ISSN 2007-9737

Reconocimiento robusto de lugares mediante redes neuronales convolucionales

Omar E. Lugo Sánchez[1], Humberto Sossa[1,2], Erik Zamora[1]

[1] Instituto Politécnico Nacional,
Centro de Investigación en Computación,
México

[2] Tecnológico de Monterrey,
México

lugo.oe@email.org, hsossa@cic.ipn.mx, ezamorag@ipn.mx

Resumen. En este trabajo se propone utilizar una red neuronal convolucional para el reconocimiento visual de lugares. El trabajo se enfoca en la identificación y extracción automática de regiones de interés a partir de una imagen consulta. Estas regiones de interés se usan para construir la codificación de la imagen mediante un vector de descriptores localmente agregados, el cual es usado para la recuperación de imágenes. A diferencia de otros métodos, donde la imagen completa es utilizada para crear la codificación, el enfoque propuesto solo usa las regiones de interés más importantes en la imagen. Esto otorga una mejor robustez ante cambios extremos de vista en la imagen, condiciones de iluminación y oclusiones. Otra aportación de este trabajo consiste en la integración de un transformador convolucional espacial de acuerdo a la arquitectura de la red neuronal convolucional. Este transformador es usado para la normalización de las regiones de interés, lo que favorece una mayor robustez en la codificación. También se propone una función de pérdida que se usa para entrenar la red neuronal artificial para identificar las regiones de manera automática. Para medir la eficiencia del modelo propuesto, se realizó una variedad de experimentos con un conjunto desafiante de datos desafiantes. Los resultados reportados muestran que el método propuesto produce resultados superiores respecto a los otros métodos del estado del arte.

Palabras clave. Red neuronal convolucional, vector de descriptores localmente agregados, reconocimiento visual de lugares.

Robust Place Recognition using Convolutional Neural Networks

Abstract. In this work, we propose using Convolutional Neural Network for place visual recognition. The work focuses on the identification and automatic extraction of interest regions from a query image. These regions are used to build an image encoding through a vector of locally aggregated descriptors, which is turn used for image recovery. Unlike other methods, where the entire image is used to create the encoding, our approach only uses the most important image interest regions. This provides better invariance to changes at extreme view points of view, lighting and occlusions. Another contribution of the work consists in the integration of a totally convolutional spatial transformer according to the convolutional neural network architecture. This transformer is used for normalizing these interest regions, which allows achieving a greater robustness during coding. A loss function is also proposed that is used to train the artificial neural network to identify automatically regions. To measure the efficiency of the proposed model, a variety of experiments were carried out with challenging data sets. The reported results show that the proposed method produces superior results than other state of the art methods.

Keywords. Convolutional neural network, vector of locally aggregated descriptors, visual place recognition.

1. Introducción

Dada una imagen de consulta, el reconocimiento visual de lugares puede ser vista como la tarea de recuperar imágenes [3, 7, 20, 35, 44, 45]. Esta tarea consiste en determinar la coincidencia entre una imagen consulta y las imágenes de lugares visitados con anterioridad, almacenadas en una gran base de datos [24]. En los últimos años esta tarea ha recibido

considerable atención por parte de la comunidad de visión por computadora [3, 6, 7, 13, 20] debido a sus importantes aplicaciones, como la navegación automática [25], más específicamente ante la falta de información sobre el posicionamiento global, así como el geo-etiquetado automático y realidad aumentada [27].

Tradicionalmente, cada imagen de la base de datos era codificaba al extraer las características locales robustas [46] mediante métodos como SIFT / SURF / ORB [5, 23, 34]. Estos rasgos eran luego agregados a una única representación vectorial para toda la imagen, esto se logra al usar métodos de codificación como la bolsa de palabras visuales [33, 38], VLAD [2, 18] y Vector de Fisher [19, 32]. La representación resultante es normalmente comprimida e indexada de manera eficiente [17, 38].

