(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 557 462**

(51) Int. CI.:

G06K 9/46 (2006.01)
G06K 9/42 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **21.06.2011 E 11727166 (8)**

(97) Fecha y número de publicación de la concesión europea: **07.10.2015 EP 2585979**

(54) Título: **Procedimiento y sistema para la identificación rápida y robusta de productos específicos en imágenes**

(30) Prioridad:

25.06.2010 ES 201030985

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
26.01.2016

(73) Titular/es:

TELEFÓNICA, S.A. (100.0%)
Gran Vía, 28
28013 Madrid, ES

(72) Inventor/es:

ADAMEK, TOMASZ y
RODRÍGUEZ BENITO, JAVIER

(74) Agente/Representante:

CARPINTERO LÓPEZ, Mario

DESCRIPCIÓN

Procedimiento y sistema para la identificación rápida y robusta de productos específicos en imágenes

Antecedentes de la invención

Campo técnico

La presente invención se refiere al campo de la recuperación de información multimedia basada en contenido [LSDJ06] y de la visión artificial. Más específicamente, la invención contribuye a las áreas de la recuperación de información multimedia basada en contenido relativas al problema de realizar búsquedas en grandes colecciones de imágenes basándose en su contenido, y también al área del reconocimiento de objetos que, en la visión artificial, es la tarea de encontrar un objeto dado en una imagen o una secuencia de vídeo.

Descripción de la técnica relacionada

La identificación de un objeto particular (idéntico) en una colección de imágenes está ahora alcanzando una cierta madurez [SZ03]. El problema sigue siendo un reto porque la apariencia visual de los objetos puede ser diferente debido a cambios en el punto de vista, condiciones de iluminación, o debido a un ocultamiento parcial, sin embargo ya existen soluciones con un rendimiento relativamente bueno con pequeñas colecciones. En la actualidad, las mayores dificultades que todavía existen parecen ser la correspondencia parcial, que permitan el reconocimiento de pequeños objetos “enterrados” dentro de fondos complejos, y la posibilidad de ampliar a escala los sistemas, necesaria para afrontar colecciones realmente grandes.

A continuación se comentarán recientes avances relevantes en el campo del rendimiento del reconocimiento, específicamente en el contexto de una rápida identificación de múltiples objetos pequeños en escenas complejas, basándose en una gran colección de imágenes de referencia de alta calidad.

A finales de la década de los noventa, David Lowe fue pionero en un nuevo enfoque del reconocimiento de objetos al proponer la transformada de características invariante a escala (conocida de manera generalizada como SIFT) [LOW99] (patente de los Estados Unidos 6711293). La idea básica que subyace al enfoque de Lowe es bastante simple. Objetos de la escena se caracterizan mediante descriptores locales que representan la apariencia de estos objetos en algunos puntos de interés (parches de imagen destacados). Los puntos de interés se extraen de una manera invariante respecto a la escala y la rotación de los objetos presentes en la escena. La figura 1 muestra ejemplos de puntos clave de interés SIFT [LOW99, LOW04] detectados en dos fotografías de la misma escena tomadas desde puntos de vista significativamente diferentes. Los puntos de interés están representados mediante círculos. Los centros de los círculos representan ubicaciones de puntos clave y sus radios representan sus escalas. Una interpretación intuitiva de los puntos de interés SIFT es que corresponden a estructuras a modo de gota o a modo de esquina y sus escalas corresponden estrechamente al tamaño de estas estructuras. Debe observarse que, independientemente de los ángulos de visión, la mayor parte de los puntos clave se detectan en la misma posición en la escena. Las imágenes originales pertenecen al conjunto de datos creado por Mikolajczyk et al., [MS04].

Los descriptores extraídos de una única imagen de entrenamiento de un objeto de referencia pueden usarse después para identificar instancias del objeto en nuevas imágenes (consultas). Los sistemas que se basan en los puntos SIFT pueden identificar de manera robusta objetos en escenas agrupadas, independientemente de su escala, orientación, ruido y también, hasta cierto punto, de cambios del punto de vista y la iluminación. El procedimiento de Lowe ha encontrado numerosas aplicaciones, entre las que se incluyen recuperación y clasificación de imágenes, reconocimiento de objetos, localización robotizada, formación de imágenes panorámicas (*stitching*) y otras muchas.

Animados por el rendimiento del procedimiento SIFT, muchos investigadores centraron su trabajo en ampliar adicionalmente las capacidades del enfoque. Por ejemplo, Mikolajczyk y Smith [MS04] propusieron detectores covariantes afines que permitían una robustez sin precedentes para cambios en los ángulos de visión. Matas et al. [MCUP02] propusieron un procedimiento alternativo para extraer puntos característicos denominados zonas extremas de máxima estabilidad que extrae puntos de interés diferentes a los seleccionados por el detector SIFT. Muy recientemente, Bay et al. [BTG06] propusieron una versión eficaz desde el punto de vista computacional del procedimiento SIFT denominada características robustas aceleradas (SURF). De manera sorprendente, el detector SURF no solo es tres veces más rápido que el detector SIFT, sino que también, en algunas aplicaciones, puede proporcionar un mayor rendimiento de reconocimiento. Uno de los ejemplos más interesantes de la aplicación de SURF es el reconocimiento de objetos de arte en un museo interior que contiene más de 200 artefactos, dando una tasa de reconocimiento del 85,7 %.

En muchas áreas de aplicación, el éxito de los enfoques de puntos característicos ha sido realmente espectacular. Sin embargo, hasta hace poco, seguía siendo imposible construir sistemas que pudieran reconocer de manera eficaz objetos en grandes colecciones de imágenes. Esta situación mejoró cuando Sivic y Zisserman propusieron usar puntos característicos imitando a los sistemas de recuperación textual [SZ03, SIV06]. En su enfoque, que denominaron “Video Google”, se cuantifican los puntos característicos de [MS04] y [MCUP02] mediante agrupamiento por k-medias en un vocabulario de lo que se denomina palabras visuales. Como resultado, cada zona

destacada puede correlacionarse fácilmente con la palabra visual más próxima, es decir, los puntos clave se representan mediante palabras visuales. Una imagen se representa entonces como una “bolsa de palabras visuales” (BoW), y éstas se introducen en un índice para su posterior consulta y recuperación. El enfoque permite un reconocimiento eficaz en colecciones de imágenes muy grandes. Por ejemplo, la identificación de una pequeña zona seleccionada por el usuario en una colección de cuatro mil imágenes tarda 0,1 segundos.

Aunque los resultados de “Video Google” eran muy admirables, especialmente en comparación con otros procedimientos disponibles en aquel momento, la búsqueda en escenas completas o incluso en zonas grandes seguía siendo prohibitivamente lenta. Por ejemplo, correlacionar escenas representadas usando imágenes con un tamaño de 720x576 píxeles en la colección de cuatro mil imágenes tardaba aproximadamente 20 segundos [SIV06]. Esta limitación la paliaron hasta cierto punto Nister y Stewenius [NS06] quienes propusieron un motor de búsqueda basado en imágenes muy optimizado que podía realizar reconocimiento de imágenes casi en tiempo real en colecciones más grandes. En particular, su sistema podía proporcionar buenos resultados de reconocimiento de 40.000 carátulas de CD en tiempo real.

Finalmente, hace muy poco, Philbin et al. [PCI+07, PCI+08] propusieron una variante mejorada del enfoque de “Video Google” y demostraron que podía recuperar rápidamente las imágenes de 11 “monumentos” de Oxford diferentes de una colección de cinco mil imágenes de alta resolución (1024 x 768) recopiladas de Flickr [FLI].

Los recientes avances espectaculares en el área del reconocimiento visual de objetos están empezando a atraer mucho el interés de la industria. En la actualidad, varias empresas ofrecen tecnologías y servicios basados, al menos en parte, en los avances anteriormente mencionados.

Kooaba [KOO], una empresa escindida de ETH Zurich fundada a finales de 2006 por los inventores del enfoque SURF [BTG06], usa la tecnología de reconocimiento de objetos para proporcionar acceso y realizar búsquedas de contenido digital de teléfonos móviles. Se accede a los resultados de búsqueda de Kooaba enviando una imagen como consulta. Defienden su tecnología diciendo que permite literalmente “hacer clic” en objetos del mundo real tales como pósters de películas, artículos enlazados en periódicos o revistas y, en el futuro, incluso en lugares de interés turístico.

Evolution Robotics en Pasadena, Calif, [EVO] ha desarrollado un motor de búsqueda visual que puede reconocer de qué tomó una foto el usuario, y entonces los publicistas pueden usar eso para enviar contenido relevante al teléfono móvil del usuario. Predicen que en los próximos 10 años se podrá levantar el teléfono móvil y se etiquetará visualmente todo lo que tenga delante. Uno de los asesores de Evolution Robotics es el Dr. David Lowe, el inventor del enfoque SIFT [LOW99].

SuperWise Technologies AG [SUP], empresa que desarrolló el sistema de reconocimiento de imágenes Apollo, ha desarrollado un novedoso programa para teléfonos móviles denominado eye-Phone, que puede proporcionar al usuario información turística esté donde esté. En otras palabras, eye-Phone puede proporcionar información sobre lo que el usuario ve cuando lo ve. El programa combina tres modernas tecnologías actuales: servicios de localización mediante navegación por satélite, reconocimiento avanzado de objetos e información relevante recuperada de Internet. Con el eye-Phone en su teléfono, por ejemplo mientras da un paseo, el usuario puede hacer fotografías con su teléfono móvil y seleccionar el elemento de interés con el cursor. La zona seleccionada se transmite entonces con datos de localización mediante navegación por satélite a un sistema central que realiza el reconocimiento de objetos y que está interconectado a bases de datos en Internet para obtener información sobre el objeto. La información encontrada se devuelve al teléfono y se muestra al usuario.

Los enfoques existentes tienen limitaciones importantes. En la actualidad, solo los procedimientos que se basan en características de imagen locales parecen estar cerca de cumplir con la mayor parte de los requisitos necesarios para que un motor de búsqueda entregue resultados en respuesta a fotografías.

Uno de los primeros sistemas que pertenecen a esta categoría de procedimientos y que realiza reconocimiento de objetos en tiempo real con una colección de decenas de imágenes fue propuesto por David Lowe, el inventor de SIFT [LOW99, LOW04]. En la primera etapa de este enfoque se hacían corresponder puntos clave de manera independiente con la base de datos de puntos clave extraídos de imágenes de referencia usando un procedimiento de aproximación para encontrar los vecinos más cercanos denominado Best-Bin-First (BBF) [BL97]. Estas correspondencias iniciales se validaban más tarde en la segunda fase mediante agrupamiento en el espacio de posición usando la transformada de Hough [HOU62]. Este sistema parece ser muy adecuado para el reconocimiento de objetos cuando estos están abarrotados y se tapan entre sí, pero no hay pruebas en la literatura de que pueda aumentarse a escala para colecciones de más de decenas de imágenes.

Para mejorar la capacidad del aumento a escala, otros investigadores han propuesto usar puntos característicos imitando a los sistemas de recuperación textual [SZ03, SIV06]. Por ejemplo, Sivic y Zisserman [SZ03, SIV06, PCI+07, PCI+08] propusieron cuantificar descriptores de puntos clave mediante agrupamiento por k-medias creando lo que se denomina “vocabulario de palabras visuales”. El reconocimiento se realiza en dos fases. La primera fase se basa en el modelo de espacio vectorial de recuperación de información [BYRN99], en el que la colección de palabras visuales se usa con la puntuación estándar de frecuencia de término - frecuencia inversa de documento

(TF-IDF) de la relevancia de una imagen para la consulta. Esto da como resultado una lista inicial de los primeros n candidatos potencialmente relevantes para la consulta. Debe observarse que, normalmente, no se usa información espacial acerca de la ubicación en la imagen de las palabras visuales en la primera fase. La segunda etapa normalmente implica algún tipo de comprobación de consistencia espacial en la que se usa información espacial de los puntos clave para filtrar la lista inicial de candidatos. La mayor limitación de los enfoques de esta categoría procede del hecho de que se basan en puntuación TF-IDF, que particularmente no es muy adecuada para identificar pequeños objetos “enterrados” en escenas abarrotadas. La identificación de múltiples objetos pequeños requiere aceptar listas mucho más largas de candidatos de correspondencia iniciales. Esto da como resultado un aumento del coste global de la correspondencia ya que la posterior validación de la consistencia espacial es costosa desde un punto de vista computacional en comparación con lo que cuesta la fase inicial. Además, nuestros experimentos indican que estos tipos de procedimientos no son adecuados para la identificación de muchos tipos de productos reales, tales como, por ejemplo, latas de refrescos o cajas de DVD, ya que la puntuación TF-IDF a menudo está sesgada por puntos clave de los bordes de los objetos que a menudo se asignan a palabras visuales habituales en escenas que contienen otros objetos manufacturados.

