrada 102 concreta, correspondiente en este caso a un dormitorio. La imagen de entrada 102 puede alimentar directamente la red neuronal detectora de objetos 310 y la segunda red neuronal predictora de escenas 340. Alternativamente, tal y como se muestra en línea discontinua en la Figura 3, la imagen de entrada 102 se puede proporcionar como entrada a una red neuronal convolucional base 302 encargada de generar características convolucionales de la imagen de entrada 102, de forma que dichas características convolucionales se alimentan por un lado a la red neuronal detectora de objetos 310 y por otro lado a la segunda red neuronal predictora de escenas 340.

La red neuronal detectora de objetos 310 se encarga de detectar los objetos 122 presentes

en la imagen de entrada 102. En una realización, la red neuronal detectora de objetos 310 se implementa mediante una red de propuesta de regiones (RPN, “Region Proposal Network”) que utiliza las características convolucionales generadas por la red neuronal convolucional base 302 para obtener una región de interés 312 asociada a cada objeto 122 detectado y unas características locales de cada objeto 122 detectado dentro de la correspondiente región de interés 312.

La red neuronal clasificadora de objetos 320 se encarga de recibir los objetos detectados por la red neuronal detectora de objetos 310 y clasificarlos, obteniendo así las etiquetas de objetos 132. En el ejemplo mostrado en la Figura 3 la red neuronal clasificadora de objetos 320 clasifica los objetos contenidos en las distintas regiones de interés 312 como “lámpara”, “cama”, “ventana” y “silla”, entre otras etiquetas de objetos 132.

A continuación, se genera una leyenda o descripción 142 de la imagen de entrada utilizando las etiquetas de objetos 132, las cuales son tratadas como palabras visuales que describen el contenido de la imagen de entrada 102. En la descripción 142 de la imagen de entrada también se incluye información relativa a la frecuencia de aparición (i.e. número de ocurrencias) de las etiquetas de objetos 132 en la imagen de entrada 102. En una realización, dicha información relativa a la frecuencia de aparición de las etiquetas de objetos 132 incluye unos cuantificadores seleccionados en función de la comparación del número de ocurrencias de las etiquetas de objetos con al menos un umbral de ocurrencias.

En una realización, se utilizan los siguientes cuantificadores en función del número de ocurrencias $N_{oc}$ de cada etiqueta de objeto 132 en la imagen de entrada 102:

- $N_{oc}=1$. Para las etiquetas que tienen una única ocurrencia o aparición en la imagen de entrada 102, no se incluye cuantificador (esto es, se incluye únicamente la etiqueta, sin cuantificador). Alternativamente, se puede incluir el cuantificador “un” o “una”, en función del género de la etiqueta, antecediendo a la etiqueta.
- $1< N_{oc} <TH_{oc}$. Para las etiquetas que tienen más de una ocurrencia y menos de un determinado umbral de ocurrencias $TH_{OC}$, se emplea el cuantificador “pocos” o ”pocas”, en función del género de la etiqueta, antecediendo a la etiqueta. La desventaja de tener cuantificadores diferentes en función del género de la etiqueta se puede solucionar si las etiquetas se expresan en

inglés; en ese caso se utilizaría un mismo cuantificador, “few”.

- $N_{oc}$ >$TH_{oc}$. Si la ocurrencia de una etiqueta es superior al umbral de ocurrencias $TH_{OC}$, se utiliza el cuantificador “muchos” o “muchas”, en función del género de la etiqueta, antecediendo a la etiqueta. De manera similar, si las etiquetas se expresan en inglés se utilizaría un único cuantificador, “many”.

En una realización, se selecciona el umbral de ocurrencias con un valor de 4 ($TH_{OC}$=4). De esta forma, si por ejemplo en la escena el clasificador reconoce 3 sillas, nueve mesas y un portátil, se genera la leyenda o descripción “pocas sillas, muchas mesas, portátil” (en inglés, “few chairs, many tables, laptop”). En el ejemplo de la Figura 3, se han detectado las siguientes etiquetas (recorriendo por ejemplo la imagen de entrada 102 de izquierda a derecha):

{*lámpara*, *cama*, *lámpara*, *silla*, *ventana*}

La descripción 142 de la imagen de entrada resultante sería la siguiente (las etiquetas podrían aparecer en otro orden):

*“pocas lámparas, cama, silla, ventana”*

Se genera a continuación una representación vectorial 152 de la descripción 142 de la imagen de entrada. En una realización, la representación vectorial 152 de la descripción 142 de la imagen de entrada se realiza mediante vectores incrustados de longitud fija, utilizando por ejemplo el método de incrustación Word2vec.

La primera red neuronal predictora de escenas 330 recibe la representación vectorial 152 y obtiene una primera predicción 162 de categorías de escena 332 asignadas a la imagen de entrada junto con sus probabilidades $P_{OC}$ asociadas. En el ejemplo de la Figura 3 la primera red neuronal predictora de escenas 330 obtiene una primera predicción 162, en la que se incluyen las 5 categorías de escena 332 con la mayor probabilidad $P_{OC}$, en particular las categorías de escena “dormitorio”, “habitación niños”, “cocina”, “salón” y “despacho” con respectivas probabilidades $P_{OC}$ de 85%, 12%, 1%, 0.5% y 0.4% (que acumulan un 98.9% de probabilidad). La primera predicción 162 incluye un conjunto de categorías de escena 332 que tienen la mayor probabilidad, donde el número de dicho conjunto puede ser un número fijo (e.g. 5) o un número variable en función de sus probabilidades $P_{OC}$ asociadas (por ejemplo, incluir categorías de escena cuya probabilidad acumulada supere un determinado umbral, e.g. 98%).

Por su parte, la segunda red neuronal predictora de escenas 340 está configurada para obtener, a partir del contenido global de la imagen de entrada (i.e. a partir de características globales de la imagen de entrada, basándose en el contexto global de la imagen de entrada), la segunda predicción 172 de categorías de escena asignadas a la imagen de entrada 102 junto con sus probabilidades asociadas.

En una realización, como en la mostrada en la Figura 3, la segunda red neuronal predictora de escenas 340 utiliza las características convolucionales generadas por la red neuronal convolucional base 302 para obtener la segunda predicción 172 de categorías de escena 342 y sus probabilidades $P_{OC}$ asociadas. En el ejemplo de la Figura 3 la segunda predicción 172 incluye las 5 categorías de escena 342 con la mayor probabilidad $P_{SC}$, en particular las categorías de escena “salón”, “bodega”, “dormitorio”, “museo” y “auditorio” con respectivas probabilidades $P_{SC}$ de 44%, 27%, 25%, 0.2% y 0.1% (acumulando un 96.3% de probabilidad). La segunda predicción 162 incluye, de manera similar a la primera predicción 162, un conjunto determinado de categorías de escena 342 que tienen la mayor probabilidad, donde el número de dicho conjunto puede ser un número fijo (e.g. 5) o un número variable en función de sus probabilidades $P_{SC}$ asociadas (por ejemplo, incluir categorías de escena cuya probabilidad acumulada supere un determinado umbral, e.g. 98%).

Finalmente, se combinan la primera predicción 162 de categorías de escena 332 y la segunda predicción 172 de categorías de escena 342 utilizando una primera función de peso 350 para obtener una clasificación de categoría de escena 182 de la imagen de entrada 102.

En una realización, se utiliza la primera función de peso 350 ($WCOS_{SR}$) para calcular 352, para un conjunto de categorías de escena 354 de la primera 162 y segunda 172 predicción (e.g. al menos un número *k* de categorías de escena (332,342) de la primera y segunda predicción con mayor probabilidad asociada), una puntuación 356 utilizando las respectivas probabilidades ($P_{OC}$,$P_{SC}$) ponderadas por unos respectivos pesos asociados, y se obtiene la clasificación de categoría de escena 182 de la imagen de entrada 102 seleccionando la categoría de escena con la puntuación 356 más alta.

En una realización, la primera función de peso 350 se implementa según la siguiente ecuación:

$$WCOS_{SR} = W_{OCi} * P_{OC} + W_{SCj} * P_{SC} \qquad (1)$$

En esta ecuación (1) la probabilidad de la primera predicción 162 de categorías de escena se representa por $P_{OC}$ (de acuerdo a un enfoque centrado en el objeto) y la probabilidad de la segunda predicción 172 de categorías de escena se representa por $P_{SC}$ (de acuerdo a un enfoque centrado en la escena global, no en los objetos). Cada una de estas probabilidades $P_{OC}$ y $P_{SC}$ tiene un respectivo peso asociado $W_{OCi}$ y $W_{SCj}$ en la primera función de peso 350. El peso $W_{OCi}$ representa el valor de la ponderación de la primera predicción 162 resultante de un enfoque centrado en el objeto, y el peso $W_{SCj}$ representa el valor de la ponderación de la segunda predicción 172 resultante de un enfoque centrado en la escena.

En una realización, los pesos $W_{OCi}$ y $W_{SCj}$ asociados a cada probabilidad $P_{OC}$ y $P_{SC}$, respectivamente, se calculan en función de la posición o rango $i$ y $j$ que ocupa la correspondiente probabilidad en la respectiva predicción.

Así, en la realización mostrada en la Figura 3 se considera una imagen de entrada 102 clasificada en una cierta categoría de escena con una probabilidad $P_{OC}$ en el rango o posición $i$ y con una probabilidad $P_{SC}$ en la posición $j$, donde se cumple que $i, j \in Z, 1 \leq i \leq 5, 1 \leq j \leq 5$. El valor de los pesos $W_{OCl}$ y $W_{SCl}$ de una determinada categoría de escena para una posición general $l$ se puede definir, por ejemplo, con la siguiente ecuación que depende del rango $l$ que ocupe dicha categoría de escena en la respectiva predicción:

$$W_{OCl} = W_{SCl} = \begin{cases} 1.1 - 0.1l \; if \; 1 \leq l \leq 5, l \in Z \\ 0 \qquad\qquad en \; otro \; caso \end{cases} \qquad (2)$$

Por ejemplo, en la Figura 3 la categoría de escena “salón" en la primera predicción 162 ocupa el rango $i$=4, por detrás de “dormitorio” (rango o posición 1), “habitación niños” (rango 2), “cocina” (rango 3), y por delante de “despacho” (rango 5). La misma categoría de escena ocupa el rango $j$=1 en la segunda predicción 172. Aplicando la ecuación (2), los pesos ($W_{OC4}$,$W_{SC1}$) asociados a la categoría de escena “salón” son los siguientes:

$$W_{OC4} = 1.1 - 0.1x4 = 0.7$$

$$W_{SC1} = 1.1 - 0.1x1 = 1$$

Aplicando la ecuación (1) se obtiene la siguiente puntuación 356 para la categoría salón:

$$WCOS_{SR} = W_{OC4}xP_{OC} + W_{SC1}xP_{SC} = 0.7x0.005 + 1x0.44 = 0.4435$$

Usando las ecuaciones (1) y (2) en el ejemplo de la Figura 3 se obtienen las diferentes puntuaciones 356 para las categorías consideradas, contenidas en el rango [0-2], tal y como se muestra en la figura. Sin embargo, se podrían utilizar ecuaciones diferentes. En este ejemplo se calculan las puntuaciones 356 para las categorías que tienen rango≤3 en ambas predicciones (162,172), esto es, “dormitorio”, “habitación niños” y “cocina” en la primera predicción 162 y “salón”, “bodega” y “dormitorio” en la segunda predicción 172, resultando las cinco puntuaciones mostradas en la Figura 3 al estar repetida la categoría de escena “dormitorio”. Alternativamente, se pueden realizar otros cálculos de puntuaciones; por ejemplo, se puede calcular las puntuaciones 356 para todas las categorías de escena incluidas en las dos predicciones (162,172).