Sin embargo, el problema del reconocimiento visual de lugares sigue siendo una tarea desafiante; se busca lograr una representación adecuada que sea lo suficientemente robusta que permita distinguir entre lugares similares aún bajo cambios extremos de punto de vista, oclusión parcial y cambios en la iluminación, pero que también sea lo suficiente eficiente y compacta.

Motivado por el éxito del aprendizaje profundo en el área de visión por computadora [9], el enfoque en el reconocimiento visual de lugares se ha desplazado del uso de técnicas tradicionales de características obtenidas a mano como SIFT / SURF / ORB [5, 23, 34], al uso de características más genéricas obtenidas mediante aprendizaje profundo y extraídas a través de redes neuronales convolucionales (RNC).

Lo que se busca con este enfoque es la utilización del aprendizaje profundo para generar representaciones adecuadas una la imagen desde una RNC previamente entrenada. En general, los enfoques basados en aprendizaje profundo actuales se dividen en dos grandes categorías:

1. La imagen completa es directamente alimentada a la RNC, se extraen las activaciones de las capas convolucionales, estas son agregadas. El resultado es una representación global de la imagen [41, 4].
2. De la imagen se extraen regiones de interés, la RNC previamente entrenada se utiliza sobre cada una de estas regiones. Los resultados de cada una de estas regiones se agregan para crear una representación de imagen final.

Por lo general, el primer enfoque, aplana o agrega directamente las activaciones de una sola capa de la RNC para crear una representación global de la imagen, la cual generalmente resulta en una representación menos robusta frente a efectos como la oclusión parcial o las variaciones severas del punto de vista. Por otro lado, la representación generada por el segundo enfoque, resulta ser más robusta e invariable ante cambios de puntos de vista extremos y oclusiones parciales. Tiende a enfocarse en áreas más importantes de la imagen [40], sin embargo, resulta computacionalmente intensiva y a usar técnicas externas de extracción de regiones de interés.

En [8, 12, 28, 22, 47, 43], considerados como trabajos del estado del arte, se identifican las regiones de interés mediante las activaciones de las capas convolucionales. En [8], los autores utilizaron la red VGG-16 [37] previamente entrenada en el conjunto de datos ImageNet [1]. A pesar de que sus resultados son mejores comparados a los otros métodos, este trabajo presenta varios inconvenientes. Uno es que el modelo RNC fue previamente entrenado con una base de datos centrados en objetos, lo que hace que la RNC intente poner más énfasis en los objetos en vez del lugar en sí, lo que conduce a más casos de falla.

Además, el enfoque propuesto de identificación de regiones de interés dificulta la identificación de regiones estáticas que pueden ser más efectivas bajo condiciones y variaciones de puntos de vista. También, afirma superar a trabajos clásicos como FABMAP [12], SEQSLAM [12] y otros métodos de codificación como Sum-Pooling [47], Max-Pooling [43] y Cross-Pooling [22].

Inspirado en estos resultados, en este artículo se propone un nuevo método de codificación, basado en la identificación automática de las características convolucionales de regiones de interés de un modelo RNC. El trabajo propone el entrenamiento del modelo RNC con una nueva función de perdida diseñada específicamente para la comparación de lugres. La ventaja de este enfoque es que, en lugar de depender de técnicas de extracción de regiones de interés externas o de

Fig. 1. Los recuadros indican regiones de correspondencia entre imágenes bajo dos vistas de la misma escena encontrados por CorVlad-Net (trabajo propuesto). Las correspondencias se ilustran con el mismo color de rectángulo

la selección manual de la capa convolucional, el método propuesto identifica las regiones de interés automáticamente en la última capa convolucional. Para lograr mejores resultados se agrega un transformador espacial totalmente convolucional para la normalización de las áreas de interés, lo cual fue inspirado por el trabajo [16]. Esto ayuda a lograr una mayor robustez, que hasta donde se sabe, este es el primer trabajo que combina estos enfoques en la tarea de recuperación de imágenes. En contraste con otros métodos que usan modelos previamente entrenados en conjunto de datos enfocado en objetos [1], el modelo de RNC descrito en este trabajo, se entrenó con un conjunto de datos centrado en lugares [30].