Debido al coste computacional de la etapa de validación de la consistencia espacial, Nister y Stewenius [NS06] se concentraron en mejorar la calidad de la fase de recuperación previa a la geométrica, que sugieren que es crucial para aumentar a escala a grandes bases de datos. Como solución, propusieron palabras visuales definidas de manera jerárquica que forman un árbol de vocabulario que permite consultar de manera más eficaz las palabras visuales. Eso permite usar vocabularios mucho más grandes que han demostrado que dan como resultado una mejora de la calidad de los resultados, sin implicar tener que considerar la disposición geométrica de las palabras visuales. Aunque este enfoque se ajusta a escala muy bien a grandes colecciones, hasta ahora se ha mostrado que solo tiene un buen rendimiento cuando los objetos que han de correlacionarse cubren la mayor parte de las imágenes. Parece que esta limitación está causada por el hecho de que se basa en una variante de la puntuación TF-IDF y por la ausencia de una validación de la consistencia espacial.

Sumario de la invención

Un objetivo de la presente invención es desarrollar un motor de búsqueda que entrega resultados en respuesta a fotografías en lugar de palabras de texto. Se plantea un escenario en el que el usuario proporciona una imagen de consulta que contiene los objetos que han de reconocerse, y el sistema devuelve una lista clasificada de imágenes de referencia que contienen los mismos objetos, recuperada de un gran corpus. En particular, un objetivo es desarrollar un procedimiento particularmente adecuado para el reconocimiento de una amplia gama de productos en 3D potencialmente relevantes para muchos escenarios de casos de uso atractivos, tales como por ejemplo libros, CD/DVD, productos envasados en tiendas de alimentación, carteles urbanos, fotografías en periódicos y revistas y cualquier objeto con marcas comerciales distintivas, etc.

Una imagen de consulta típica se espera que contenga múltiples objetos para reconocer situados en una escena compleja. Además, es habitual que una imagen de consulta tenga mala calidad (por ejemplo tomada por la cámara de un teléfono móvil). Por otro lado, se supone que cada imagen de referencia contendrá solo un objeto de referencia bien posicionado y un fondo relativamente simple. Es deseable que el sistema permita la indexación de un gran número de imágenes de referencia (>1000), y que pueda realizar una rápida (< 5 segundos) identificación de objetos presentes en una imagen de consulta comparándola con las imágenes indexadas. El motor de búsqueda debería proporcionar resultados significativos independientemente de la ubicación, la escala y la orientación de los objetos en la imagen de consulta y debería ser robusto frente a ruido y, hasta cierto punto, frente a cambios en el punto de vista y la iluminación. Finalmente, el motor de búsqueda debería permitir una rápida inserción (sobre la marcha) de nuevos objetos en la base de datos.

Con el fin de alcanzar al menos parte de estos objetivos, según la invención se proporcionan un procedimiento y un sistema según las reivindicaciones independientes. Las realizaciones favorables se definen en las reivindicaciones dependientes.

La idea básica es identificar objetos a partir de la imagen de consulta en una única etapa, realizando una validación parcial de la consistencia espacial entre palabras visuales que se han hecho corresponder mediante un uso directo del vocabulario de palabras visuales y nuestra extensión de la estructura de archivo invertido.

En otras palabras, el procedimiento y sistema propuesto combina la excepcional capacidad de ajuste a escala de los procedimientos que se basan en agrupamiento de descriptores en vocabularios de palabras visuales [SZ03, SIV06, NS06, PCI+07, PCI+08], con la robustez frente al abarrotamiento y ocultamiento parcial de los procedimientos que se basan en la validación de la consistencia espacial usando la transformada de Hough [HOU62, LOW99, LOW04]. Según un punto de vista, la invención puede verse como un intento de eliminar la fase de reconocimiento inicial basada en el modelo de espacio vectorial (puntuación TF-IDF) de los enfoques que se basan en vocabularios de palabras visuales y, en lugar de ello, realizar el reconocimiento en una única etapa que implica la validación de la consistencia espacial entre palabras visuales que se han hecho corresponder. Por otro lado, la invención también puede verse como un intento de sustituir la búsqueda aproximada de los vecinos más cercanos del procedimiento propuesto en [LOW99, LOW04] por la correspondencia usando vocabularios de palabras visuales.

La presente invención pretende aprovechar el hecho de que, en muchos escenarios de aplicación, es aceptable suponer que cada imagen de referencia solo contiene un objeto de referencia bien posicionado (es decir modelo) y un fondo relativamente simple. Debe observarse que no se adopta ninguna suposición con respecto al número de objetos y la complejidad del fondo en la imagen de consulta. Esto contrasta con los procedimientos existentes, en los que normalmente tanto la imagen de consulta como la de referencia se procesan efectivamente del mismo modo. Además, la intención era desarrollar un procedimiento muy adecuado para el reconocimiento de una amplia gama de productos en 3D potencialmente relevantes para muchos escenarios de casos de uso atractivos, como por ejemplo libros, CD/DVD, productos envasados en tiendas de alimentación, carteles urbanos, fotografías en periódicos y revistas, y cualquier objeto con marcas comerciales, etc. En casos en los que una imagen de consulta contiene un objeto para reconocer que pertenece a una familia de productos con un subconjunto común de marcas comerciales, por ejemplo muchos productos de Coca-Cola contienen el logotipo de Coca-Cola, el sistema debería devolver una lista clasificada de todos los productos relevantes que tienen marcas comerciales similares.

Los experimentos indican que la invención da como resultado un importante avance en cuanto a rendimiento de reconocimiento, específicamente en el contexto de una rápida identificación de múltiples objetos pequeños en escenas complejas basándose en una gran colección de imágenes de referencia de alta calidad.

El presente enfoque se basa en características de imagen locales. Se escanean todas las imágenes para determinar zonas “destacadas” (puntos clave) y se calcula un descriptor de muchas dimensiones para cada zona. Los puntos clave detectados a escalas muy bajas y muy altas se eliminan y, en el caso de imágenes de referencia, se normalizan las escalas de puntos clave con respecto a un tamaño estimado del objeto de referencia representado. En un proceso fuera de línea, se agrupa un gran número de ejemplos de descriptores en el vocabulario de palabras visuales, que define una cuantificación del espacio de descriptor. Desde este momento, cada punto clave puede correlacionarse con la palabra visual más próxima.

Las imágenes no se representan como bolsas de palabras visuales. En lugar de ello, proponemos extender la estructura de archivo invertido propuesta en [SZ03] para soportar agrupamiento de correspondencias en el espacio de posición, de una manera que se asemeja a la transformada de Hough ampliamente conocida. Con el fin de mantener el coste computacional bajo, se propone limitar el espacio de posición únicamente a orientación y escala. La estructura de archivo invertido tiene una lista de coincidencias para cada palabra visual, que almacena todas las apariciones de la palabra en todas las imágenes de referencia. A diferencia de otros enfoques, cada coincidencia almacena no solo un identificador de la imagen de referencia en la que se detectó originalmente el punto clave, sino también información acerca de su escala y orientación. Además, cada coincidencia tiene una fuerza asociada de la evidencia con la que puede confirmar la existencia del objeto correspondiente. La fuerza de la coincidencia se calcula basándose en su escala (los puntos clave detectados a mayor escala son más distintivos), y el número de coincidencias asignadas a la misma palabra visual y que tienen una orientación y escala similares. De forma similar, cada punto clave a partir de la imagen de consulta también tiene una fuerza asociada de la certeza que puede proporcionar. En este caso, la fuerza depende solo del número de puntos clave a partir de la consulta asignados a la misma palabra visual y que tienen orientación y escala similares. El reconocimiento empieza asignando puntos clave a partir de la imagen de consulta a las palabras visuales más próximas. En realidad, esta etapa es equivalente a asignar cada punto clave de consulta a una lista completa de coincidencias correspondientes a la misma palabra visual. Después, cada emparejamiento del punto clave y una de sus coincidencias de la lista da un voto a un acumulador de posición correspondiente a la imagen de referencia en la que se encontró la coincidencia. Cada par punto clave/coincidencia predice una orientación y escala específicas del modelo representado por la imagen de referencia. La fuerza de cada voto se calcula como un producto escalar de las fuerzas del punto clave y la coincidencia. Una vez dados todos los votos, todos los grupos de acumuladores que han recibido al menos un voto se escanean para identificar los grupos con el máximo número de votos. Los valores acumulados en estos grupos se toman como las puntuaciones de relevancia finales para las correspondientes imágenes de referencia. Finalmente, las imágenes de referencia se ordenan según las puntuaciones de relevancia y se seleccionan los objetos más relevantes basándose en una extensión del procedimiento de aplicación de umbral dinámico de [ROS01].

Estos y otros aspectos de la invención resultarán evidentes a partir de y se dilucidarán con referencia a las realizaciones descritas a continuación.

Breve descripción de los dibujos

La invención se entenderá mejor y sus numerosos objetivos y ventajas serán más evidentes para los expertos en la técnica mediante referencia a los siguientes dibujos, junto con la memoria descriptiva que los acompaña, en los que:

La figura 1 muestra la detección de puntos clave en una imagen según la técnica anterior.

La figura 2 muestra una perspectiva general del procedimiento según una realización de la presente invención que muestra la relación entre sus principales componentes.

La figura 3 muestra una perspectiva general del proceso de reconocimiento de objetos del procedimiento representado en la figura 2.

La figura 4 muestra una perspectiva general del proceso de indexación del procedimiento representado en la

figura 2.

La figura 5 muestra un ejemplo de la estructura de archivo invertido usada en el procedimiento según la presente invención.

Figura 6 muestra un ejemplo de identificación de un pequeño objeto con el procedimiento según la presente invención.

Figura 7 muestra un ejemplo de identificación de un objeto con una posición difícil con el procedimiento según la presente invención.

Figura 8 muestra un ejemplo de identificación de un objeto oculto con el procedimiento según la presente invención.

Figura 9 muestra un ejemplo de identificación de un pequeño objeto en una escena abarrotada con el procedimiento según la presente invención.

Figura 10 muestra un ejemplo de identificación de múltiples objetos pequeños con el procedimiento según la presente invención.

Figura 11 muestra un ejemplo de una aplicación industrial del procedimiento según la presente invención.

Descripción detallada de la presente invención

Se describirá una realización ejemplar del procedimiento de identificación de objetos específicos en imágenes según la invención.

El enfoque propuesto consiste en cuatro componentes (fases) principales:

1. *Extracción de características* implica la identificación de zonas de imagen “destacadas” (puntos clave) y el cálculo de sus representaciones (descriptores) – véase el ejemplo en la figura 1. Esta fase incluye también un postprocesamiento de puntos clave en el que los puntos clave que no son útiles para el proceso de reconocimiento se eliminan. Debe observarse que la extracción de características se realiza para: tanto imágenes que representan objetos de referencia (imágenes de referencia) como imágenes que representan objetos desconocidos que han de identificarse (imágenes de consulta).

2. *Construcción de un vocabulario de palabras visuales* es un proceso fuera de línea, en el que se agrupa un gran número de ejemplos de descriptores en vocabularios de palabras visuales. El papel de estos vocabularios es cuantificar el espacio de los descriptores. Una vez creado el vocabulario, los puntos clave de las imágenes de referencia y de consulta pueden correlacionarse para dar las palabras visuales más próximas. En otras palabras, los puntos clave pueden representarse mediante identificadores de palabras visuales, en lugar de descriptores multidimensionales.

3. *Indexación de imágenes de referencia* implica la extracción de características locales para imágenes de referencia y su organización en una estructura que permite su rápida correspondencia con características extraídas de imágenes de consulta. Este proceso consiste en (i) extracción de puntos clave y (ii) postprocesamiento, (iii) asignación de puntos clave a palabras visuales, (iv) estimación de pesos de votación y (v) adición de puntos clave a una estructura de archivo invertido como las denominadas coincidencias – véase una perspectiva general del proceso de indexación en la figura 4. La adición de un nuevo objeto de referencia a la base de datos implica añadir coincidencias que representan puntos clave a la estructura de archivo invertido. En la estructura de archivo invertido hay una lista (lista de coincidencias) para cada palabra visual que almacena todas las apariciones (coincidencias) de la palabra en imágenes de referencia – véase la figura 5. Cada coincidencia corresponde a un punto clave de una imagen de referencia y almacena el identificador de la imagen de referencia en la que se detectó el punto clave e información acerca de su escala y orientación. Además, cada coincidencia tiene un peso asociado (fuerza) con el que puede confirmar la existencia del correspondiente objeto de referencia en respuesta a una aparición de la palabra visual en una imagen de entrada.