La clasificación de categoría de escena 182 de la imagen de entrada 102 es aquella cuya puntuación 356 es mayor. En este caso se selecciona “dormitorio” como la clasificación de categoría de escena 182, al tener una puntuación de 1.05, superior a las puntuaciones del resto de categorías de escena.

La **Figura 4** ilustra una implementación de los diferentes módulos utilizados en el método de clasificación de escenas de interior 100, representando en detalle las redes neuronales empleadas en cada uno de dichos módulos de acuerdo a una realización. Como red neuronal convolucional base 302 se utiliza una red neuronal convolucional VGG-16 preentrenada con el conjunto de datos Places365 [37].

Un módulo centrado en objetos 410 incluye un detector de objetos y un clasificador de objetos, ambos implementados mediante redes neuronales (red neuronal detectora de objectos 310 y red neuronal clasificadora de objetos 320, respectivamente), y obtiene las etiquetas de objetos 132 y las características locales 412 de los objetos detectados.

El módulo centrado en objetos 410 comprende por tanto dos redes de extremo a extremo: detección de objetos y clasificación. En una realización, para detectar objetos se emplea una red de propuesta de regiones (RPN), y en concreto una red simple de tres capas que tiene una sola capa convolucional (capa de entrada) conectada a dos capas de salida, y que es entrenada utilizando por ejemplo el conjunto de datos Open Image. La primera capa de salida tiene 1xm neuronas de salida, donde m es el número de posibles elementos detectados (anclajes, del inglés “anchors”) y cada neurona proporciona un grado de objeto o valor de

objetividad (probabilidad de que sea un objeto) de cada uno de ellos. La otra capa de salida tiene un tamaño 4xm que representan los cuadros delimitadores (en inglés, “bounding boxes”) de los anclajes, definiendo las coordenadas de los elementos detectados y las puntuaciones de objetividad y, con ello, las propiedades locales de una escena interior. La capa de entrada acepta características convolucionales de la red neuronal convolucional base 302 y las dos capas de salida devuelven cuadros de anclaje y puntuaciones del grado de objeto. En otras palabras, este módulo actúa como un clasificador binario que determina la presencia de objetos junto con una puntuación de grado de objeto para cada objeto detectado. Los cuadros de anclaje de los objetos localizados se denominan regiones de interés (ROI).

Las ROI y las características producidas por la red RPN se utilizan además para entrenar, mediante aprendizaje supervisado, una red clasificadora de objetos que, en una realización, consta de una capa densa de 1024 neuronas conectadas con dos unidades de salida diferentes para generar las etiquetas de objetos y sus probabilidades. Las etiquetas de los objetos hacen referencia a la clase a la que pertenecen los objetos.

El siguiente módulo, módulo objetos a escena 420, se encarga de recibir la información relativa a los objetos en la imagen de entrada 102 procedente del módulo centrado en objetos 410 y convertirla en la primera predicción 162 de categorías de escena. El módulo objetos a escena 420 genera la descripción 142 de la imagen de entrada y representa dicha descripción en una representación vectorial 152. En una realización, las etiquetas incluidas en la descripción se transforman en palabras visuales utilizando la incrustación de palabras Word2vec [38]. El algoritmo Word2vec da como resultado una representación vectorial, también llamados vectores incrustados de longitud fija que representan la descripción. En una realización, se obtiene una representación vectorial de 390 elementos, valor seleccionado empíricamente para un mejor rendimiento.

Utilizando la descripción representada vectorialmente, se obtiene a continuación una primera propuesta de categorías de escena junto con sus probabilidades (primera predicción 162). En una realización, las representaciones vectorizadas de la descripción usando Word2Vec se dan como entradas a una red neuronal de tres capas (correspondiente a la primera red neuronal predictora de escenas 330), que dan como resultado la primera predicción 162 de categorías de escenas. En una realización, la red encargada de calcular la primera predicción de categorías de escena utiliza para el entrenamiento los conjuntos de datos de reconocimiento de escenas del MIT (MIT-67 Indoor) y de la Universidad de Nueva York (NYU).

Sin embargo, las etiquetas de los objetos son relevantes para realizar una primera propuesta de categorías, pero a medida que aumenta el número de clases aparecen más objetos comunes en diferentes clases de escenas. Como consecuencia, la precisión del modelo disminuye. Para abordar este problema, se combinan las predicciones obtenidas a partir el módulo centrado en objetos 410 (primera predicción 162) con unas predicciones obtenidas por un módulo centrado en la escena 430. De este modo, se superan las limitaciones individuales de cada uno de los módulos y se mejora la precisión de la predicción de la escena resultante. Esta decisión se fundamenta en el hecho de que usar solo el enfoque centrado en objetos puede no producir los resultados deseados cuando las escenas entre clases tienen objetos comunes y las características globales centradas en la escena son más genéricas debido a la presencia de un diseño similar en diferentes clases de imágenes.

La imagen de entrada 102 también se envía al módulo centrado en la escena 430 para realizar la segunda predicción 172 de categorías de escena utilizando una red neuronal entrenada con aprendizaje supervisado (segunda red neuronal predictora de escenas 340). En una realización, la red neuronal consiste en dos capas totalmente conectadas añadidas a la red neuronal base. Antes de poder utilizarse, la red neuronal ha de entrenarse. En una realización, el entrenamiento se realiza a través de aprendizaje por transferencia, utilizando por ejemplo los conjuntos de datos de reconocimiento de escenas del MIT (MIT-67 Indoor) y de la Universidad de Nueva York (NYU). Aplicando la función SoftMax 432 como una capa final del clasificador, se obtienen predicciones con valores finales en el rango [0-1] en la segunda predicción 172 de categorías de escena. Además, mientras que el módulo centrado en objetos 410 obtiene características locales 412 de los objetos de la imagen de entrada, el módulo centrado en la escena 430 obtiene características globales 434 de la imagen de entrada 102.

Las características locales 412 son vectores que describen una pequeña parte de la imagen, que contienen información relevante de los objetos dentro de la imagen. En cambio, las características globales 434 son vectores que describen toda la imagen, que contienen información relevante de la imagen completa.

La combinación de ambos tipos de información (local y global) confiere a la invención una ventaja técnica que se explica a continuación. El reconocimiento de escenas en interiores depende principalmente de la detección de objetos que se encuentran comúnmente en escenas interiores. Sin embargo, a veces debido a la presencia de objetos desordenados,

podrían no ser detectados, no generándose etiquetas para algunos objetos. Además, en e caso de que existan objetos similares en escenas diferentes, si solo se utilizan las características locales 412 de los objetos, dichas características podrían no ser lo suficientemente robustas para representar todas las características de la escena. En cambio, si también se extraen las características globales 434 de la imagen de entrada 102, extraídas del módulo centrado en la escena 430, y se combinan con las características locales 412, se genera una descripción de la imagen 352 suficientemente robusta, la cual permite compensar errores por situaciones como la descrita.

La presente invención también se refiere a un método de recuperación de escenas de interior. La **Figura 5** representa los tres módulos principales (410,420,430) empleados en el método de clasificación de escenas de interior para obtener la clasificación de la categoría de escena 182 de la imagen de entrada 102, utilizando una combinación ponderada con pesos ($W_1$), o primera función de peso 350, de las predicciones (162,172) obtenidas por el módulo objetos a escena 420 y por el módulo centrado en la escena 430.

El módulo centrado en objetos 410 detecta objetos y genera etiquetas de objetos y características locales, mientras que el módulo centrado en la escena 430 genera características globales y una predicción de categorías de escena basada en la escena global.

El método de recuperación de escenas de interior utiliza por su parte información obtenida por el módulo centrado en objetos 410 y por el módulo centrado en la escena 430 para obtener, mediante una combinación ponderada con pesos ($W_2$) o segunda función de peso 502, un conjunto de imágenes recuperadas 504, de entre un conjunto de imágenes de referencia almacenadas en una base de datos, que tienen mayor similitud con respecto a la imagen de entrada 102.

La **Figura 6** representa las etapas de un método de recuperación de escenas de interior 600 de acuerdo a una realización. El método comprende las siguientes etapas:

- Recibir 610 una imagen de entrada 102 (o imagen de consulta).
- Detectar 620, mediante una red neuronal detectora de objetos, unos objetos 122 presentes en la imagen de entrada 102.
- Clasificar 630, mediante una red neuronal clasificadora de objetos, los objetos 122 detectados, obteniendo unas etiquetas de objetos 132.

- Extraer 640 unas características locales 412 de cada objeto 122 detectado.
- Extraer 650 unas características globales 434 de la imagen de entrada 102 utilizando una red neuronal.
- Crear 660 un diccionario 662 de la imagen de entrada 102 que incluye las características locales 412 y la etiqueta 132 de cada objeto 122 detectado y las características globales 434 de la imagen de entrada 102.
- Recibir 670 una pluralidad de diccionarios 672 de unas imágenes de referencia, donde cada diccionario 672 incluye unas características locales y una etiqueta de cada objeto detectado en la imagen de referencia y unas características globales de la correspondiente imagen de referencia. Los diccionarios 672 de las imágenes de referencia se pueden obtener, por ejemplo, mediante el acceso a una base de datos almacenada en una memoria.
- Calcular 680 una puntuación de similitud 682 de cada imagen de referencia con respecto a la imagen de entrada 102, donde la puntuación de similitud 682 es obtenida en base a la similitud entre las características locales 412 de los objetos del diccionario 662 de la imagen de entrada 102 y del diccionario 672 de la correspondiente imagen de referencia que tienen asociada una misma etiqueta 132, y en base a la similitud entre las características globales 434 del diccionario 662 de la imagen de entrada 102 y del diccionario 672 de la correspondiente imagen de referencia.
- Recuperar 690 una o varias imágenes de referencia (imágenes recuperadas 504) con la puntuación de similitud más alta.