El enfoque propuesto representa cada imagen como un conjunto de regiones de interés (es decir, parches rectangulares) que luego se codifican con el método VLAD. La motivación para usar VLAD [18, 2] viene de los resultados superiores obtenidos en comparación a otros enfoques como bolsa de palabras [38] y otros como [12, 28, 22, 47, 43]. La evaluación de estas regiones permite la comparación de dos imágenes y determinar su similitud.

Esto se ilustra en la Figura 1. Para determinar la eficiencia trabajo propuesto, se evaluó comparándolo con otros algoritmos de reconocimiento de lugares considerado en el estado del arte, en varios conjuntos de datos de referencia, los cuales exhiben variaciones de apariencia, iluminación, oclusión y punto de vista.

A partir de este momento, y por facilidad, nos referiremos a nuestro trabajo como CorVlad-Net (*Correspondence Vlad Network*).

En particular, se realizan las siguientes contribuciones principales:

- Aprendizaje mediante el uso de la función de pérdida de correspondencias propuesta en este trabajo, la cual permite identificar las áreas de interés directamente, las cuales se usan para crear la codificación de la imagen.
- La integración de un transformador espacial totalmente convolucional en la arquitectura de la RNC, el cual se utiliza en la normalización de las regiones de interés, lo cual logra una codificación más robusta.
- Un modelo RNC entrenado y ajustado con un conjunto de datos centrado en lugares [30], a diferencia de otros enfoques que entrenados con datos centrados en objetos [1].
- Un sistema de aprendizaje profundo de reconocimiento visual de lugares basado en la región de interés que puede abordar variaciones tanto en el punto de vista, oclusiones e iluminación.

El resto del artículo está organizado de la siguiente manera. En la sección 2 se presenta el trabajo relacionado en el reconocimiento de imágenes. En la sección 3, se presenta a detalle la metodología propuesta. La sección 4 ilustra y discute los resultados logrados en varios conjuntos de datos de referencia. En la sección 5 se enumeran las conclusiones obtenidas.

2. Trabajos relacionados

En esta sección se describen los principales trabajos con el reconocimiento visual de lugares, así como los métodos que se han desarrollado para codificar características basadas en RNC.

Los primeros trabajos [12, 28] utilizaban características locales como SIFT / SURF / ORB [5, 23, 34]. Esta información luego era codificada principalmente mediante la representación de bolsa de palabras [33]. Estos trabajos mostraron buenos resultados bajo cambios de punto de vista, ya que aprovechaban las propiedades de invariancia de los descriptores. Se convirtieron en el estado del arte en el área. Las investigaciones se enfocaron en diseñar mejores características

de este tipo. Sin embargo, motivado por el éxito del aprendizaje profundo en el ara de visión por computador, los primeros trabajos que atacaron el problema haciendo uso de un modelo RNC fueron [9, 41, 36, 40, 31]. En [9, 36], los autores utilizaron modelos previamente entrenados en el conjunto de datos centrado en objetos ImageNet [1]. Para realizar la comparación entre imágenes utilizaban diferentes medidas de distancia usando las activaciones de los mapas de características de la RNC. Esto trabajos han mostrado resultados muy superiores a los que métodos clásicos que utilizaban características diseñadas a mano como SIFT / SURF / ORB [5, 23, 34].

En [10], los autores introdujeron dos modelos tipo RNC diseñados específicamente para la tarea de reconocimiento de lugares, los cuales han sido entrenados y ajustados en un conjunto de datos SPED [8], y centrados en el reconocimiento de lugares. El conjunto de datos SPED consiste en miles de imágenes de lugares con variaciones de el mismo lugar durante diferentes épocas del año. Los resultados mostrados han sido muy superiores respecto a otros trabajos.

Los enfoques mencionados anteriormente extraen representaciones globales de la imagen completa, las cuales mostraban resultados superiores, pero presentaban problemas en los casos de oclusiones parciales de la estructura de la escena o variaciones severas del punto de vista. En cambio, los enfoques que dividían la imagen en regiones más pequeñas (siendo estas las que mejor caracterizaban a la imagen) obtenían mejores resultados en esos casos.