4. *Reconocimiento de objetos* presentes en la imagen de consulta consiste en las siguientes etapas: (i) extracción de puntos clave y (ii) postprocesamiento, (iii) asignación de puntos clave a palabras visuales, (iv) cálculo de pesos de votación (fuerzas) correspondientes a cada punto clave, (v) agregación de evidencias proporcionadas por pares (punto clave de consulta, coincidencia) en acumuladores de votos, (vi) identificación de las puntuaciones de correspondencia correspondientes a cada imagen de referencia, y finalmente (vii) ordenación y selección de los resultados más relevantes basándose en una extensión del procedimiento de aplicación de umbral dinámico de [ROS01]. Una perspectiva general del proceso de reconocimiento puede verse en la figura 3.

La relación entre los componentes o “fases” principales del enfoque se exhibe en la figura 2. Debe observarse que la creación de vocabularios, la indexación y el reconocimiento requieren la etapa de extracción de características.

Asimismo, la indexación y el reconocimiento requieren usar un vocabulario de palabras visuales creado a partir de la gran colección de imágenes de entrenamiento. Todas las fases anteriores se comentan con más detalle a continuación en el presente documento.

Extracción de características y postprocesamiento

- **Características locales**

En el enfoque propuesto las imágenes se representan mediante un conjunto de características locales muy distintivas (puntos clave). Estas características locales pueden verse como parches de imagen destacados que tienen características específicas e invariantes que pueden almacenarse en la base de datos y compararse. En otras palabras, el motor de búsqueda propuesto requiere que cada imagen se represente como un conjunto de puntos clave, cada uno con una ubicación, escala, orientación y descriptor específicos.

Para que sean útiles para el reconocimiento de objetos, los puntos clave tienen que poder detectarse de manera consistente independientemente de la ubicación, el tamaño, la orientación, el ruido, el abarrotamiento y cambios en la iluminación de los objetos y el punto de vista de la cámara. El número de puntos detectados en cada imagen tiene que ser suficiente para representar todos los elementos potencialmente interesantes de la escena. Además, los descriptores de puntos clave tienen que ser razonablemente distintivos para facilitar la identificación de puntos clave correspondientes a partir de diferentes imágenes. Finalmente, la extracción de características tiene que ser eficaz desde un punto de vista computacional porque el reconocimiento de objetos implica detección de puntos clave en línea en imágenes de consulta. Un ejemplo de puntos clave útiles se muestra en la figura 1.

En el prototipo desarrollado, las características locales se extraen usando la transformada de características invariantes a escala (SIFT) [LOW99, LOW04] (patente de los Estados Unidos 6711293). Sin embargo, el motor de búsqueda propuesto debería proporcionar un rendimiento similar o mejor cuando se usa con otras representaciones alternativas, tales como por ejemplo características robustas aceleradas (SURF) [BTG06] (patente europea EP1850270), zonas extremas con máxima estabilidad [MCUP02] o detectores covariantes afines [MS04].

- **Postprocesamiento de puntos clave**

Los experimentos realizados indican que no todos los puntos clave son igual de útiles para la identificación de objetos. Por ejemplo, en casos de imágenes de alta resolución muchos de los puntos clave detectados a las escalas más bajas no representan ningún patrón discriminatorio, sino que corresponden simplemente a diferentes tipos de ruido o artefactos.

Los detectores usados más habitualmente, tales como por ejemplo SIFT, permiten controlar el número de puntos clave y el rango de escalas analizadas principalmente ajustando la resolución de las imágenes de entrada. Este mecanismo no permite relacionar el rango de escalas que está usándose con el tamaño de los objetos que están representándose. Esto significa que todas las imágenes de referencia deben tener aproximadamente la misma resolución para garantizar comparaciones significativas.

Para paliar este problema, se propone realizar una etapa de postprocesamiento adicional que: (i) normaliza las escalas de los puntos clave según el tamaño de los objetos de referencia y (ii) elimina los puntos clave que no pueden contribuir de manera eficaz al proceso de reconocimiento basándose en sus escalas normalizadas. Se supone que cada imagen de referencia debería contener solo un ejemplo de un objeto de referencia y un fondo relativamente simple y uniforme. La mayoría de los puntos clave deberían detectarse en áreas correspondientes al objeto de referencia, mientras que el fondo no debería generar un número significativo de puntos clave. En tales imágenes es posible detectar automáticamente lo que se denomina zona de interés (ROI) basándose en las ubicaciones de los puntos clave detectados. Para mayor simplicidad solo se consideran ROI rectangulares.

En el caso de imágenes de referencia, el centro de la ROI se estima como el centro de la masa del conjunto de todas las ubicaciones de puntos clave detectadas. Su anchura y altura iniciales se calculan de manera independiente en las direcciones horizontal y vertical como cuatro veces los valores de la desviación estándar de ubicaciones de puntos clave. Con el fin de minimizar la influencia de zonas con ruido, las ubicaciones de puntos clave se ponderan según escalas de puntos clave. Finalmente, los límites iniciales se ajustan (“reducen”) siempre que cubren áreas sin puntos clave.

La longitud de la diagonal de la ROI se usa para normalizar las escalas de todos los puntos clave. Debe observarse que puesto que las ROI dependen solo de los tamaños de los objetos representados, proporcionan referencias ideales para normalizar las escalas de los puntos clave de manera independiente a la resolución de la imagen.

Una vez identificada la ROI, los puntos clave ubicados fuera de la ROI se eliminan. Entonces, los puntos clave con escala normalizada inferior a un valor predefinido también se eliminan. Todos los demás puntos clave se clasifican según sus escalas normalizadas y solo se conserva un número predefinido de puntos con las escalas más grandes. En la mayoría de las aplicaciones, limitar el número de puntos clave en imágenes de referencia a 800 da buenos resultados.

Puesto que, en el caso de las imágenes de consulta, no puede esperarse un fondo simple, las ROI se ajustan para cubrir imágenes completas. El postprocesamiento siguiente de puntos clave sigue el mismo esquema que en el caso de las imágenes de referencia. Los experimentos realizados indican que limitar el número de puntos clave en las imágenes de consulta a 1200 es suficiente para garantizar el reconocimiento de pequeños objetos “enterrados en escenas abarrotadas”.

Debe enfatizarse que la etapa anterior de postprocesamiento y normalización de escala desempeña un papel importante en el proceso de correspondencia global y es crucial para garantizar un alto rendimiento de reconocimiento.

Construcción de vocabularios de palabras visuales

El reconocimiento de objetos requiere establecer correspondencias entre puntos clave a partir de la imagen de consulta y todas las imágenes de referencia. En casos de grandes colecciones de imágenes de referencia, una búsqueda exhaustiva de las correspondencias entre puntos clave no es viable desde el punto de vista del coste computacional. En la solución propuesta, la búsqueda exhaustiva entre todas las posibles correspondencias de puntos clave se evita cuantificando el espacio de descriptor en agrupamientos de manera similar a la comentada en [SZ03, SIV06]. En la literatura tales agrupamientos se denominan a menudo “palabras visuales” y las colecciones de todas las palabras visuales se denominan a menudo vocabularios. Los vocabularios permiten la asignación de puntos clave a palabras visuales con los descriptores más similares. Esta operación asigna de manera eficaz cada punto clave de la imagen de consulta a una lista completa de puntos clave de imágenes de referencia que corresponden a la misma palabra visual.

En el prototipo implementado, la cuantificación se lleva a cabo mediante el agrupamiento de *K-medias* ampliamente conocido. Sin embargo, también es posible incorporar otros procedimientos de agrupamiento, tales como k-medias jerárquico *de* [NS06] (patente de los Estados Unidos 20070214172).

El agrupamiento se realiza fuera de línea usando puntos clave de imágenes típicas de un escenario de aplicación dado. Usar colecciones más grandes de imágenes produce diccionarios más genéricos y lleva a un mejor rendimiento de reconocimiento. Sin embargo, puesto que el coste computacional de crear diccionarios visuales depende del número de puntos clave, a menudo es necesario seleccionar de manera aleatoria solo un subconjunto imágenes disponibles [SZ03].

El número de agrupamientos (es decir, el tamaño del diccionario) afecta al rendimiento de reconocimiento y a la velocidad del reconocimiento y la indexación. Diccionarios más grandes (células de cuantificación muy pequeñas) proporcionan una mejor distintividad pero, al mismo tiempo, pueden disminuir la repetibilidad en presencia de ruido. Además, diccionarios más grandes son costosos de crear desde el punto de vista computacional, y dan como resultado un reconocimiento mucho más lento. Siguiendo a [SZ03] se ha elegido usar diccionarios que contienen 10.000 palabras visuales que proporcionan un buen equilibrio entre distintividad, repetibilidad y rapidez de reconocimiento.

En principio, las adiciones de nuevas imágenes de referencia no requieren una actualización del diccionario visual. Por otro lado, volver a crear el diccionario después de cambios significativos en la colección de las imágenes de referencia puede mejorar el rendimiento de reconocimiento. Tal recreación del diccionario implica una nueva indexación de todas las imágenes de referencia. Tanto la actualización del diccionario como la nueva indexación pueden realizarse fuera de línea.

Siguiendo las sugerencias de [SZ03, SIV06, NS06], se ha incorporado un mecanismo que excluye del proceso de reconocimiento puntos clave asignados a palabras visuales muy comunes. En la literatura, estas palabras visuales muy comunes se denominan habitualmente “palabras vacías visuales”, debido a su cierta analogía con el problema en la recuperación textual en el que las palabras muy comunes, tales como ‘and’ o ‘the’ del inglés, no son discriminatorias. La frecuencia de las palabras visuales se calcula basándose en sus apariciones en toda la colección de imágenes de referencia. Las frecuencias pueden actualizarse siempre que haya cambios significativos en la colección de imágenes de referencia. Un porcentaje predefinido (normalmente el 1 %) de palabras visuales son vacías. En otras palabras, los puntos clave de las imágenes de consulta asignados a las palabras visuales más comunes (en el presente caso 100) no se tienen en cuenta en el proceso de reconocimiento. Debe observarse que el mecanismo usado para excluir las palabras vacías difiere ligeramente del propuesto en [SZ03, SIV06, NS06]. En el presente caso, las palabras vacías se incluyen para la indexación de imágenes de referencia. Las palabras vacías se tienen en cuenta solo en la fase de reconocimiento, cuando los puntos clave de la imagen de consulta asignados a palabras vacías se excluyen del proceso de correspondencia. Esta solución permite evitar una nueva indexación frecuente de toda la base de datos cuando cambian las palabras vacías debido a adiciones a la colección. Aunque los experimentos que se realizaron indican ciertas mejoras en el rendimiento de reconocimiento debido a la incorporación del mecanismo de palabras vacías, esta extensión no es crucial para el rendimiento del motor de reconocimiento propuesto.

Indexación de imágenes de referencia

En términos generales, la indexación de imágenes de referencia implica una extracción de características locales y

su organización en una estructura que permita su rápida correspondencia con características extraídas de imágenes de consulta.

Una visión general del proceso de indexación se muestra en la figura 4. La indexación de una nueva imagen de referencia empieza con (i) extracción de puntos clave y (ii) postprocesamiento descritos en la sección "Postprocesamiento de puntos clave". En la siguiente etapa, (iii) los puntos clave extraídos se asignan a las palabras visuales más próximas (es decir, las palabras que las representan mejor). Específicamente, cada punto clave se asigna a la palabra visual (agrupamiento) del vocabulario que tiene el descriptor más similar. Una vez representados todos los puntos clave con palabras visuales correspondientes, la etapa consecutiva (iv) estima su importancia individual (pesos) en el proceso de reconocimiento. Los pesos se estiman basándose en escalas de puntos clave y también en los números de puntos clave en la misma imagen pertenecientes a la misma palabra visual y que tienen orientación y escala similares. Finalmente, (v) todos los puntos clave y sus pesos se añaden a la estructura de archivo invertido como las denominadas coincidencias.

Puesto que las dos primeras etapas se han descrito en la sección "Extracción de características y postprocesamiento", el resto de esta sección describe en detalle solo las últimas tres etapas específicas del proceso de indexación.

- **Clasificación de puntos clave**

En esta etapa, cada punto clave de la imagen se asigna a una palabra visual con el descriptor más similar. Esto implica la comparación de descriptores de puntos clave con descriptores de palabras visuales. En la implementación actual, la asignación se lleva a cabo mediante una búsqueda exhaustiva en todo el vocabulario [SZ03, SIV06]. Debe observarse que actualmente ésta es la etapa más intensiva desde un punto de vista computacional del proceso de indexación y reconocimiento. Sin embargo, en el futuro debería ser posible incorporar los procedimientos más recientes para la clasificación rápida de puntos clave tal como el propuesto en [NS06].