En la **Figura 7** se representa, de acuerdo a una posible realización, un sistema de recuperación de escenas de interior 700. El sistema comprende una unidad de procesamiento de datos 202 (por ejemplo, el mismo dispositivo empleado en el sistema de clasificación de escenas de interior 200) configurada para ejecutar el método de recuperación de escenas de interior 600, obteniendo una o varias imágenes recuperadas 504 a partir de una imagen de entrada 102 recibida. La unidad de procesamiento de datos 202 puede estar distribuida en diferentes elementos hardware (por ejemplo, diferentes procesadores) que, en conjunto y combinadas entre sí, lleven a cabo las diferentes etapas del método de recuperación.

El sistema de recuperación de escenas de interior 700 comprende preferentemente al menos una memoria 204 configurada para almacenar datos utilizados durante el proceso de

recuperación, incluyendo datos de entrada del proceso (i.e. los diccionarios 672 de las imágenes de referencia, que pueden estar almacenados en una base de datos). En la memoria 204 también se puede almacenar las propias imágenes de referencia 706, de forma que la unidad de procesamiento de datos 202 pueda recuperar directamente las imágenes de referencia 706 con mayor puntuación de similitud mediante el acceso a la memoria 204. Así mismo, el sistema puede comprender un dispositivo de visualización 206 (e.g. una pantalla o monitor) configurado para mostrar al usuario las imágenes recuperadas 504.

En la **Figura 8** se ilustra un ejemplo de realización del método de recuperación de escenas de interior 600 aplicado a una imagen de entrada 102. La imagen de entrada 102 puede alimentar directamente el módulo centrado en objetos 410 y el módulo centrado en la escena 430. Alternativamente, la imagen de entrada 102 se puede proporcionar como entrada a una red neuronal convolucional base 302 (representada en línea discontinua) encargada de generar características convolucionales de la imagen de entrada 102, de forma que dichas características convolucionales se alimentan al módulo centrado en objetos 410 y al módulo centrado en la escena 430.

En una realización, el método de recuperación de escenas de interior 600 comprende una comprobación previa que incluye determinar la clasificación de categoría de escena 182 de la imagen de entrada 102 utilizando el método de clasificación de escenas de interior 100, y comprobar si la categoría de escena de la imagen de entrada coincide con alguna de las categorías de escena de las imágenes de referencia, etiquetadas previamente por el procedimiento de clasificación de escenas de interior 100. Las categorías de escena de las imágenes de referencia están preferentemente almacenadas en memoria (e.g. en una base de datos). Si ninguna de las categorías de escena de las imágenes de referencia 706 coincide con la categoría de escena de la imagen de entrada 102 (e.g. dormitorio), el método de recuperación de escenas de interior 600 se detiene, de forma que no hay ninguna imagen recuperada 504. En caso de que sí existan imágenes de referencia 706 etiquetadas con la misma categoría que la imagen de entrada 102, el método de recuperación de escenas de interior 600 continua para obtener las imágenes de referencia con mayores similitudes con respecto a la imagen de entrada 102.

El módulo centrado en objetos 410 comprende una red neuronal detectora de objetos 310 y una red neuronal clasificadora de objetos 320, a partir de las cuales se obtienen las características locales 412 y las etiquetas 132 de los objetos detectados en la imagen de

entrada 102, de forma similar a como se ha explicado para el método de clasificación de escenas de interior 100. El módulo centrado en la escena 430 incluye una red neuronal encargada de analizar la imagen de entrada 102 y obtener, a partir del contenido global de la misma, unas categorías globales 434 de la imagen de entrada 102. A diferencia del método de clasificación de escenas de interior 100, el módulo centrado en la escena 430 no necesita en este caso obtener una segunda predicción 172 de categorías de escena asignadas a la imagen de entrada 102 con sus probabilidades asociadas.

Una vez obtenidas las etiquetas 132 y características locales 412 de los objetos y las características globales 434 de la imagen de entrada 102, se genera un diccionario 662 de la imagen de entrada 102 que incluye, para cada uno de los M objetos ($ID_1$, $ID_2$, …, $ID_M$) detectados y etiquetados, su correspondiente etiqueta 132 y sus características locales 412. El diccionario 662 incluye además las características globales 434 de la imagen de entrada 102. En el caso de que no se detecten objetos en la imagen de entrada 102, el diccionario contendría únicamente las características globales 434 de la imagen.

Por otro lado, se obtienen los diccionarios 672 de las imágenes de referencia. En la realización mostrada en la Figura 8, la unidad de procesamiento de datos 202 está configurada para recibir las P imágenes de referencia 706 ($IR_1$, $IR_2$, …, $IR_P$) y obtener, a partir de ella, los diccionarios 672 asociados a cada una de ellas, realizando un análisis similar al explicado para la imagen de entrada 102 para obtener sus etiquetas 132', características locales 412' y características globales 434'. Cada diccionario 672 de una imagen de referencia $IR_i$ tendrá un número determinado de objetos 122' detectados (en el ejemplo de la Figura 8 el diccionario de la imagen de referencia $IR_1$ tiene K objetos ($ID'_1$, $ID'_2$, …, $ID'_K$) detectados y etiquetados).

De manera ventajosa, los diccionarios 672 de las imágenes de referencia solo precisan obtenerse una única vez. Por ello, en una realización preferida estos diccionarios 672 están almacenados en una memoria 204, habiendo sido obtenidos en un proceso previo de manera similar a como se obtiene el diccionario 662 de la imagen de entrada 102. En este caso, el método de recuperación de escenas de interior 700 puede recibir u obtener los diccionarios 672 de las imágenes de referencia 706 accediendo a una base de datos almacenada en una memoria 204.

Para recuperar las imágenes similares, se obtienen puntuaciones ponderadas utilizando una métrica de similitud mediante una segunda función de peso 502 ($W_2$). En una realización, se

utiliza la similitud del coseno, aunque otras métricas de distancia o similitud podrían ser utilizadas.

En la **Figura 9** se ilustra un ejemplo de las medidas de similitud ($CS_1$,$CS_2$,$CS_3$,$CS_4$) realizadas para la obtención de la puntuación de similitud 682 entre la imagen de entrada y una imagen de referencia $IR_i$. Cada una de las características del objeto se indexa en función de las etiquetas. En primer lugar, se comparan las características locales 412' del objeto 122' en el diccionario 672 de la imagen de referencia $IR_i$ con respecto a las características locales 412 del objeto 122 en el diccionario 662 de la imagen de entrada 102 que tenga asociada la misma etiqueta (132,132') y, a continuación, se comparan las características globales 434' de la imagen de referencia $IR_i$ con las características globales 434 de la imagen de entrada 102. En cada comparación se obtiene una medida de similitud utilizando una métrica de similitud. Finalmente, se obtiene la puntuación de similitud 682 usando dichas medidas de similitud, por ejemplo mediante un promedio.

En el ejemplo de la Figura 9, se obtiene una medida de similitud $CS_1$ entre las características locales 412' del objeto $ID'_1$ del diccionario 672 de la imagen de referencia $IR_i$ y las características locales 412 del objeto $ID_1$ del diccionario 662 de la imagen de entrada 102, ya que ambos objetos tienen asociados una misma etiqueta (132,132') (i.e. "cama"). También se comparan las características locales 412 de los objetos etiquetados como "lámpara" (similitud entre el objeto $ID_2$ en el diccionario 662 de la imagen de entrada 102 con respecto a los objetos $ID'_2$ y $ID'_3$ en el diccionario 672 de la imagen de referencia $IR_i$), obteniendo las medidas de similitud $CS_2$ y $CS_3$. No se obtiene ninguna medida de similitud de los objetos "silla" y "cortina", ya que estas etiquetas no están incluidas simultáneamente en ambos diccionarios (662,672). Por último, se obtiene la medida de similitud $CS_4$ al comparar las características globales (434,434') de ambos diccionarios (662,672).

A continuación, se indica la ecuación de la métrica de similitud empleada para obtener la medida de similitud CS entre características locales 412 de ambos diccionarios:

$$CS = \frac{H_{qk} m_i}{||H_{qk}||\,||m_i||} = \frac{\sum_{j=1}^{N} H_{qkj} m_{ij}}{\sqrt{\sum_{j=1}^{N} H_{qkj}^2}\,\sqrt{\sum_{j=1}^{N} m_{ij}^2}}$$

Esta ecuación corresponde a la métrica de similitud del coseno (si bien se pueden utilizar otras métricas de similitud), donde N es la dimensión del descriptor de características, $H_{qk}$ es el

descriptor del objeto k-ésimo en la imagen de entrada y $m_{ij}$ es el descriptor del objeto j-ésimo con la misma etiqueta en el i-ésimo diccionario de imágenes de referencia 672 (diccionario $IR_i$).

En una realización, las medidas de similitud ($CS_1$, $CS_2$, $CS_3$) entre las características locales de los objetos se utilizan para calcular $ObjectBinsAvg_{score}$, que es la similitud de coseno promedio de las características locales de los objetos del diccionario 662 de la imagen de entrada 102 con respecto a las características locales de los objetos de un diccionario 672 de una imagen de referencia 706 que tienen etiquetas coincidentes:

$$ObjectBinsAvg_{score} = (CS_1 + CS_2 + CS_3)/3$$

Después, se calcula la medida de similitud $CS_4$ (e.g. similitud del coseno) entre las características globales (434,434') para obtener $GF_{score}$:

$$GF_{score} = CS_4$$

Por último, se combina $ObjectBinsAvg_{score}$ y $GF_{score}$ para calcular la puntuación de similitud 682 ($WCOS_{SRet}$), como una combinación ponderada de características de objetos y escenas, donde se pueden incluir pesos específicos diferentes para cada término:

$$WCOS_{SRet} = (ObjectBinsAvg_{score} + GF_{score})/2$$

Las imágenes se recuperan en función de la puntuación de similitud $WCOS_{SRet}$ más alta. Por ejemplo, se pueden recuperar las k imágenes que tengan las puntuaciones más altas.

## REIVINDICACIONES

1. Un método de clasificación de escenas de interior (100), caracterizado por que comprende:

- recibir (110) una imagen de entrada (102);
- detectar (120), mediante una red neuronal detectora de objetos (310), unos objetos (122) presentes en la imagen de entrada (102);
- clasificar (130), mediante una red neuronal clasificadora de objetos (320), los objetos detectados, obteniendo unas etiquetas de objetos (132);
- generar (140) una descripción (142) de la imagen de entrada (102) que incluye las etiquetas de objetos (132) e información relativa a la frecuencia de aparición de las etiquetas de objetos en la imagen de entrada (102);
- generar (150) una representación vectorial (152) de la descripción (142) de la imagen de entrada (102);
- obtener (160), a partir de dicha representación vectorial (152), una primera predicción (162) de categorías de escena asignadas a la imagen de entrada (102) con sus probabilidades ($P_{OC}$) asociadas utilizando una primera red neuronal predictora de escenas (330);
- obtener (170), a partir del contenido global de la imagen de entrada (102), una segunda predicción (172) de categorías de escena asignadas a la imagen de entrada (102) con sus probabilidades ($P_{SC}$) asociadas utilizando una segunda red neuronal predictora de escenas (340); y
- combinar (180) la primera (162) y segunda (172) predicción de categorías de escena mediante una primera función de peso (350) para obtener una clasificación de categoría de escena (182) de la imagen de entrada (102).