En el trabajo reportado en [26], los autores hacían uso de este enfoque, el cual mostraba ser más robusto frente a las variaciones de escala y punto de vista. Los enfoques presentados en [40, 42], combinaban un detector de regiones de interés externo, los cuales también mostraban buenos resultados. Nuestro trabajo CorVlad-Net se inspira en estos métodos, como la representación de la imagen basada en regiones de interés. En contraste, nuestro enfoque no requiere ninguna técnica externa de extracción de regiones de interés.

El enfoque propuesto identifica directamente las regiones de interés en la última capa convolucional. Unos de los inconvenientes en este enfoque es que las RNCs solo son robustas para algunos tipos de transformaciones como la traslación y en menor medida al escalado, debido a las operaciones en las capas como convolución y agrupación. Sin embargo, la gestión explícita de tales variaciones se ha manejado aumentando más los datos de entrenamiento con tales variaciones [15].

Por lo que es deseable poder agregar más invariabilidad a estas áreas de interés sin tener que aumentar los datos. En [16] atacan este problema y proponen una red de tranformadores espaciales que permite a la RNC aprender a escalar, rotar o aplicar transformaciones arbitrarias a un objeto de interés. Este enfoque, a nuestro conocimiento, nunca había sido utilizado en el reconocimiento visual de lugares previamente. Esta es una de las ideas en que se basa este trabajo, al implementar un módulo de trasformación especial, para lograr una normalización de las regiones de interés, lo cual logra una codificación más robusta.

Normalmente una vez extraídas las características como SIFT y SURF [5, 23] o las regiones de interés [48, 29, 9], se codifican mediante el método de bolsa de palabras (BoW) [38].

Se han desarrollado otros métodos de codificación los cuales utilizan directamente las activaciones de las capas convolucionales, Sum-Pooling [47], Max-Pooling [43] y Cross-Pooling [22], son algunos de estos métodos. Estas codificaciones han mostrado resultados superiores y prometedores en varias tareas de visión por computadora como la clasificación de imágenes y detección de objetos [43, 22].

En otros trabajos [22, 21] se utilizaron otras representaciones como el vector Fisher [32] y el vector de descriptores localmente agregados (del inglés, VLAD) [2], los cuales logran mucho mejores resultados que los otros métodos de codificación mencionados, además de utilizar vocabularios de palabras visuales mucho más reducidos. Por esta razón, en este trabajo se utiliza la codificación VLAD en nuestra propuesta CorVlad-Net.

3. Método propuesto

En esta sección, se describen en detalle los pasos de la metodología propuesta. Se comienza

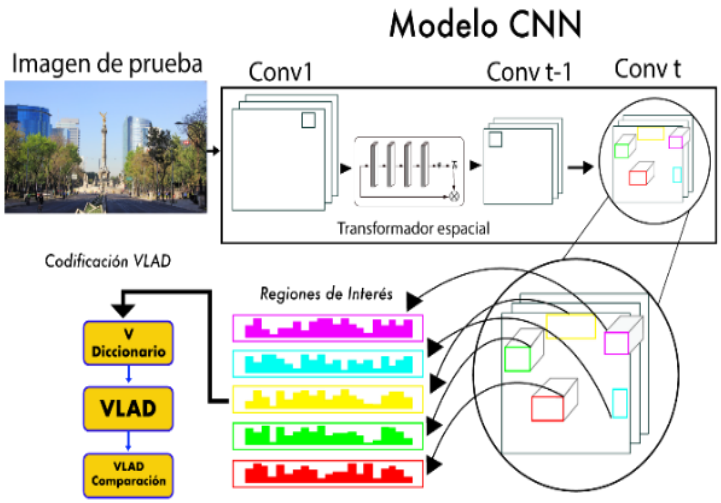

Fig. 2. Esquema del método propuesto. Las imágenes de prueba / referencia se presentan al modelo RNC, luego se identifican las regiones de interés en la última capa de convolución y se construye una representación compacta con VLAD, la cual se almacena para su posterior comparación