- **Estimación de pesos de puntos clave**

En el enfoque propuesto, cada punto clave tiene un factor de ponderación (fuerza) asociado que refleja su importancia en el proceso de correspondencia. En la implementación actual los pesos están basados en dos factores principales: (i) la escala a la que se detectó el punto clave, y (ii) el número de puntos clave en la imagen asignados a la misma palabra visual que el punto clave considerado y que tienen orientación y escala similares.

La incorporación de escalas de puntos clave en los pesos viene motivada por el hecho de que los puntos clave detectados a escalas mayores son más discriminatorios que los puntos clave detectados a escalas muy bajas. De hecho, muchos puntos clave detectados a escalas muy pequeñas corresponden a elementos insignificantes de la escena. A menudo tales puntos clave son muy comunes en muchas imágenes de referencia diferentes y por tanto no son muy discriminatorios. Al mismo tiempo, los puntos clave detectados a escalas mayores corresponden normalmente a partes más grandes de la escena y son mucho más discriminatorios.

Basándose en la observación anterior, los pesos se eligieron para ser proporcionales a escalas a las que se detectaron los puntos clave. Específicamente, el factor de ponderación w_S^i correspondiente a la escala s_i a la que se detectó el punto clave i se calcula como:

$$w_S^i = \min(s_i, T_s),$$

donde T_s es un umbral elegido empíricamente que limita la influencia de puntos clave detectados a escalas muy grandes.

El segundo factor de ponderación w_M^i se introduce para limitar la influencia de grupos de puntos clave de la misma imagen que están asignados a la misma palabra visual y tienen orientación y escala similares. Específicamente, el peso w_M^i para el punto clave i se calcula como:

$$w_M^i = \frac{1}{N_S^i},$$

donde N_S^i indica el número de puntos clave de la misma imagen asignados a la misma palabra visual que i y que tienen la misma orientación y escala. Se considera que dos puntos clave tienen la misma orientación y escala si la diferencia entre sus orientaciones y el factor de ajuste a escala están por debajo de algunos umbrales definidos empíricamente.

Aunque los casos en los que está representado más de un punto clave en la imagen mediante la misma palabra

visual y con orientación y escala similares no son muy habituales, el peso w_M^i desempeña un papel importante a la hora de ajustar la influencia de tales grupos en el proceso de reconocimiento. Su papel exacto se explica con más detalle en la sección que describe el esquema de votación.

El peso de votación final w_K^i asignado al punto clave i se calcula como un producto escalar de los pesos correspondientes a los dos factores de ponderación anteriores: $w_K^i = w_S^i w_S^i$.

La introducción de los pesos anteriores resultó ser muy eficaz en la solución propuesta. Sin embargo, es probable que otros factores de ponderación y/o combinaciones puedan conseguir un efecto similar.

Finalmente, el esquema de ponderación propuesto permite la fácil adición de nuevos factores de ponderación. En el futuro esto podría permitir la incorporación de ubicación espacial de puntos clave (por ejemplo, podría asignarse más importancia a coincidencias que se encontraran más cerca del centro de la imagen) u orientación (por ejemplo, podría asignarse menos importancia a puntos clave con una orientación muy común dentro de la imagen).

- **Construcción de estructura de archivo invertido**

El objetivo de la fase de indexación es organizar las características locales extraídas de imágenes de referencia de manera que se permita su rápida correspondencia con características extraídas de las imágenes de consulta. Como se demuestra en [SZ03, NS06] una de las claves para el reconocimiento rápido de objetos es la organización de las características locales en lo que se denomina la estructura de archivo invertido. De manera interesante, esta solución estuvo motivada por los motores de búsqueda textual populares, tales como el descrito en [BP98]. En el caso de la recuperación de texto, el archivo invertido tiene una entrada (lista de coincidencias) para cada palabra textual, almacenando cada lista todas las apariciones de la palabra en todos los documentos. En el caso de la búsqueda visual, la estructura tiene una lista de coincidencias para cada palabra visual que almacena todas las apariciones de la palabra en todas las imágenes de referencia. Debe observarse que, si el diccionario es suficientemente grande en comparación con el número de imágenes de referencia, las listas de coincidencias son relativamente cortas, lo que lleva a una correspondencia muy rápida.

En el presente enfoque se incorporaron algunas extensiones de la estructura de archivo invertido que son favorables para la solución de correspondencia. Como en [SZ03, NS06], en el archivo invertido hay una lista para cada palabra visual que almacena todas las apariciones (coincidencias) de la palabra visual en todas las imágenes de referencia – véase la figura 5. Como en enfoques anteriores, cada coincidencia corresponde a un punto clave de una imagen de referencia, es decir cada coincidencia almacena el identificador de la imagen que describe. Sin embargo, en el presente caso, cada coincidencia almacena también información adicional acerca de la escala de punto clave, la orientación y la fuerza de votación.

Debe recalcarse que la información almacenada en las coincidencias no solo se usa para limitar el número de imágenes comparadas (como se describe en [SZ03, NS06]), sino que desempeña un papel central fundamental en el proceso de reconocimiento de objetos.

Reconocimiento de objetos

La identificación de objetos presentes en la imagen de consulta se inicia con las mismas cuatro etapas que la indexación de imágenes de referencia – véase la perspectiva general del proceso de reconocimiento en la figura 3. El proceso comienza con (i) extracción de puntos clave y (ii) postprocesamiento, como se describe en la sección “Extracción de características y postprocesamiento”. A continuación, los puntos clave extraídos (iii) se asignan a palabras visuales (véase la sección “Clasificación de puntos clave” para más detalles) y (iv) se calculan los pesos de votación para todos los puntos clave. Debe observarse que la asignación de un punto clave de consulta a una palabra visual es efectivamente equivalente a asignar el punto clave a una lista completa de coincidencias asociadas con la misma palabra visual. Una vez completadas las cuatro etapas, se inicia una (v) agregación de votos para diferentes imágenes de referencia. Cada emparejamiento de un punto clave de la imagen de consulta y una de las coincidencias asignadas a la misma palabra visual da un voto a un acumulador de posición correspondiente a la imagen de referencia en la que se encontró la coincidencia. En otras palabras, cada par (punto clave de consulta, coincidencia) vota por la presencia de un objeto de referencia que aparece con una rotación y factor de escala específicos. La fuerza de cada voto se calcula como un producto escalar de los pesos del punto clave de consulta y la coincidencia. Una vez dados todos los votos, (vi) los acumuladores que han recibido al menos un voto se escanean para identificar grupos con el máximo número de votos. Los valores acumulados en estos grupos se toman como las puntuaciones de relevancia finales para las imágenes de referencia correspondientes. Finalmente, (vii) se ordenan las imágenes de referencia según sus puntuaciones de correspondencia y los objetos más relevantes se seleccionan basándose en una extensión del procedimiento de aplicación de umbral dinámico de [ROS01].

A continuación, se describirán con más detalle las etapas específicas del proceso de correspondencia.

- **Estimación de pesos de puntos clave**

En el caso de imágenes de consulta, los pesos de votación asociados con puntos clave se calculan basándose únicamente en el número de puntos clave en la misma imagen asociados con la misma palabra visual y que tienen escala y orientación similares.

Por tanto, el factor de ponderación w_{QK}^{i} para un punto clave i se calcula como:

$$w_{QK}^{i} = \frac{1}{N_S^i},$$

donde N_S^i indica el número de puntos clave de la imagen de consulta que se asignan a la misma palabra visual que i y tienen orientación y escala similares.

Debe observarse que la exclusión de escalas de la ponderación en el caso de imágenes de consulta permite el reconocimiento de objetos presentes en la escena independientemente de su tamaño. Al mismo tiempo, la inclusión de escalas en la ponderación de coincidencias de imágenes de referencia da más importancia a las coincidencias que normalmente son más discriminatorias sin afectar a la capacidad de reconocer pequeños objetos – véase la sección “Estimación de pesos de puntos clave” para la indexación de imágenes de referencia.

- **Votación**

La fase de votación es el componente más distintivo del enfoque propuesto en comparación con los procedimientos descritos en la literatura. La idea principal es imponer cierta consistencia de posición (rotación y factor de escala) entre los puntos clave que se han hecho corresponder usando el vocabulario de palabras visuales y la estructura de archivo invertido. Esta solución es posible debido a que, en el presente caso, las coincidencias almacenan no solo identificadores de las imágenes de referencia correspondientes, sino también la orientación y la escala de los puntos clave originales. Esta información adicional permite la estimación de la rotación y el ajuste a escala entre puntos clave de la imagen de consulta y las coincidencias correspondientes a diferentes imágenes de referencia. En otras palabras, para cada hipótesis de correspondencia (par de un punto clave de consulta y una coincidencia) puede crearse la entrada de transformada que predice la rotación y el ajuste a escala del objeto de referencia.

Antes de que pueda iniciarse la votación, un acumulador de votos vacío se asigna a cada imagen de referencia. Los acumuladores se implementan como tablas bidimensionales en las que cada celda (grupo) corresponde a una rotación y ajuste a escala particulares del objeto de referencia. Esta estructura simplemente cuantifica los parámetros de transformación de posición de objetos de referencia. Una dimensión del acumulador corresponde a la rotación del objeto de referencia y la otra a su ajuste a escala.

Como se explicó anteriormente, la asignación de una palabra visual a un punto clave de la imagen de consulta es equivalente efectivamente a la asignación de una lista completa de coincidencias de imágenes de referencia correspondientes a la misma palabra visual. Pares (punto clave de consulta, coincidencia) resultantes de la asignación proporcionan hipótesis de correspondencia.

Durante el proceso de votación, cada hipótesis de correspondencia (emparejamiento de un punto clave de la consulta y una de las coincidencias asignadas a la misma palabra visual) da un voto al acumulador correspondiente a la imagen de referencia en la que se encontró la coincidencia. Además, cada par de este tipo (punto clave de consulta, coincidencia) vota no solo por la presencia de un objeto de referencia, sino de hecho por su apariencia con una transformación de rotación y ajuste a escala específicas.

Como ya se explicó anteriormente, el esquema de ponderación tiene en cuenta la presencia de grupos de puntos clave asignados a la misma palabra visual y que tienen orientación y escala similares. El motivo de este factor de ponderación adicional puede explicarse mejor analizando en detalle el esquema de votación. Idealmente, un par de puntos clave correspondientes (uno de la consulta y el otro de la imagen de referencia) darían un voto al acumulador correspondiente a la imagen de referencia. Sin embargo, en casos en los que se asignan múltiples coincidencias de una imagen de referencia a la misma palabra visual y con orientación y escala similares, cada punto clave de la imagen de consulta asignado a la misma palabra visual dará múltiples votos (uno con cada una de tales coincidencias) al mismo grupo de acumuladores. Por ejemplo, si un objeto de referencia genera tres puntos clave representados por la misma palabra visual y con la misma orientación y escala, entonces cada punto clave de la consulta que también se haya asignado a la misma palabra visual dará tres votos (en lugar de uno) al mismo grupo de acumuladores. El esquema de ponderación simplemente garantiza que los múltiples votos dados por tales grupos desempeñen el papel adecuado en el cálculo de las puntuaciones de correspondencia.

- **Cálculo de puntuaciones**

Una vez dados todos los votos, los acumuladores se escanean para identificar los grupos con el máximo número de votos. Los votos acumulados en estos máximos se toman como las puntuaciones de correspondencia finales, es decir, las puntuaciones que indican en qué medida las imágenes de referencia correspondientes a los acumuladores

en los que se encontraron estos máximos corresponden a la imagen de consulta. En otras palabras, para una consulta dada, la puntuación de correspondencia para cada imagen de referencia se obtiene tomando los votos acumulados en el grupo con el máximo número de votos encontrado en el acumulador correspondiente a esta imagen de referencia. Debe observarse que estos grupos representan las transformaciones de posición más probables (es decir, rotación y ajuste a escala) entre las imágenes de consulta y las imágenes de referencia correspondientes.

Debe observarse que el enfoque propuesto tiene como objeto principalmente detectar la presencia o ausencia de objetos de referencia en la imagen de consulta. Por tanto, basta con identificar solo el grupo más votado en cada acumulador e ignorar las múltiples apariciones del mismo objeto de referencia. Debe observarse que la identificación de posiciones de todas las instancias del mismo objeto de referencia requeriría la identificación de todos los máximos locales en el acumulador correspondiente.

- **Ordenación y selección de los objetos de referencia relevantes**

La última fase de la búsqueda implica la ordenación y selección de los resultados que son relevantes para la imagen de consulta. En muchas aplicaciones, esta tarea puede reducirse a una selección trivial del objeto de referencia que obtuvo la puntuación más alta.

Por el contrario, el presente enfoque puede identificar múltiples objetos relevantes presentes en la consulta, véanse los resultados ejemplares en la figura 10. La lista de objetos devuelta se ordena según las puntuaciones obtenidas. Además, el sistema no devuelve ningún resultado en casos en los que no hay presentes objetos relevantes en la imagen de consulta.