2. El método según la reivindicación 1, donde combinar (180) la primera (162) y segunda (172) predicción de categorías de escena mediante una primera función de peso (350) comprende:

- calcular (352), para un conjunto de categorías de escena (354) de la primera (162) y segunda (172) predicción, una puntuación (356) utilizando las respectivas probabilidades ($P_{OC}$,$P_{SC}$) ponderadas por unos respectivos pesos ($W_{OCi}$, $W_{SCj}$) asociados; y

- obtener la clasificación de categoría de escena (182) de la imagen de entrada (102) mediante la selección de la categoría de escena con la puntuación (356) más alta.

3. El método según la reivindicación 2, donde los pesos ($W_{OCi}, W_{SCj}$) asociados a cada probabilidad ($P_{OC}$,$P_{SC}$) se calculan en función del rango ($i$, $j$) que ocupa la correspondiente probabilidad en la respectiva predicción.

4. El método según cualquiera de las reivindicaciones anteriores, donde la imagen de entrada (102) se proporciona como entrada a una red neuronal convolucional base (302) encargada de generar características convolucionales de la imagen de entrada (102).

5. El método según la reivindicación 4, donde la red neuronal detectora de objetos (310) es una red de propuesta de regiones, RPN, que utiliza las características convolucionales generadas por la red neuronal convolucional base (302) para obtener una región de interés (312) asociada a cada objeto (122) detectado y unas características locales (412) de cada objeto (122) detectado dentro de la correspondiente región de interés (312).

6. El método según cualquiera de las reivindicaciones 4 a 5, donde la segunda red neuronal predictora de escenas (340) utiliza las características convolucionales generadas por la red neuronal convolucional base (302) para obtener la segunda predicción (172) de categorías de escena.

7. El método según cualquiera de las reivindicaciones anteriores, donde la información relativa a la frecuencia de aparición de las etiquetas de objetos (132) en la descripción (142) de la imagen de entrada (102) incluye unos cuantificadores seleccionados en función de la comparación del número de ocurrencias ($N_{oc}$) de las etiquetas de objetos (132) con al menos un umbral de ocurrencias ($TH_{oc}$).

8. El método según cualquiera de las reivindicaciones anteriores, donde la representación vectorial (152) de la descripción (142) de la imagen de entrada (102) se realiza mediante vectores incrustados de longitud fija.

9. Un sistema de clasificación de escenas de interior (200), caracterizado por que comprende una unidad de procesamiento de datos (202,212) configurada para:

- recibir (110) una imagen de entrada (102);
- detectar (120), mediante una red neuronal detectora de objetos (310), unos objetos (122) presentes en la imagen de entrada (102);
- clasificar (130), mediante una red neuronal clasificadora de objetos (320), los objetos detectados, obteniendo unas etiquetas de objetos (132);
- generar (140) una descripción (142) de la imagen de entrada (102) que incluye las etiquetas de objetos (132) e información relativa a la frecuencia de aparición de las etiquetas de objetos en la imagen de entrada (102);
- generar (150) una representación vectorial (152) de la descripción (142) de la imagen de entrada (102);
- obtener (160), a partir de dicha representación vectorial (152), una primera predicción (162) de categorías de escena asignadas a la imagen de entrada con sus probabilidades ($P_{OC}$) asociadas utilizando una primera red neuronal predictora de escenas (330);
- obtener (170), a partir del contenido global de la imagen de entrada (102), una segunda predicción (172) de categorías de escena asignadas a la imagen de entrada (102) con sus probabilidades ($P_{SC}$) asociadas utilizando una segunda red neuronal predictora de escenas (340); y
- combinar (180) la primera (162) y segunda (172) predicción de categorías de escena mediante una primera función de peso (350) para obtener una clasificación de categoría de escena (182) de la imagen de entrada (102).

10. El sistema según la reivindicación 9, donde para combinar (180) la primera (162) y segunda (172) predicción de categorías de escena mediante una primera función de peso (350) la unidad de procesamiento de datos (202,212) está configurada para:

- calcular (352), para un conjunto de categorías de escena (354) de la primera (162) y segunda (172) predicción, una puntuación (356) utilizando las respectivas probabilidades ($P_{OC}$,$P_{SC}$) ponderadas por unos respectivos pesos ($W_{OCi}, W_{SCj}$) asociados; y

- obtener la clasificación de categoría de escena (182) de la imagen de entrada (102) mediante la selección de la categoría con la puntuación (356) más alta.

11. El sistema según la reivindicación 10, donde la unidad de procesamiento de datos (202,212) está configurada para calcular los pesos ($W_{OCi}, W_{SCj}$) asociados a cada probabilidad ($P_{OC}$,$P_{SC}$) en función del rango ($i, j$) que ocupa la correspondiente probabilidad en la respectiva predicción.

12. El sistema según cualquiera de las reivindicaciones 9 a 11, donde la unidad de procesamiento de datos (202,212) está configurada para procesar la imagen de entrada (102) recibida mediante una red neuronal convolucional base (302) encargada de generar características convolucionales de la imagen de entrada (102).

13. El sistema según la reivindicación 12, donde la red neuronal detectora de objetos (310) es una red de propuesta de regiones, RPN, que utiliza las características convolucionales generadas por la red neuronal convolucional base (302) para obtener una región de interés (312) asociada a cada objeto (122) detectado y unas características locales (412) de cada objeto (122) detectado dentro de la correspondiente región de interés (312).

14. El sistema según cualquiera de las reivindicaciones 12 a 13, donde la segunda red neuronal predictora de escenas (340) está encargada de recibir y procesar las características convolucionales generadas por la red neuronal convolucional base (302) para obtener la segunda predicción (172) de categorías de escena.

15. El sistema según cualquiera de las reivindicaciones 9 a 14, donde la información relativa a la frecuencia de aparición de las etiquetas de objetos (132) en la descripción (142) de la imagen de entrada (102) incluye unos cuantificadores seleccionados en función de la comparación del número de ocurrencias ($N_{oc}$) de las etiquetas de objetos con al menos un umbral de ocurrencias ($TH_{oc}$).

16. El sistema según cualquiera de las reivindicaciones 9 a 15, donde la representación vectorial (152) de la descripción (142) de la imagen de entrada (102) se realiza mediante vectores incrustados de longitud fija.

17. Un producto de programa que comprende medios de instrucciones de programa para llevar a cabo el método de clasificación de escenas de interior (100) definido en cualquiera de las reivindicaciones 1 a 8 cuando el programa se ejecuta en un procesador.

18. Un medio de soporte de programas, que almacena el producto de programa según la reivindicación 17.

19. Un método de recuperación de escenas de interior (600), caracterizado por que comprende:

- recibir (610) una imagen de entrada (102);
- detectar (620), mediante una red neuronal detectora de objetos (310), unos objetos (122) presentes en la imagen de entrada (102).
- clasificar (630), mediante una red neuronal clasificadora de objetos (320), los objetos (122) detectados, obteniendo unas etiquetas de objetos (132);
- extraer (640) unas características locales (412) de cada objeto (122) detectado;
- extraer (650) unas características globales (434) de la imagen de entrada (102) utilizando una red neuronal;
- crear (660) un diccionario (662) de la imagen de entrada (102) que incluye las características locales (412) y la etiqueta (132) de cada objeto (122) detectado y las características globales (434) de la imagen de entrada (102);
- recibir (670), una pluralidad de diccionarios (672) de unas imágenes de referencia (706), donde cada diccionario (672) incluye unas características locales (412') y una etiqueta (132') de cada objeto (122') y unas características globales (434') de la correspondiente imagen de referencia (706);
- calcular (680) una puntuación de similitud (682) de cada imagen de referencia (706) con respecto a la imagen de entrada (102), donde la puntuación de similitud (682) es obtenida en base a la similitud entre las características locales (412,412') de los objetos (122,122') del diccionario (662) de la imagen de entrada (102) y del diccionario (672) de la correspondiente imagen de referencia (706) que tienen asociada una misma etiqueta (132,132'), y en base a la similitud entre las características globales (434,434') del diccionario (662) de la imagen de entrada (102) y del diccionario (672) de la correspondiente imagen de referencia (706); y

- recuperar (690) al menos una imagen de referencia (504) con la puntuación de similitud (682) más alta.

20. El método según la reivindicación 19, donde las características locales (412,412') de los objetos (122,122') están incluidas en un vector de características de longitud fija.

21. El método según cualquiera de las reivindicaciones 19 a 20, donde la puntuación de similitud (682) se calcula usando una métrica de similitud de coseno.

22. Un sistema de recuperación de escenas de interior (700), caracterizado por que comprende una unidad de procesamiento de datos (202) configurada para:

- recibir (610) una imagen de entrada (102);
- detectar (620), mediante una red neuronal detectora de objetos (310), unos objetos (122) presentes en la imagen de entrada (102);
- clasificar (630), mediante una red neuronal clasificadora de objetos (302), los objetos (122) detectados, obteniendo unas etiquetas de objetos (132);
- extraer (640) unas características locales (412) de cada objeto (122) detectado;
- extraer (650) unas características globales (434) de la imagen de entrada (102) utilizando una red neuronal;
- crear (660) un diccionario (662) de la imagen de entrada (102) que incluye las características locales (412) y la etiqueta (132) de cada objeto (122) detectado y las características globales (434) de la imagen de entrada (102);
- recibir (670) una pluralidad de diccionarios (672) de unas imágenes de referencia (706), donde cada diccionario (672) incluye unas características locales (412') y una etiqueta (132') de cada objeto (122') y unas características globales (434') de la correspondiente imagen de referencia (706);
- calcular (680) una puntuación de similitud (682) de cada imagen de referencia (706) con respecto a la imagen de entrada (102), donde la puntuación de similitud (682) es obtenida en base a la similitud entre las características locales (412,412') de los objetos (122,122') del diccionario (622) de la imagen de entrada (102) y del diccionario (672) de la correspondiente imagen de referencia (706) que tienen asociada una misma etiqueta (132,132'), y en base a la similitud entre las características globales

(434,434') del diccionario (662) de la imagen de entrada (102) y del diccionario (672) de la correspondiente imagen de referencia (706); y

- recuperar (690) al menos una imagen de referencia (504) con la puntuación de similitud (682) más alta.

23. El sistema según la reivindicación 22, que comprende al menos una memoria (204) en la que se almacena las imágenes de referencia (706) y los diccionarios (672) de las imágenes de referencia (706).