En otras palabras, el objetivo de esta fase es usar las puntuaciones de correspondencia producidas en fases anteriores para identificar solo los objetos más destacados presentes en la consulta y al mismo tiempo evitar devolver resultados irrelevantes. La idea básica que subyace a este enfoque es ordenar las imágenes de referencia según sus puntuaciones de correspondencia y a continuación seleccionar solo los primeros objetos de la lista clasificada usando una extensión del procedimiento de aplicación de umbral dinámico de [ROS01].

Debe observarse que la motivación subyacente a incorporar el umbral dinámico la proporcionó el hecho de que las puntuaciones típicas obtenidas por objetos relevantes pueden variar en un amplio intervalo de valores (desde ~40 para consultas con pocos puntos clave hasta ~300 para consultas con gran número de puntos clave). Puesto que es imposible elegir un umbral fijo que proporcionará resultados significativos para tales casos extremos se propone usar la forma de la curva creada por la lista ordenada de puntuaciones para identificar el umbral más adecuado.

La selección del umbral dinámico comienza con la clasificación de las imágenes de referencia según las puntuaciones de correspondencia obtenidas y la aplicación del procedimiento de aplicación de umbral propuesto en [ROS01]. Esto da como resultado una separación inicial de la lista ordenada en dos grupos: (i) objetos potencialmente relevantes al principio de la lista, y (ii) probablemente objetos irrelevantes en el resto de la lista. Esta etapa va seguida del cálculo de un valor promedio de puntuaciones de la segunda parte de la lista que contiene los objetos potencialmente irrelevantes. Este valor (indicado como T_{ir}) proporciona una puntuación de referencia típica para objetos que son irrelevantes para la imagen de consulta actual. El umbral dinámico T_d se calcula como $T_d = \alpha T_{ir}$, donde el valor de α se ajusta empíricamente a 4. El umbral final T_c se calcula como $T_c = \max(T_d, T_f)$, donde T_f indica un umbral fijo, empíricamente ajustado a 30, que proporciona un valor mínimo del umbral por debajo del cual no es probable encontrar resultados relevantes. T_f garantiza resultados significativos para consultas que dan como resultado normalmente puntuaciones muy bajas y para los que el umbral dinámico puede devolver resultados irrelevantes.

Una vez calculado el umbral final T_c, el sistema clasifica los primeros objetos de referencia que obtuvieron puntuaciones por encima del umbral como que están presentes en la imagen de consulta.

La presente invención se implementa preferentemente por medio de un programa informático adecuado cargado en un procesador de propósito general.

Resultados

Las figuras 6 a 10 contienen resultados seleccionados que demuestran las capacidades más interesantes de la invención. Todos los experimentos se llevaron a cabo con una colección de 70 imágenes de referencia. Normalmente el tiempo requerido para una identificación con éxito no supera los 2 segundos cuando se ejecuta en un PC convencional. Además, el tiempo de reconocimiento aumenta muy lentamente con el tamaño de la colección de imágenes de referencia.

La figura 6 muestra un ejemplo de identificación de un objeto pequeño. La primera columna contiene la imagen de

consulta y las siguientes columnas contienen productos recuperados ordenados de izquierda a derecha según sus puntuaciones.

La figura 7 muestra un ejemplo de identificación de un objeto con posición difícil (inclinación de aproximadamente 45 grados). La primera columna contiene la imagen de consulta y las restantes columnas contienen productos recuperados ordenados de izquierda a derecha según sus puntuaciones. Debe observarse que el segundo producto recuperado tiene una marca comercial idéntica a la de la consulta ("Juver").

La figura 8 muestra un ejemplo de identificación de un objeto oculto. La primera columna contiene la imagen de consulta y las restantes columnas contienen productos recuperados ordenados de izquierda a derecha según sus puntuaciones.

La figura 9 muestra un ejemplo de identificación de un objeto pequeño en una escena abarrotada. La primera columna contiene la imagen de consulta y las restantes columnas contienen productos recuperados ordenados de izquierda a derecha según sus puntuaciones.

La figura 10 muestra un ejemplo de identificación de múltiples objetos pequeños. La primera columna contiene la imagen de consulta y las columnas restantes contienen productos recuperados ordenados de izquierda a derecha y de arriba a abajo según sus puntuaciones.

Aplicación industrial

La invención propuesta permite un tipo novedoso de motores de reconocimiento eficaces que suministran resultados en respuesta a fotografías en lugar de a palabras textuales. Tales motores tienen potencial para convertirse en la llave que permita a la tecnología multitud de aplicaciones industriales.

- **Aplicaciones para teléfonos móviles**

La motivación principal de la presente invención la proporcionó la creencia en un enorme potencial comercial para sistemas que permiten a los usuarios simplemente hacer una fotografía con una cámara de teléfono móvil, enviarla, y recibir servicios relacionados –véase una realización ejemplar de la invención ("búsqueda visual para móviles") en la figura 11. El sistema permite a los usuarios simplemente hacer una fotografía con una cámara de teléfono móvil, enviarla, y recibir servicios relacionados.

Se han realizado muchos esfuerzos para garantizar que la invención propuesta se adecua bien al reconocimiento de una amplia gama de productos en 3D (por ejemplo libros, CD/DVD, productos envasados en tiendas de alimentación), carteles urbanos, fotografías en periódicos y revistas, marcas comerciales, etc. La capacidad anterior permite el desarrollo de una amplia variedad de servicios novedosos para usuarios de teléfonos móviles, que se capitalizarán en la curiosidad del usuario y/o facilitarán lo que se denomina la compra compulsiva. Es fácil imaginar muchos escenarios de casos de uso atractivos en los que los usuarios comprueban información acerca de determinados productos (por ejemplo, comparación de precios) o incluso realizan compras directamente haciendo una foto de un objeto particular. Algunos ejemplos de esta categoría incluyen la compra de contenidos audiovisuales haciendo fotos de sus anuncios en revistas, o la compra de tiques para un concierto musical simplemente haciendo una foto de un cartel urbano. Además, la invención propuesta puede desempeñar un papel enorme en el desarrollo de modelos novedosos de publicidad interactiva, por ejemplo, los usuarios pueden participar en un sorteo haciendo una foto de un anuncio que se hayan encontrado en la calle.

En el futuro, la tecnología propuesta podría combinarse con geolocalización, y tecnologías de realidad aumentada que permiten a los usuarios etiquetar y recuperar información acerca de escenas reales simplemente tomando sus teléfonos móviles y haciendo fotografías.

- Otras aplicaciones

* **Detección de duplicados falsos**

La invención podría usarse para la detección de fotos duplicadas falsas, que tiene aplicación en la detección de violación de derechos de autor y archivo de fotos, por ejemplo, organización de colecciones de fotos.

* **Publicidad contextual**

La invención podría usarse para la detección de marcas comerciales que aparezcan en imágenes y vídeos, que podrían aplicar los proveedores de contenido para introducir nuevos modelos de publicidad contextual.

*** Monitorización de publicidad a través de diversos medios**

La invención podría usarse como tecnología principal para herramientas que proporcionen monitorización automática de campañas comerciales a través de diversos tipos de medios tales como, por ejemplo, TV e Internet. Tales herramientas podrían monitorizar automáticamente programas de TV e Internet (tanto contenido generado por el usuario como revistas *online*) en busca de apariciones de marcas comerciales o anuncios particulares de

empresas específicas, por ejemplo, para analizar el impacto de una campaña de comercialización particular.

Aunque la invención se ha ilustrado y descrito en detalle en los dibujos y la descripción anterior, tal ilustración y descripción deben considerarse ilustrativas o ejemplares y no restrictivas; la invención no se limita a las realizaciones divulgadas.

Otras variaciones de las realizaciones divulgadas pueden entenderse y llevarse a cabo por los expertos en la técnica al poner en práctica la invención reivindicada, a partir de un estudio de los dibujos, la divulgación y las reivindicaciones adjuntas. En las reivindicaciones, la expresión “que comprende” no excluye otros elementos o etapas, y el artículo indefinido “un” o “una” no excluye una pluralidad. Un único procesador u otra unidad puede cumplir las funciones de varios elementos mencionados en las reivindicaciones. El mero hecho de que se mencionen ciertas medidas en reivindicaciones dependientes diferentes entre sí no indica que una combinación de estas medidas no pueda usarse de manera ventajosa. Un programa informático puede almacenarse/distribuirse en un medio adecuado, tal como un medio de almacenamiento óptico o un medio de estado sólido suministrado junto con o como parte de otro hardware, aunque también puede distribuirse en otras formas, tales como a través de Internet u otros sistemas de telecomunicación por cable o inalámbricos.

REFERENCIAS

[0097]

[BL97] J. Beis y D. G. Lowe. Shape indexing using approximate nearest neighbor search in high-dimensional spaces. In *Conference on Computer Vision and Pattern Recognition*, *Puerto Rico*, 1997.

[BP98] S. Brin y L. Page. The anatomy of a large-scale hypertextual web search engine. In *Computer Networks and ISDN Systems*, 1998.

[BTG06] Herbert Bay, Tinne Tuytelaars, y Luc Van Gool. Surf: Speeded up robust features. In *ECCV*, 2006.

[BYRN99] R. Baeza-Yates y B. Ribeiro-Neto. Modern information retrieval. In *ACM Press, ISBN: 020139829*, 1999.

[EVO] Evolution. www.evolution.com.

[FLI] Flickr. http://www.flickr.com/.

[HOU62] P.V.C. Hough. Method and means for recognizing complex patterns. In *U.S. Patent 3069654*, 1962.

[KOO] Kooaba. http://www.kooaba.com.

[LOW99] D. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[LOW04] D. Lowe. Distinctive image features from scale-invariant keypoints, cascade Þltering approach. In *IJCV*, 2004.

[LSDJ06] M. Lew, N. Sebe, Ch. Djeraba, y R. Jain. Content-based multimedia information retrieval: State of the art and challenges. In *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2006.

[MCUP02] J. Matas, O. Chum, M. Urban, y T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. In *Proc. of the British Machine Vision Conference, Cardiff, UK*, 2002.

[MS04] K. Mikolajczyk y C. Schmid. Scale and affine invariant interest point detectors.In *IJCV*, 2004.

[NS06] D. Nister y H. Stewenius. Scalable recognition with a vocabulary tree. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[PCI+07] J. Philbin, O. Chum, M. Isard, J. Sivic, y A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. CVPR*, 2007.

[PCI+08] J. Philbin, O. Chum, M. Isard, J. Sivic, y A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. 2008.

[ROS01] P. Rosin. Unimodal thresholding. In *Pattern Recognition, vol. 34, no. 11, págs. 2083-2096*, 2001.

[SIV06] Josef Sivic. Efficient visual search of images and videos. In *PhD thesis at University of Oxford*,

2006.

[SUP] Superwise. www.superwise-technologies.com.

[SZ03] J. Sivic y A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, 2003.

REIVINDICACIONES

1. Procedimiento de identificación de objetos en imágenes, **caracterizado porque** comprende las siguientes fases:

(i) una fase de extracción de características que incluye las siguientes etapas para: tanto imágenes de referencia, es decir imágenes que representan, cada una, al menos un objeto de referencia individual, como al menos una imagen de consulta, es decir una imagen que representa objetos desconocidos que han de identificarse:

(a) identificación de puntos clave, es decir zonas de imagen destacadas;

(b) postprocesamiento de puntos clave

(c) cálculo de los descriptores, es decir representaciones, de los puntos clave,

(ii) una fase de indexación de imágenes de referencia que incluye las siguientes etapas:

(a) extracción de puntos clave;

(b) postprocesamiento de puntos clave

(c) asignación de puntos clave a palabras visuales de un vocabulario de palabras visuales creado a partir de una colección de imágenes de entrenamiento, en la que las palabras visuales son centros de agrupamientos de descriptores de puntos clave;

(d) adición de puntos clave a una estructura de archivo invertido, en la que la estructura de archivo invertido comprende una lista de coincidencias para cada palabra visual que almacena todas las apariciones de la palabra en las imágenes de referencia y en la que cada coincidencia almacena un identificador de la imagen de referencia en la que se ha detectado el punto clave; y

(iii) una fase de reconocimiento de objetos presentes en la imagen de consulta que incluye las siguientes etapas:

(a) extracción de puntos clave;

(b) postprocesamiento de puntos clave

(c) asignación de puntos clave a palabras visuales del vocabulario de palabras visuales;

(d) para cada emparejamiento de un punto clave de la imagen de consulta y una de las coincidencias asignadas a la misma palabra visual, agregar un voto en un acumulador correspondiente a la imagen de referencia de la coincidencia; e

(e) identificación de las puntuaciones de correspondencia correspondientes a las imágenes de referencia basándose en los votos de los acumuladores.

caracterizado porque el postprocesamiento comprende:

- normalizar escalas de puntos clave según el tamaño de los objetos de referencia; y

- eliminar puntos clave con escalas normalizadas inferiores a un valor predefinido.