24. El sistema según cualquiera de las reivindicaciones 22 a 23, donde la unidad de procesamiento de datos (202) está configurada para incluir las características locales (412,412') de los objetos (122,122') en un vector de características de longitud fija.

25. El sistema según cualquiera de las reivindicaciones 22 a 24, donde la unidad de procesamiento de datos (202) está configurada para calcular (680) la puntuación de similitud (682) usando una métrica de similitud de coseno.

26. Un producto de programa que comprende medios de instrucciones de programa para llevar a cabo el método definido en cualquiera de las reivindicaciones 19 a 21 cuando el programa se ejecuta en un procesador.

27. Un medio de soporte de programas, que almacena el producto de programa según la reivindicación 26.

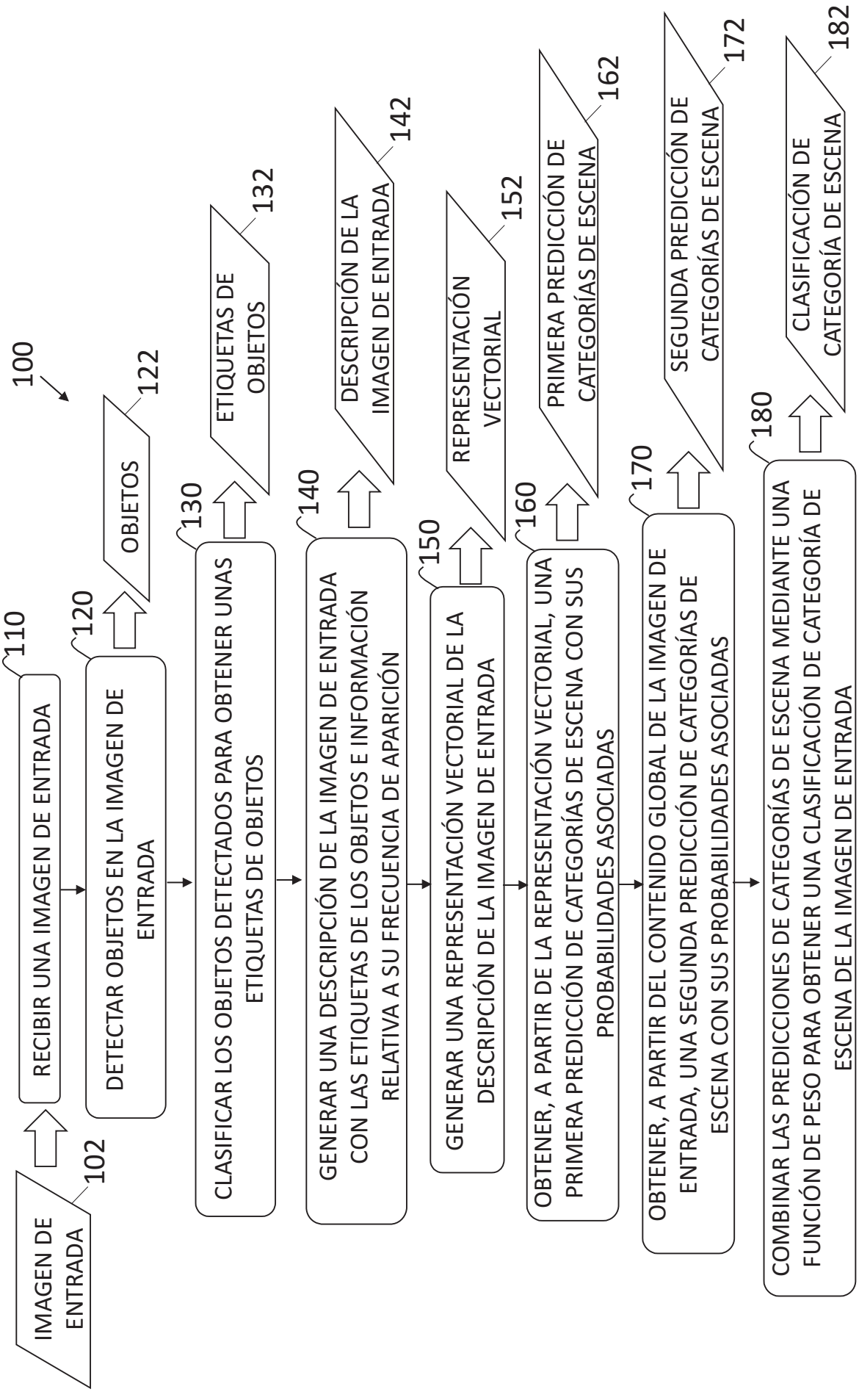
100
IMAGEN DE ENTRADA
102
RECIBIR UNA IMAGEN DE ENTRADA
110
DETECTAR OBJETOS EN LA IMAGEN DE ENTRADA
120
OBJETOS
122
CLASIFICAR LOS OBJETOS DETECTADOS PARA OBTENER UNAS ETIQUETAS DE OBJETOS
130
ETIQUETAS DE OBJETOS
132
GENERAR UNA DESCRIPCIÓN DE LA IMAGEN DE ENTRADA CON LAS ETIQUETAS DE LOS OBJETOS E INFORMACIÓN RELATIVA A SU FRECUENCIA DE APARICIÓN
140
DESCRIPCIÓN DE LA IMAGEN DE ENTRADA
142
GENERAR UNA REPRESENTACIÓN VECTORIAL DE LA DESCRIPCIÓN DE LA IMAGEN DE ENTRADA
150
REPRESENTACIÓN VECTORIAL
152
OBTENER, A PARTIR DE LA REPRESENTACIÓN VECTORIAL, UNA PRIMERA PREDICCIÓN DE CATEGORÍAS DE ESCENA CON SUS PROBABILIDADES ASOCIADAS
160
PRIMERA PREDICCIÓN DE CATEGORÍAS DE ESCENA
162
OBTENER, A PARTIR DEL CONTENIDO GLOBAL DE LA IMAGEN DE ENTRADA, UNA SEGUNDA PREDICCIÓN DE CATEGORÍAS DE ESCENA CON SUS PROBABILIDADES ASOCIADAS
170
SEGUNDA PREDICCIÓN DE CATEGORÍAS DE ESCENA
172
COMBINAR LAS PREDICCIONES DE CATEGORÍAS DE ESCENA MEDIANTE UNA FUNCIÓN DE PESO PARA OBTENER UNA CLASIFICACIÓN DE CATEGORÍA DE ESCENA DE LA IMAGEN DE ENTRADA
180
CLASIFICACIÓN DE CATEGORÍA DE ESCENA
182

FIG. 1

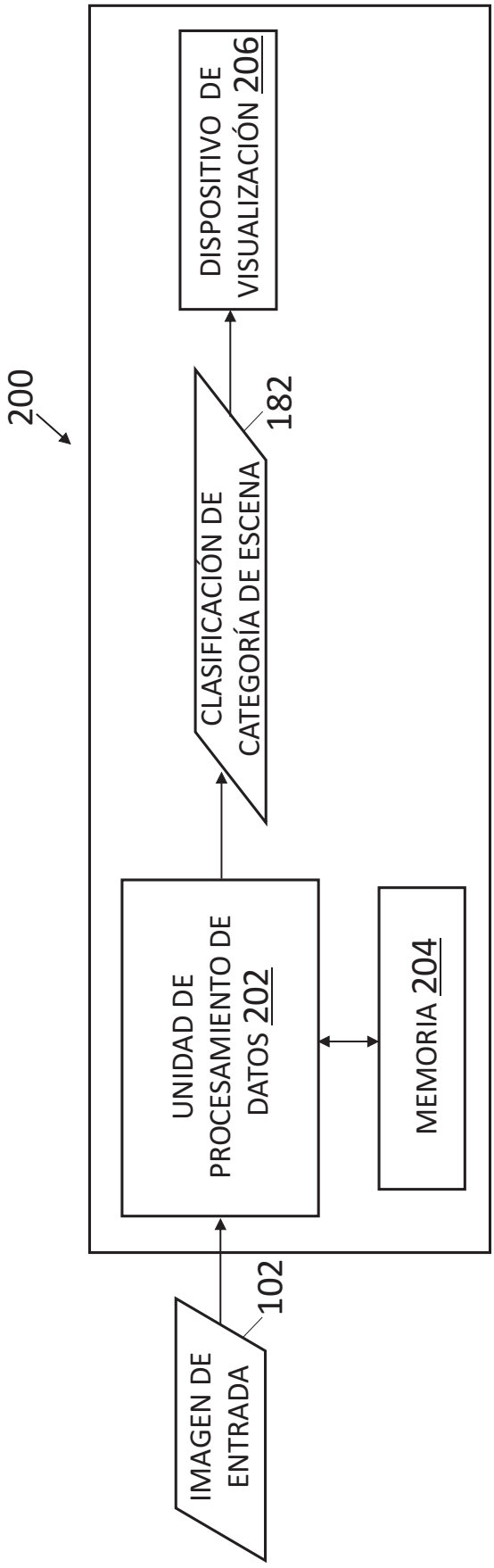
200
IMAGEN DE ENTRADA
102
UNIDAD DE PROCESAMIENTO DE DATOS 202
MEMORIA 204
CLASIFICACIÓN DE CATEGORÍA DE ESCENA
182
DISPOSITIVO DE VISUALIZACIÓN 206

FIG. 2A

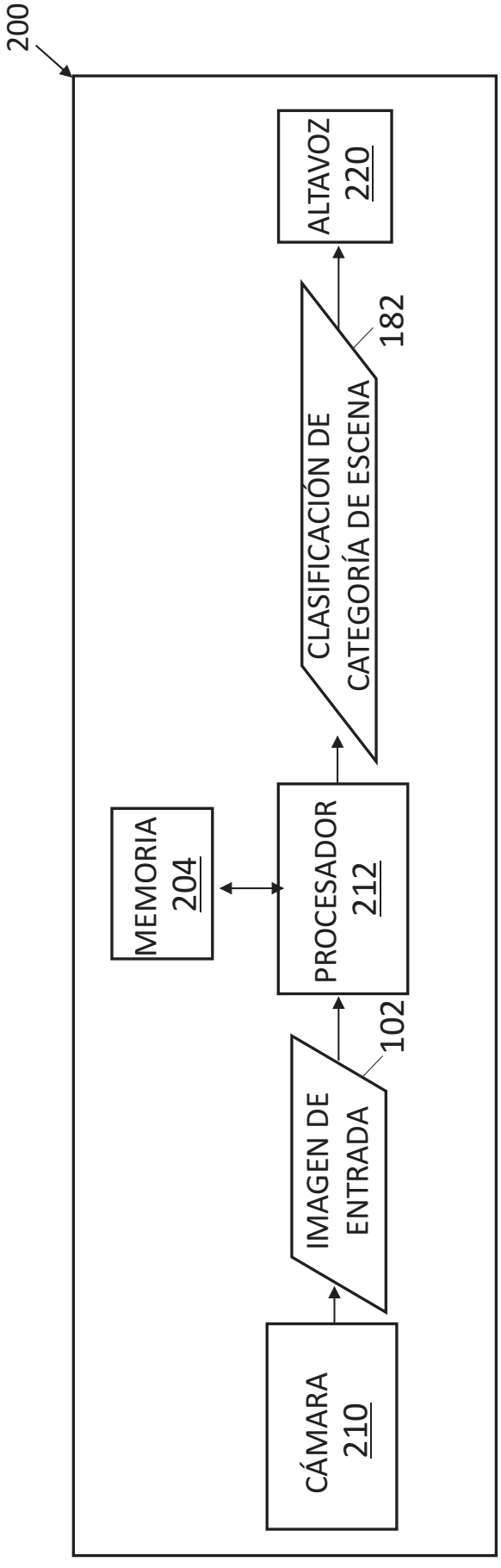
200
ALTAVOZ 220
182
CLASIFICACIÓN DE CATEGORÍA DE ESCENA
MEMORIA 204
PROCESADOR 212
102
IMAGEN DE ENTRADA
CÁMARA 210

FIG. 2B

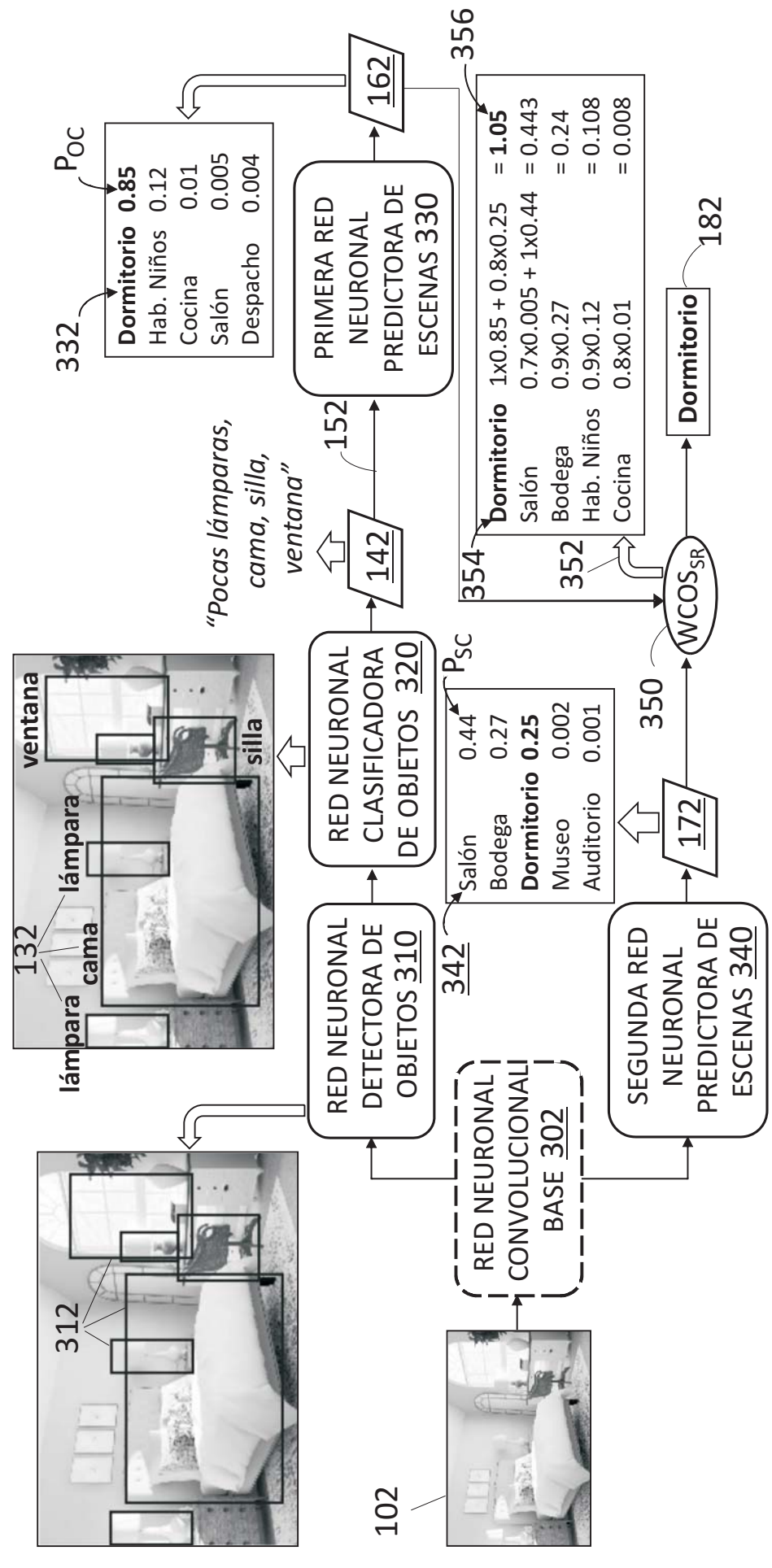
102
312
RED NEURONAL CONVOLUCIONAL BASE 302
RED NEURONAL DETECTORA DE OBJETOS 310
342
Salón 0.44
Bodega 0.27
Dormitorio 0.25
Museo 0.002
Auditorio 0.001
$P_{SC}$
SEGUNDA RED NEURONAL PREDICTORA DE ESCENAS 340
172
350
132
lámpara
cama
lámpara
ventana
silla
RED NEURONAL CLASIFICADORA DE OBJETOS 320
142
152
"Pocas lámparas, cama, silla, ventana"
PRIMERA RED NEURONAL PREDICTORA DE ESCENAS 330
162
$P_{OC}$
332
Dormitorio 0.85
Hab. Niños 0.12
Cocina 0.01
Salón 0.005
Despacho 0.004
356
354
Dormitorio 1x0.85 + 0.8x0.25 = 1.05
Salón 0.7x0.005 + 1x0.44 = 0.443
Bodega 0.9x0.27 = 0.24
Hab. Niños 0.9x0.12 = 0.108
Cocina 0.8x0.01 = 0.008
352
$WCOS_{SR}$
182
Dormitorio

FIG. 3

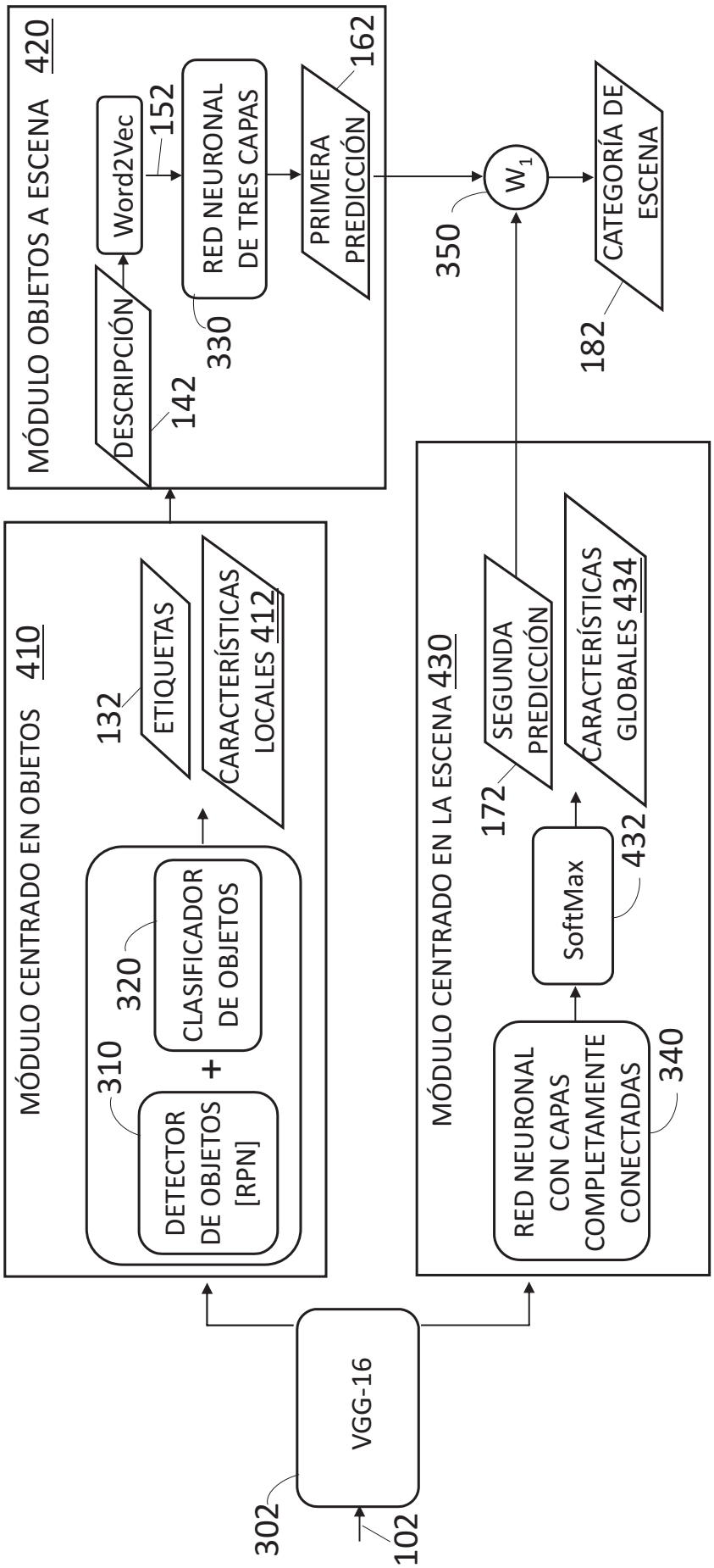
MÓDULO OBJETOS A ESCENA 420
Word2Vec
152
DESCRIPCIÓN
142
RED NEURONAL DE TRES CAPAS
330
PRIMERA PREDICCIÓN
162
350
$W_1$
CATEGORÍA DE ESCENA
182
MÓDULO CENTRADO EN OBJETOS 410
132
ETIQUETAS
CARACTERÍSTICAS LOCALES 412
310
DETECTOR DE OBJETOS [RPN]
+
320
CLASIFICADOR DE OBJETOS
MÓDULO CENTRADO EN LA ESCENA 430
172
SEGUNDA PREDICCIÓN
CARACTERÍSTICAS GLOBALES 434
SoftMax
432
RED NEURONAL CON CAPAS COMPLETAMENTE CONECTADAS
340
VGG-16
302
102