2. Procedimiento según la reivindicación 1, en el que la fase de reconocimiento (iii) de objetos comprende la etapa adicional de seleccionar un objeto u objetos que son relevantes para la consulta según sus puntuaciones de correspondencia.

3. Procedimiento según la reivindicación 1 o 2, en el que el postprocesamiento incluye la detección automática de zonas de interés basándose en las ubicaciones de puntos clave detectados.

4. Procedimiento según la reivindicación 3, en el que, en el caso de imágenes de referencia, el centro de la zona de interés se considera como el centro de la masa del conjunto de todas las ubicaciones de puntos clave detectadas, su anchura y su altura iniciales se calculan de manera independiente en las direcciones horizontal y vertical en función de la desviación estándar de las ubicaciones de puntos clave, siendo las ubicaciones de puntos clave ponderadas de acuerdo con las escalas de puntos clave normalizadas, y en el que la anchura y la altura iniciales se reducen siempre que la zona de interés abarque áreas sin puntos clave.

5. Procedimiento según la reivindicación 3 o 4, en el que las escalas de los puntos clave se normalizan en función del tamaño de la zona de interés, y se eliminan los puntos clave ubicados fuera de la zona de interés y los puntos clave con una escala normalizada inferior a un valor predeterminado.

6. Procedimiento según la reivindicación 1, en el que las fases (ii) y (iii) incluyen asociar un factor de ponderación a cada punto clave que refleje su importancia en el proceso de reconocimiento de objetos, factor de ponderación que se basa en la escala de puntos clave normalizada.

7. Procedimiento según la reivindicación 6, en el que el factor de ponderación se basa en la escala de puntos clave detectados, siendo dicha escala de puntos clave la escala de puntos clave normalizada y el número de puntos clave de la misma imagen asignados a la misma palabra visual como el punto clave considerado y con orientación y escala similares.

8. Procedimiento según la reivindicación 6 o 7, en el que, en la etapa (iii) (d), el factor de ponderación se usa en el proceso de agregación de votos, factor de ponderación que se basa en la escala de puntos clave normalizada.

9. Procedimiento según la reivindicación 1 o 2, en el que, en la etapa (ii) (d), cada coincidencia almacena, además del identificador de la imagen de referencia en la que se ha detectado el punto clave, información acerca de su escala y orientación y cada coincidencia tiene una fuerza asociada de la evidencia con la que puede confirmar una existencia del correspondiente objeto en respuesta a una aparición de la palabra visual en una imagen de entrada.

10. Procedimiento según la reivindicación 9, en el que, en la etapa (iii) (d), el acumulador correspondiente a la imagen de referencia se implementa como una tabla bidimensional en la que una dimensión del acumulador corresponde a la rotación del objeto de referencia y la otra dimensión al ajuste a escala del objeto de referencia, de modo que cada celda corresponde a una rotación y ajuste a escala particular del objeto de referencia y en el que un voto es para la apariencia del objeto de referencia con una transformación de rotación y ajuste a escala específicas.

11. Procedimiento según la reivindicación 10, en el que, en la etapa (iii) (e), se identifica la celda con el máximo número de votos en cada acumulador.

12. Procedimiento según la reivindicación 11, en el que, en la etapa (iii) (f), la imagen de referencia correspondiente con la mayor puntuación de correspondencia se selecciona como el objeto más relevante.

13. Procedimiento según la reivindicación 10, en el que se escanean los acumuladores con el fin de identificar grupos con el máximo número de votos y los votos acumulados en estos máximos se toman como las puntuaciones de correspondencia finales, es decir puntuaciones que indican en qué medida las imágenes de referencia correspondientes a los acumuladores en los que se encontraron estos máximos corresponden a la imagen de consulta.

14. Programa informático que comprende medios de código de programa informático adaptados para realizar las etapas según una cualquiera de las reivindicaciones 1 a 13 cuando dicho programa se ejecuta en un ordenador.

15. Sistema que comprende medios adaptados para realizar las etapas según una cualquiera de las reivindicaciones 1 a 14.

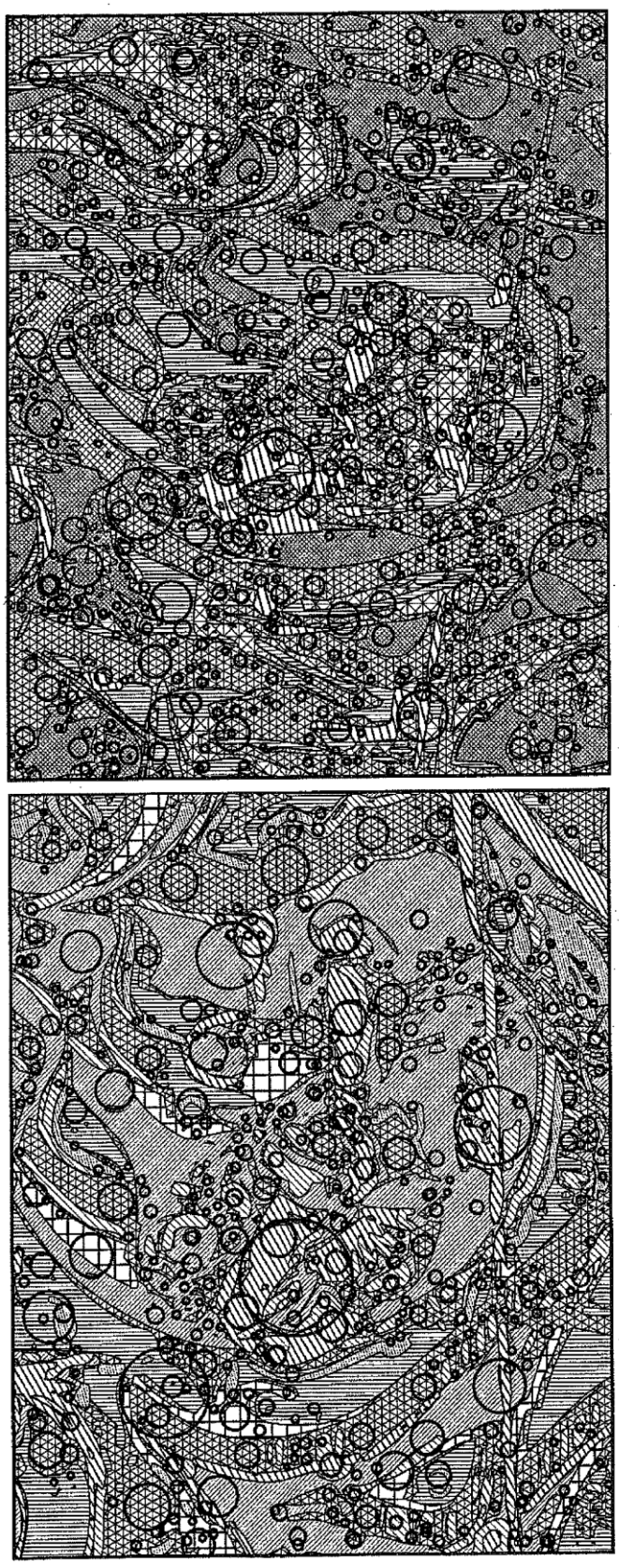

FIG. 1

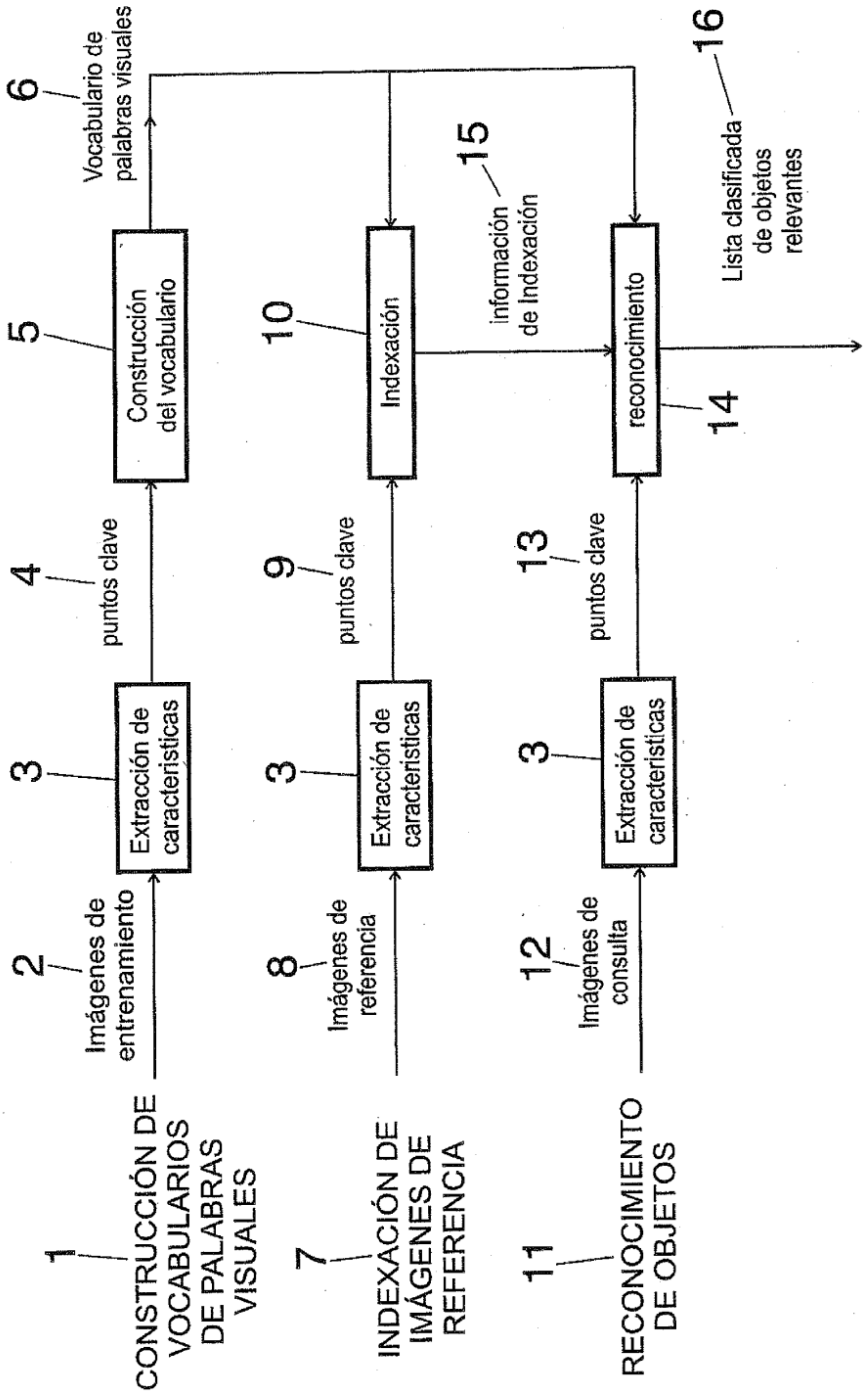
1
CONSTRUCCIÓN DE VOCABULARIOS DE PALABRAS VISUALES
2
Imágenes de entrenamiento
3
Extracción de características
4
puntos clave
5
Construcción del vocabulario
6
Vocabulario de palabras visuales
7
INDEXACIÓN DE IMÁGENES DE REFERENCIA
8
Imágenes de referencia
9
puntos clave
10
Indexación
11
RECONOCIMIENTO DE OBJETOS
12
Imágenes de consulta
13
puntos clave
14
reconocimiento
15
información de Indexación
16
Lista clasificada de objetos relevantes