FIG. 4

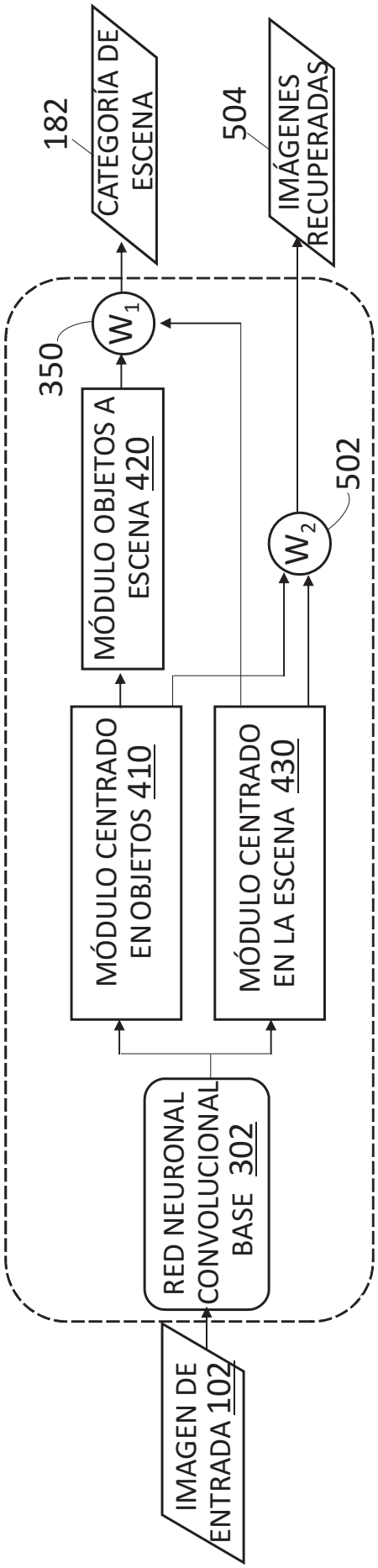
IMAGEN DE ENTRADA 102
RED NEURONAL CONVOLUCIONAL BASE 302
MÓDULO CENTRADO EN OBJETOS 410
MÓDULO CENTRADO EN LA ESCENA 430
MÓDULO OBJETOS A ESCENA 420
350
$W_1$
502
$W_2$
182
CATEGORÍA DE ESCENA
504
IMÁGENES RECUPERADAS

FIG. 5

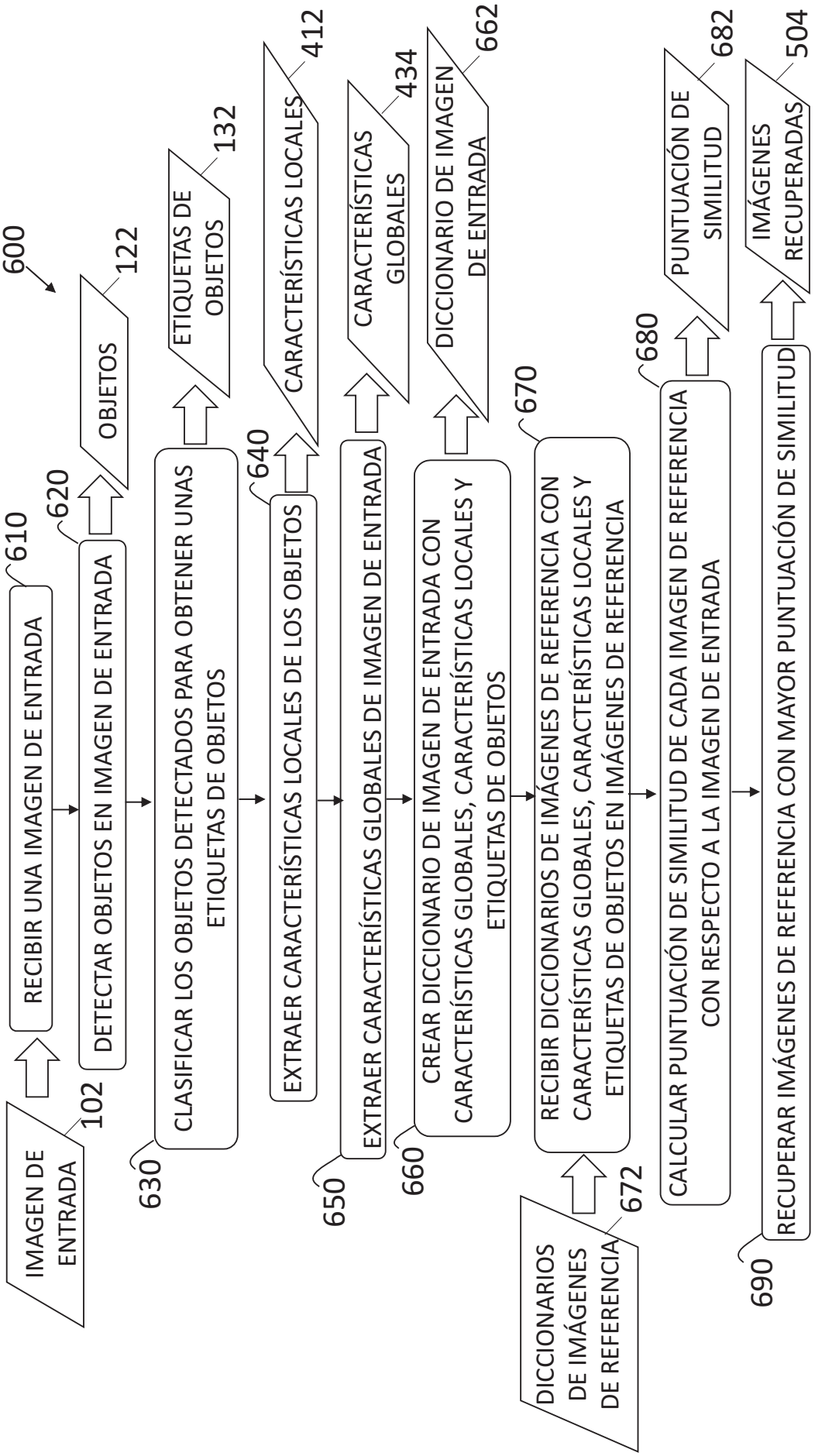
600
IMAGEN DE ENTRADA
102
RECIBIR UNA IMAGEN DE ENTRADA
610
DETECTAR OBJETOS EN IMAGEN DE ENTRADA
620
OBJETOS
122
CLASIFICAR LOS OBJETOS DETECTADOS PARA OBTENER UNAS ETIQUETAS DE OBJETOS
630
ETIQUETAS DE OBJETOS
132
EXTRAER CARACTERÍSTICAS LOCALES DE LOS OBJETOS
640
CARACTERÍSTICAS LOCALES
412
EXTRAER CARACTERÍSTICAS GLOBALES DE IMAGEN DE ENTRADA
650
CARACTERÍSTICAS GLOBALES
434
CREAR DICCIONARIO DE IMAGEN DE ENTRADA CON CARACTERÍSTICAS GLOBALES, CARACTERÍSTICAS LOCALES Y ETIQUETAS DE OBJETOS
660
DICCIONARIO DE IMAGEN DE ENTRADA
662
DICCIONARIOS DE IMÁGENES DE REFERENCIA
672
RECIBIR DICCIONARIOS DE IMÁGENES DE REFERENCIA CON CARACTERÍSTICAS GLOBALES, CARACTERÍSTICAS LOCALES Y ETIQUETAS DE OBJETOS EN IMÁGENES DE REFERENCIA
670
CALCULAR PUNTUACIÓN DE SIMILITUD DE CADA IMAGEN DE REFERENCIA CON RESPECTO A LA IMAGEN DE ENTRADA
680
PUNTUACIÓN DE SIMILITUD
682
RECUPERAR IMÁGENES DE REFERENCIA CON MAYOR PUNTUACIÓN DE SIMILITUD
690
IMÁGENES RECUPERADAS
504

FIG. 6

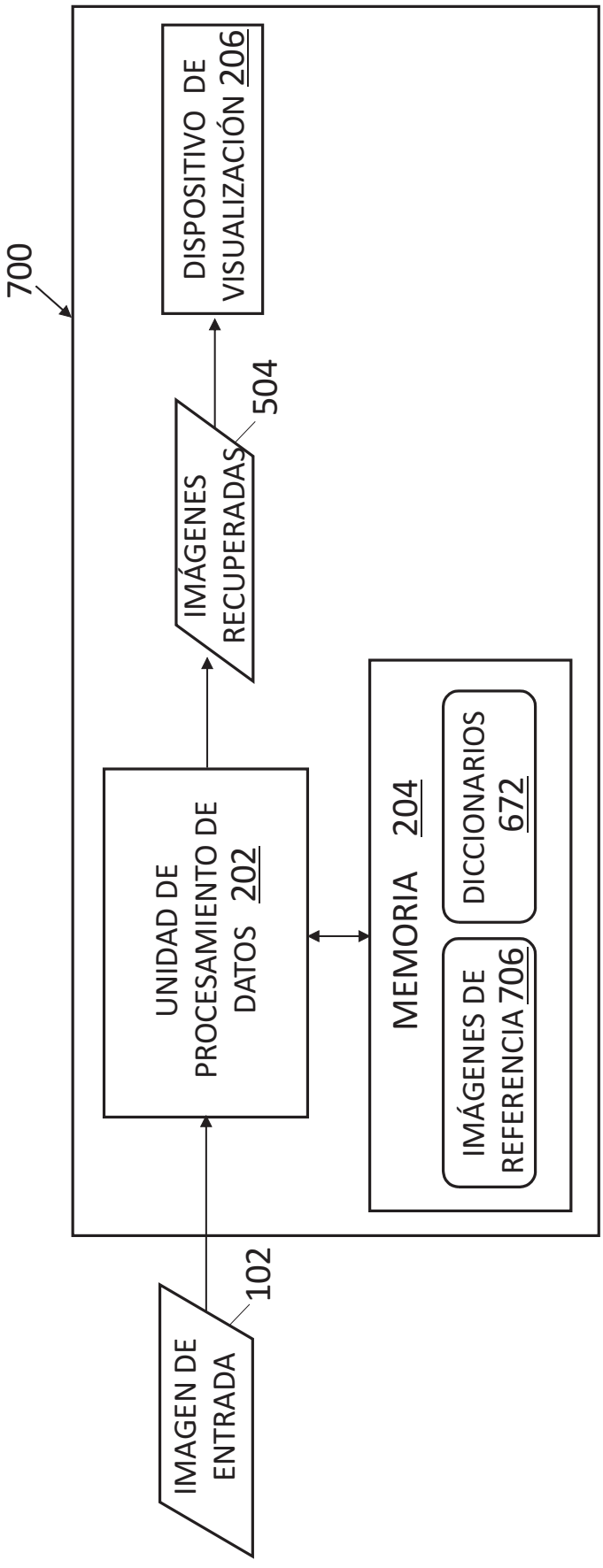
700
DISPOSITIVO DE VISUALIZACIÓN 206
IMÁGENES RECUPERADAS
504
UNIDAD DE PROCESAMIENTO DE DATOS 202
MEMORIA 204
DICCIONARIOS 672
IMÁGENES DE REFERENCIA 706
IMAGEN DE ENTRADA
102

FIG. 7

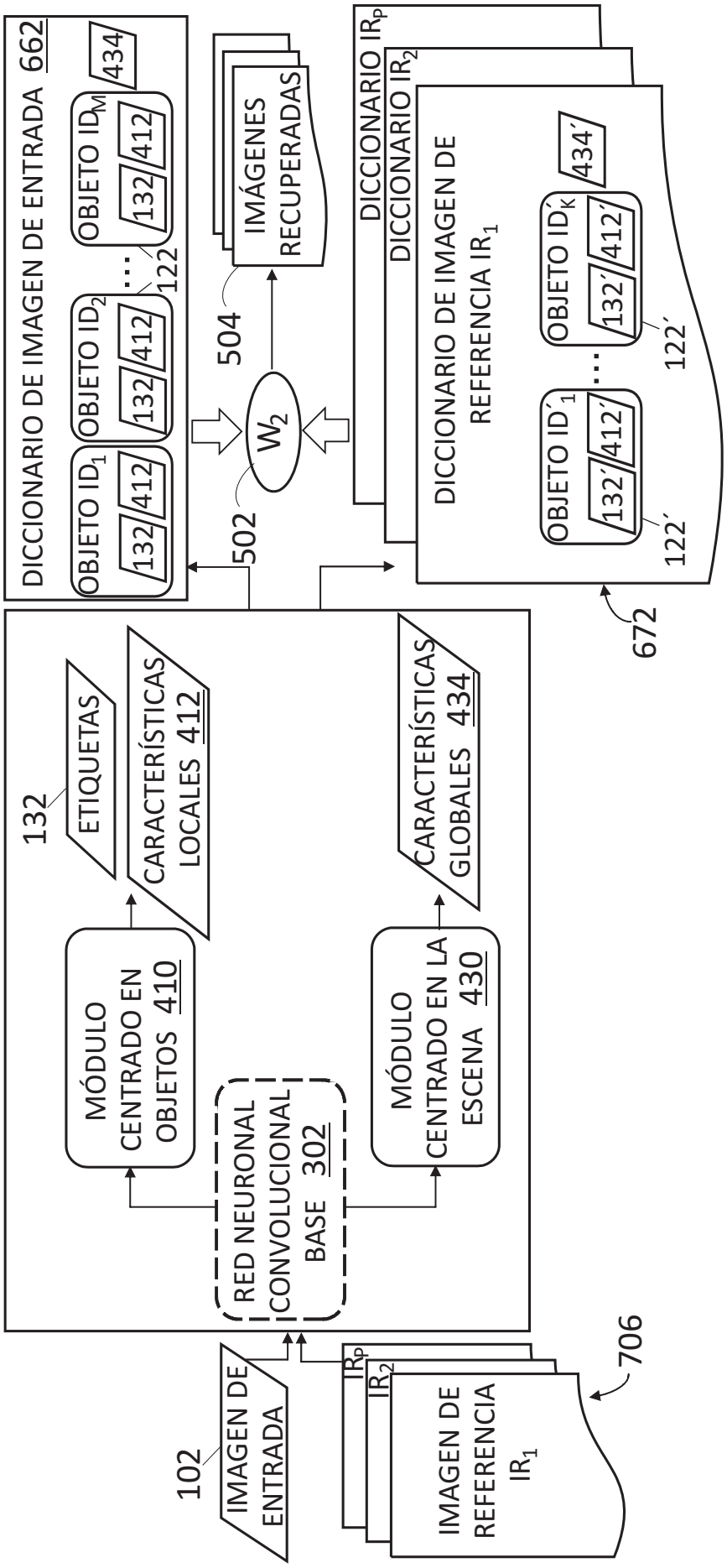
102
IMAGEN DE ENTRADA
706
IMAGEN DE REFERENCIA $IR_1$
$IR_2$
$IR_P$
672
132
ETIQUETAS
CARACTERÍSTICAS LOCALES 412
CARACTERÍSTICAS GLOBALES 434
MÓDULO CENTRADO EN OBJETOS 410
MÓDULO CENTRADO EN LA ESCENA 430
RED NEURONAL CONVOLUCIONAL BASE 302
DICCIONARIO DE IMAGEN DE ENTRADA 662
OBJETO $ID_1$
132/412
OBJETO $ID_2$
132/412
OBJETO $ID_M$
132/412
434
122
502
$W_2$
504
IMÁGENES RECUPERADAS
DICCIONARIO $IR_P$
DICCIONARIO $IR_2$
DICCIONARIO DE IMAGEN DE REFERENCIA $IR_1$
OBJETO $ID'_1$
132'/412'
OBJETO $ID'_K$
132'/412'
434'
122'
122'

FIG. 8

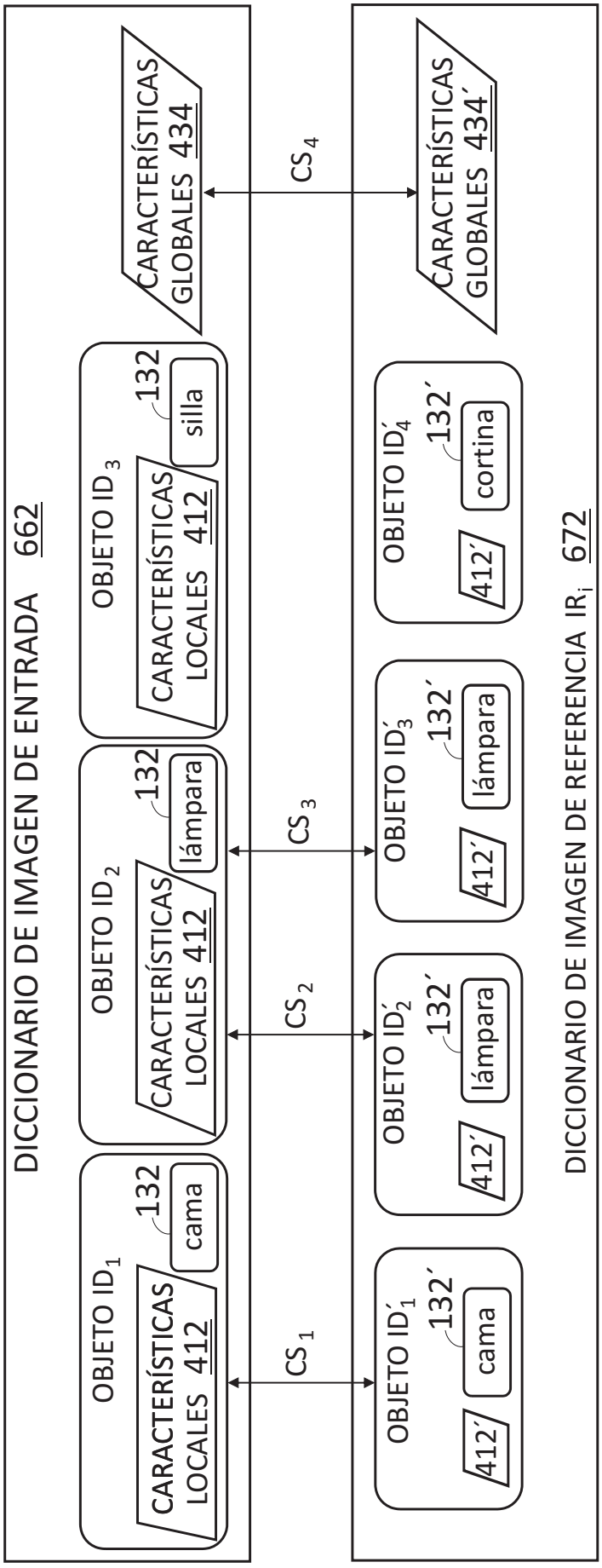
DICCIONARIO DE IMAGEN DE ENTRADA 662
OBJETO ID 1
CARACTERÍSTICAS LOCALES 412
132
cama
OBJETO ID 2
CARACTERÍSTICAS LOCALES 412
132
lámpara
OBJETO ID 3
CARACTERÍSTICAS LOCALES 412
132
silla
CARACTERÍSTICAS GLOBALES 434
CS 1
CS 2
CS 3
CS 4
DICCIONARIO DE IMAGEN DE REFERENCIA IR i 672
OBJETO ID´1
132´
412´
cama
OBJETO ID´2
132´
412´
lámpara
OBJETO ID´3
132´
412´
lámpara
OBJETO ID´4
132´
412´
cortina
CARACTERÍSTICAS GLOBALES 434´

FIG. 9

OFICINA ESPAÑOLA
DE PATENTES Y MARCAS

ESPAÑA

(21) N.º solicitud: 202231063

(22) Fecha de presentación de la solicitud: 13.12.2022

(32) Fecha de prioridad:

# INFORME SOBRE EL ESTADO DE LA TECNICA

(51) Int. CI.: ***G06V20/50*** (2022.01)
*G06F18/24* (2023.01)

DOCUMENTOS RELEVANTES

| Categoría | (56) Documentos citados | Reivindicaciones afectadas |
|---|---|---|
| X | CN 113239891 A (UNIV SHANGHAI MARITIME) 10/08/2021, resumen WPI, resumen EPODOC; párrafos 9, 38, 39, 53, 54; reivindicaciones; figuras 1, 2. | 1-27 |
| A | CN 115294441 A (UNIV NANJING POSTS & TELECOMMUNICATIONS) 04/11/2022, Todo el documento. | 1-27 |
| A | CN 115205660 A (SHENZHEN SENSETIME TECH CO LTD) 18/10/2022, Todo el documento. | 1-27 |
| A | US 2020202128 A1 (LIU, QINGFENG et al.) 25/06/2020, Todo el documento. | 1-27 |

Categoría de los documentos citados
X: de particular relevancia
Y: de particular relevancia combinado con otro/s de la misma categoría
A: refleja el estado de la técnica
O: referido a divulgación no escrita
P: publicado entre la fecha de prioridad y la de presentación de la solicitud
E: documento anterior, pero publicado después de la fecha de presentación de la solicitud

**El presente informe ha sido realizado**

☒ para todas las reivindicaciones ☐ para las reivindicaciones nº:

| Fecha de realización del informe | Examinador | Página |
|---|---|---|
| 26.10.2023 | M. T. Ibáñez Blanco | 1/2 |

Documentación mínima buscada (sistema de clasificación seguido de los símbolos de clasificación)

G06V, G06F

Bases de datos electrónicas consultadas durante la búsqueda (nombre de la base de datos y, si es posible, términos de búsqueda utilizados)

INVENES, EPODOC