FIG. 2

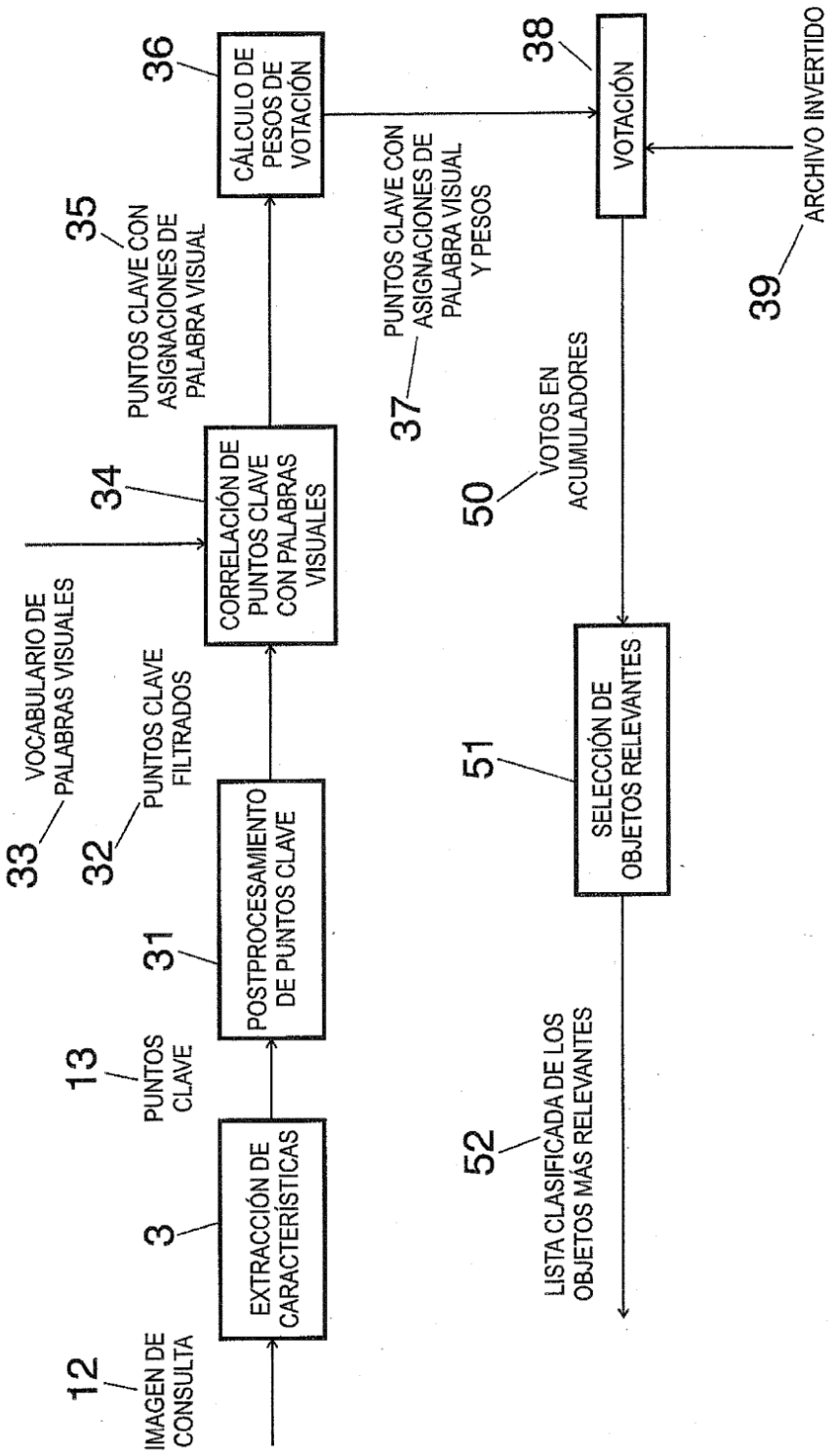
12
IMAGEN DE CONSULTA
3
EXTRACCIÓN DE CARACTERÍSTICAS
13
PUNTOS CLAVE
31
POSTPROCESAMIENTO DE PUNTOS CLAVE
32
PUNTOS CLAVE FILTRADOS
33
VOCABULARIO DE PALABRAS VISUALES
34
CORRELACIÓN DE PUNTOS CLAVE CON PALABRAS VISUALES
35
PUNTOS CLAVE CON ASIGNACIONES DE PALABRA VISUAL
36
CÁLCULO DE PESOS DE VOTACIÓN
37
PUNTOS CLAVE CON ASIGNACIONES DE PALABRA VISUAL Y PESOS
38
VOTACIÓN
39
ARCHIVO INVERTIDO
50
VOTOS EN ACUMULADORES
51
SELECCIÓN DE OBJETOS RELEVANTES
52
LISTA CLASIFICADA DE LOS OBJETOS MÁS RELEVANTES

FIG. 3

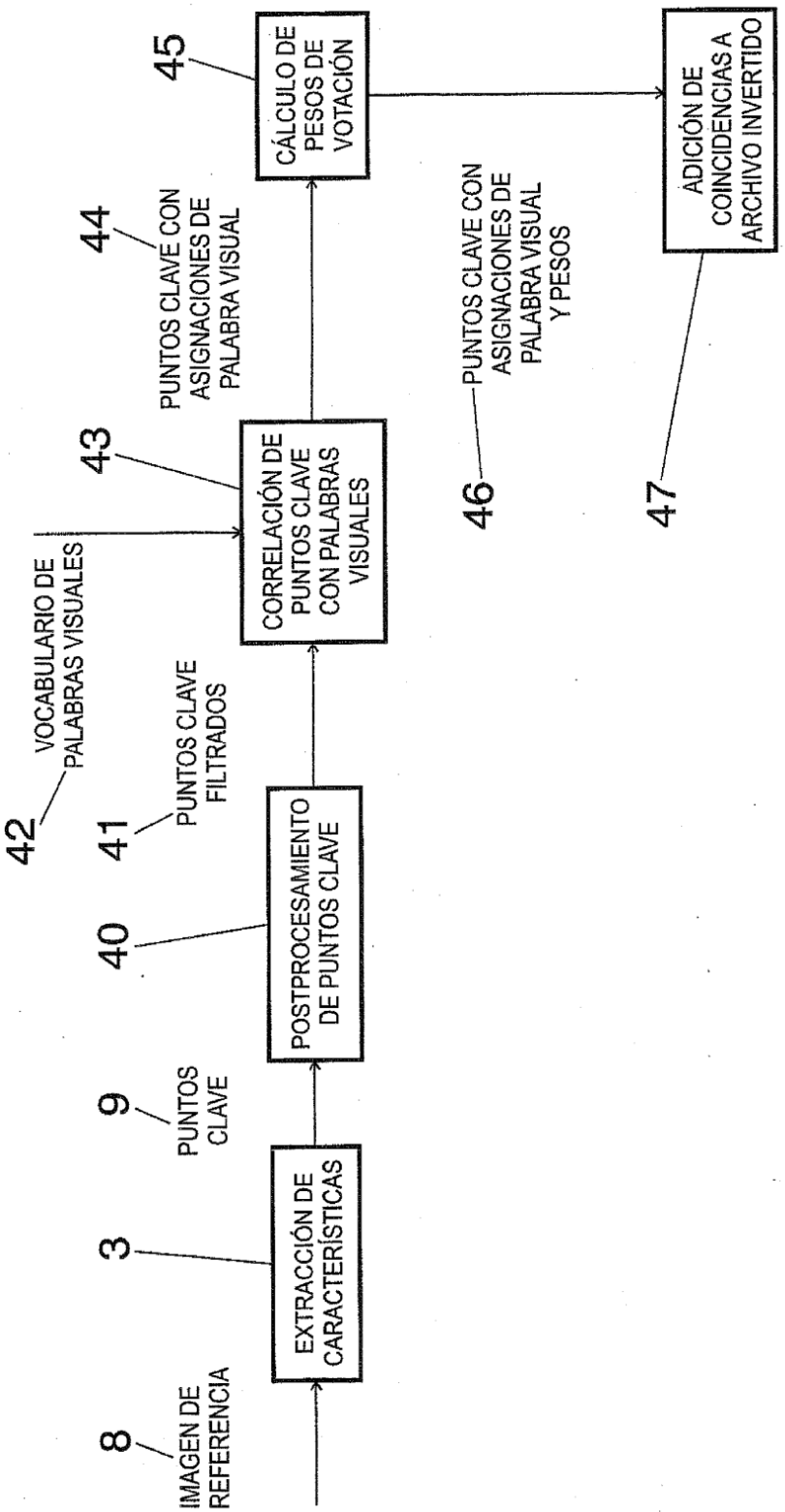
8
IMAGEN DE REFERENCIA
3
EXTRACCIÓN DE CARACTERÍSTICAS
9
PUNTOS CLAVE
40
POSTPROCESAMIENTO DE PUNTOS CLAVE
41
PUNTOS CLAVE FILTRADOS
42
VOCABULARIO DE PALABRAS VISUALES
43
CORRELACIÓN DE PUNTOS CLAVE CON PALABRAS VISUALES
44
PUNTOS CLAVE CON ASIGNACIONES DE PALABRA VISUAL
45
CÁLCULO DE PESOS DE VOTACIÓN
46
PUNTOS CLAVE CON ASIGNACIONES DE PALABRA VISUAL Y PESOS
47
ADICIÓN DE COINCIDENCIAS A ARCHIVO INVERTIDO

FIG. 4

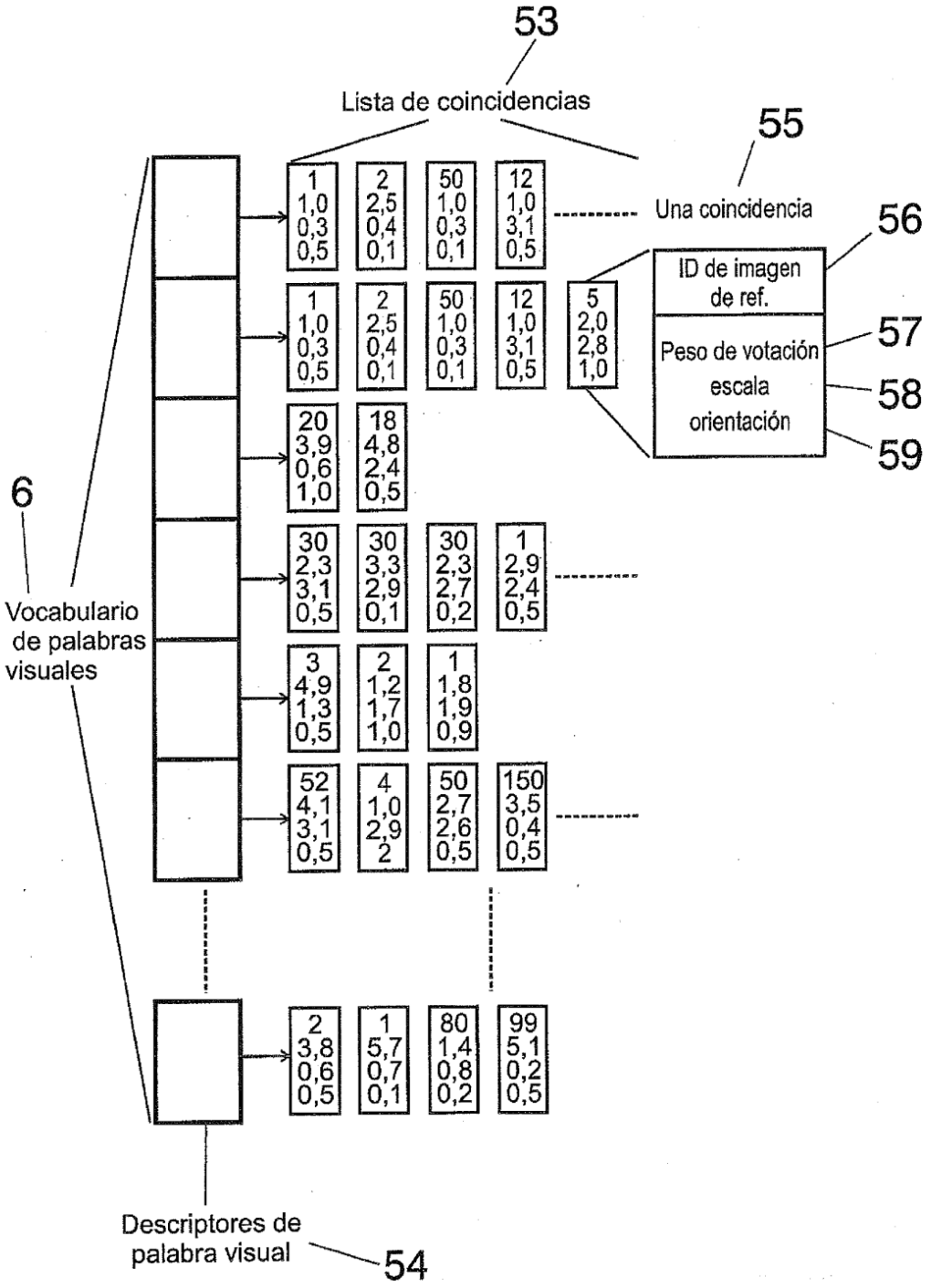
53
Lista de coincidencias
55
Una coincidencia
56
ID de imagen de ref.
57
Peso de votación
58
escala
59
orientación
6
Vocabulario de palabras visuales
Descriptores de palabra visual
54

FIG. 5

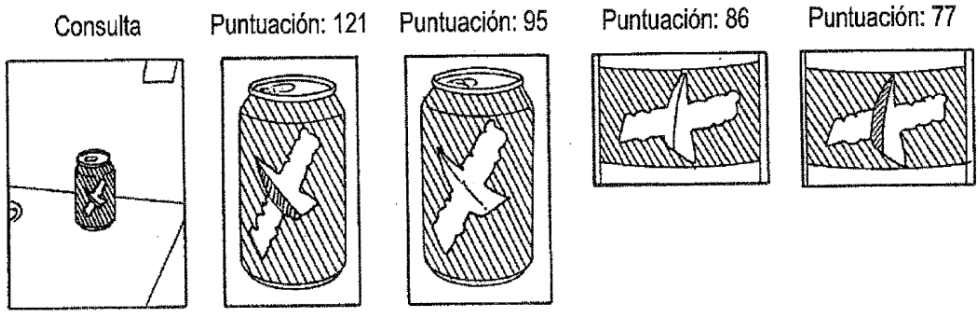
Consulta
Puntuación: 121
Puntuación: 95
Puntuación: 86
Puntuación: 77

FIG. 6

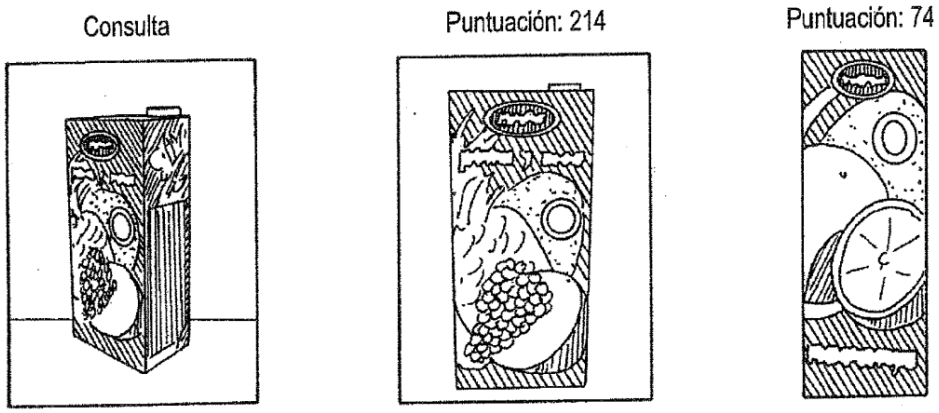
Consulta
Puntuación: 214
Puntuación: 74

FIG. 7

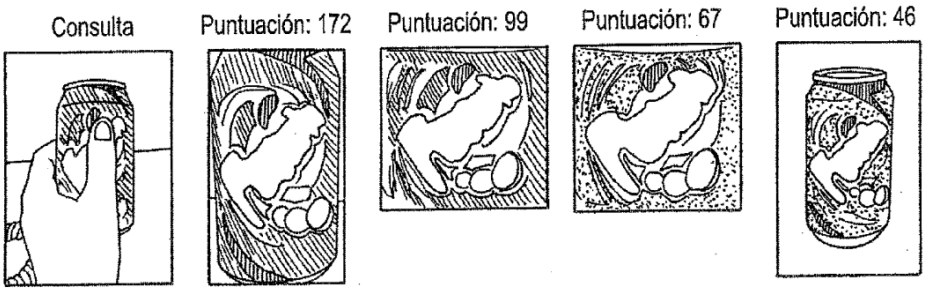
Consulta
Puntuación: 172
Puntuación: 99
Puntuación: 67
Puntuación: 46

FIG. 8

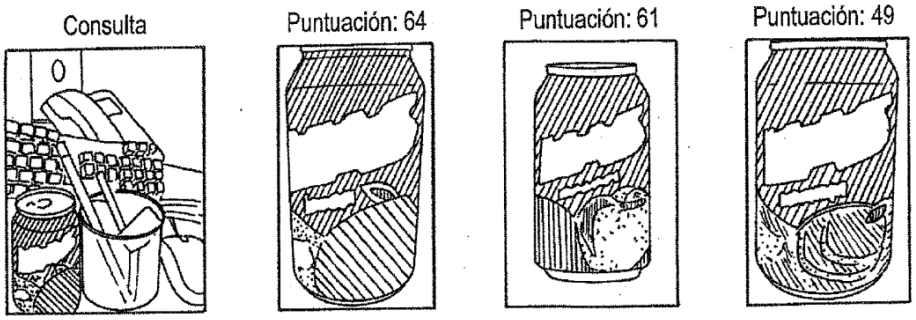
Consulta
Puntuación: 64
Puntuación: 61
Puntuación: 49

FIG. 9

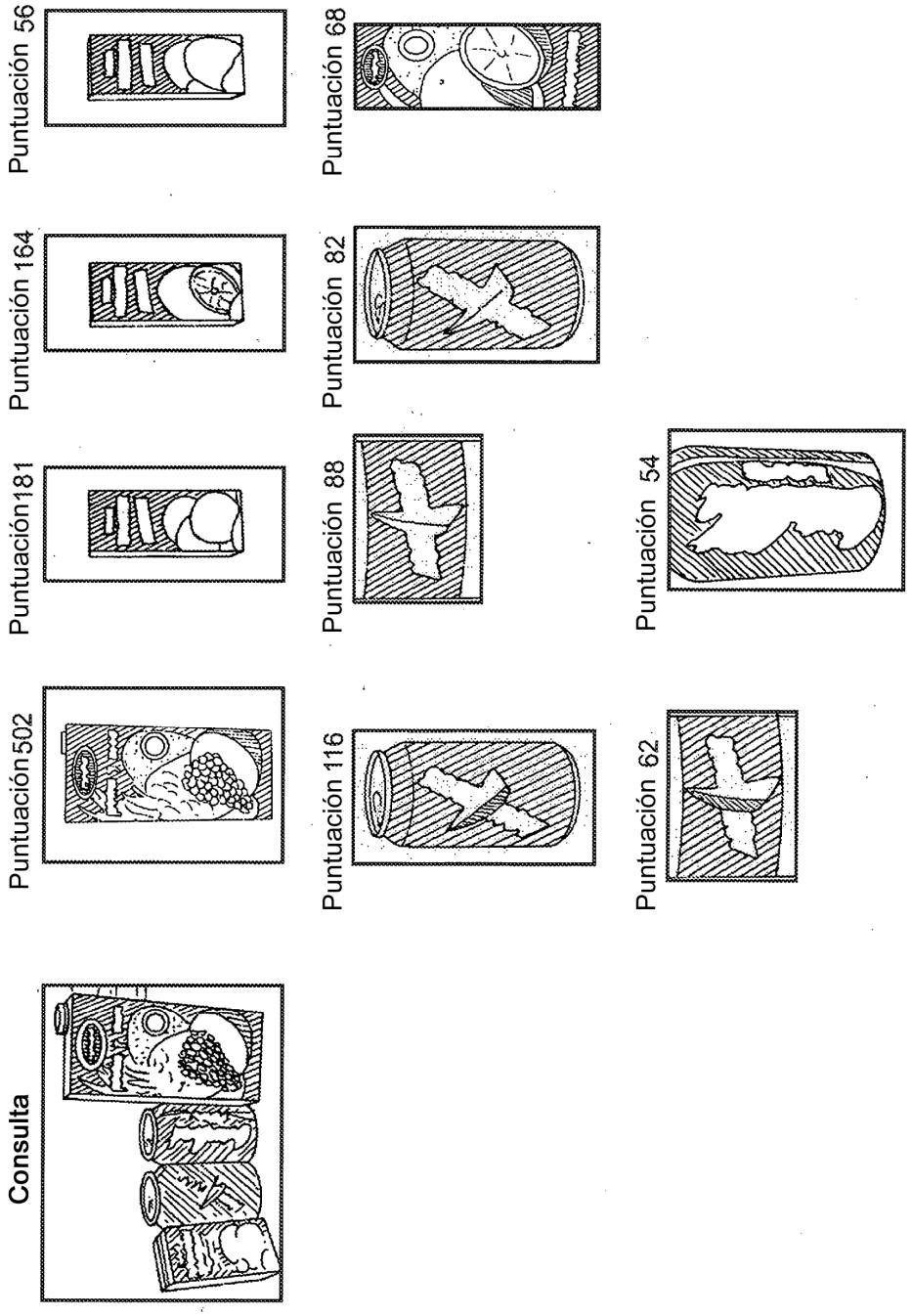
Consulta
Puntuación 502
Puntuación 181
Puntuación 164
Puntuación 56
Puntuación 116
Puntuación 88
Puntuación 82
Puntuación 68
Puntuación 62
Puntuación 54

FIG. 10

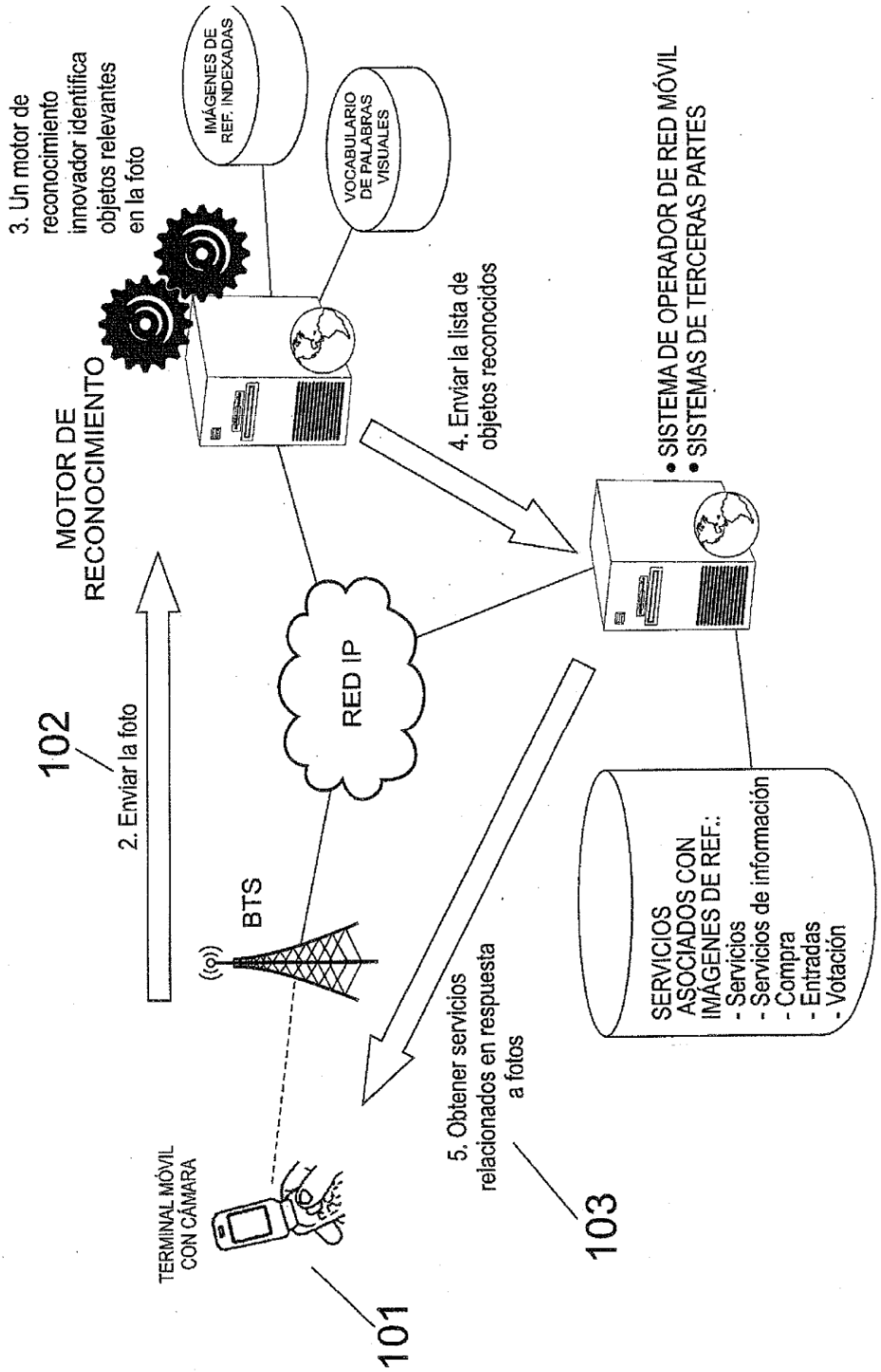
3. Un motor de reconocimiento innovador identifica objetos relevantes en la foto
IMÁGENES DE REF. INDEXADAS
VOCABULARIO DE PALABRAS VISUALES
MOTOR DE RECONOCIMIENTO
4. Enviar la lista de objetos reconocidos
• SISTEMA DE OPERADOR DE RED MÓVIL
• SISTEMAS DE TERCERAS PARTES
102
2. Enviar la foto
RED IP
BTS
SERVICIOS ASOCIADOS CON IMÁGENES DE REF.:
- Servicios
- Servicios de información
- Compra
- Entradas
- Votación
5. Obtener servicios relacionados en respuesta a fotos
TERMINAL MÓVIL CON CÁMARA
101
103

FIG. 11