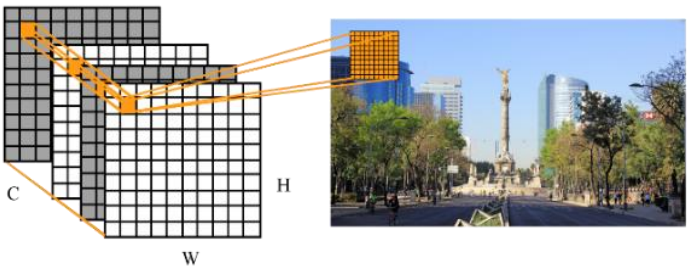

Fig. 3. Hay C mapas de características el cual tiene la forma de una matriz de $H \times W$. A cada activación en el mapa de características le corresponde a una posición en la imagen original

con la idea de la correspondencia espacial entre el mapa de características de una capa convolucional y la imagen y cómo se puede usar como codificación de la misma. Luego se muestran los pasos para la extracción de regiones de interés.

Después se muestran los pasos seguidos para la codificación VLAD y las medidas de similitud usadas para comparar dos imágenes. Finalmente, se muestra la función de perdida usada en el entrenamiento de CorVlad-Net para obtener las regiones de interés. El flujo de trabajo de la metodología propuesta se muestra en la Figura 2.

3.1. Uso del método de características de una imagen

Dada una imagen I, la salida de una capa convolucional de una RNC es un tensor 3D T de dimensiones $H \times W \times C$, con H y W, el ancho y la altura de cada mapa de características, respectivamente y C denota el número de mapas de características.

Como se muestra en la Figura 3, las activaciones en una ubicación espacial en todos los mapas de características C puede ser concatenadas en un descriptor d local C-

dimensional que representa una región en la imagen. El tamaño de esta región es igual al campo receptivo del filtro.

De esta manera, se pueden apilar las activaciones de la capa convolucional, y se pueden considerar como una matriz 2D, donde cada elemento es un descriptor d local C-dimensionales, muy parecido a la idea de los descriptores locales tradicionales como SURF / SIFT. Si R^C define el conjunto $H \times W$ de descriptores d para el mapa de características $c = \{1, \ldots, C\}$ y J representa el conjunto de los descriptores d_w C-dimensionales de la capa convolucional. Esto queda expresado por la siguiente ecuación:

$$J = \{d_w \in R^C | \forall w \in \{1, \ldots, H \times W\}, \forall c \in \{1, \ldots, C\}\}. \quad (1)$$

3.2. Minería de patrones distintivos de una red neuronal convolucional

Las regiones de interés para caracterizar una imagen son extraídas directamente de la última capa convolucional. Generalmente, las activaciones en esta capa indican que existen patrones útiles; son más escasas y corresponden a patrones visuales de la imagen que son semánticamente significativas. Por lo tanto, aún bajo cambios de vista extremos u oclusiones, estas se conservarán y serán muy parecidas en los mismos lugares en las imágenes.

Para cada mapa de características C, se asignan las activaciones a_i de los mapas de características a los grupos $G_j, \forall j \in \{1, \ldots, E\}$, donde E indica el número de grupos. Consideramos como grupo a todas aquellas activaciones a_i diferentes a cero que sea vecino con dos o más activaciones. Una vez identificados los grupos, se calcula la activación promedio de cada uno de ellos, como se muestra en la siguiente ecuación:

$$H_j = \frac{\sum a_j^i}{|S_j|}, \forall a_j^i \in G_j, \quad (2)$$

donde a_j^i denota la $i-$ésima activación en el grupo G_j y H_j representa la activación promedio del grupo G_j. Dado las activaciones promedio de esos grupos H_j, se ordenan y se seleccionan los Q grupos con las activaciones promedio más altas. Llamaremos a los grupos seleccionados como las regiones de interés R. La siguiente ecuación muestra este proceso:

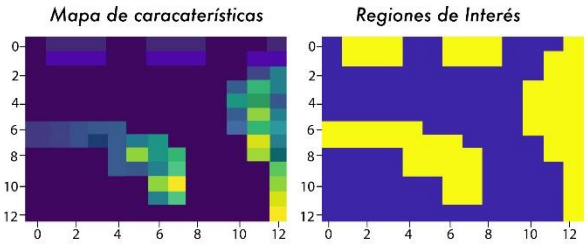

Fig. 4. Izquierda: una capa el mapa de características y sus activaciones, derecha: regiones de interés identificadas mediante el método propuesto

$$R = G_k, \forall k \in \{1, \ldots, Q\}. \quad (3)$$

Los descriptores dados por la ecuación 1 que caigan en estas regiones de interés R son agregados en la ecuación 4. Cada área de interés es representada por un descriptor f_k de tamaño $1 \times C$. La imagen es representada por el conjunto de descriptores F dado en las siguientes ecuaciones:

$$f_k = \sum_{q \in B_q} J_q, \quad (4)$$

$$F = \{f_k, \forall k \in \{1, \ldots, Q\}\}, \quad (5)$$

Un ejemplo del as regiones de interés generado con este procedimiento se muestra en la Figura 4.

3.3. Vocabulario basado en regiones uso de VLAD para el apareamiento de imágenes

La técnica de vectores de descriptores localmente agregados (del inglés VLAD) [2] es un método de agrupación de descriptores bastante popular para la recuperación de imágenes [2]. El método captura información sobre los descriptores locales, agregándolos al diccionario de palabras visuales. A diferencia del popular método de bolsa de palabras (BoW) [38], que agrega la información manteniendo el recuento de palabras visuales, VLAD almacena la suma de los residuos (vector

de diferencia entre el descriptor y su centro de agrupación correspondiente) para cada palabra visual. Para crear este diccionario de palabras visuales C, se empleó la totalidad de las imágenes que se usaron en el entrenamiento. Dichas imágenes incluyen entornos suburbanos en múltiples momentos del día. Las imágenes consisten en recorridos que cuentan con grandes variaciones en puntos de vista y oclusiones. Se utilizó el algoritmo $K-$medias para obtener los grupos de palabras visuales usadas en VLAD tomando en cuenta la siguiente formulación:

$$C = \{m_k, \forall k \in \{1, \dots, V\}, V \in \{64\}\}. \quad (6)$$

VLAD codifica F como se muestra en la siguiente ecuación, al utilizar el diccionario 6 de palabras visuales. Sea q_{ik} la fuerza de la asociación del descriptor f_k al grupo m_k, de modo que $q_{ik} \geq 0$ y $\sum_{k=1}^{K} q_{ik} = 1$. Esta asociación es obtenida por cuantificación vectorial mediante $K-$medias. m_k representa los grupos de agrupación (palabras visuales), los cuales son de la misma dimensión que f_k. VLAD codifica F como se muestra a continuación:

$$v_k = \sum_{k=1}^{M} q_{ik}(f_i - m_k)\,. \quad (7)$$

Estos residuos luego se apilan para obtener un descriptor de tamaño $Q \times K$. La implementación de VLAD admite varias estrategias de normalización. Las utilizadas en este trabajo se basan en la siguiente formulación, la cual se aplica a todos los componentes escalares del descriptor VLAD:

$$v_k = \text{sign}(v_k)\,\|v_k\|. \quad (8)$$

Esto es seguido por una normalización l_2, como sigue:

$$v_k = \frac{v_k}{\sqrt{{v_k}^T v_k}}\,. \quad (9)$$

Finalmente, la representación VLAD viene dada como:

$$P = \{v_k, \forall k \in \{1, \dots, V\}\}\,. \quad (10)$$

3.4. Similitud entre imágenes

Para determinar la similitud entre dos imágenes A y B, se realiza una comparación cruzada entre todos los elementos del vector VLAD P^A y P^B que se extrajeron de ambas imágenes. La similitud entre el vector i de A y el vector j de B se calcula como:

$$w_{i,j} = \frac{P_i^A P_j^B}{\|P_i^A\| \|P_j^B\|}. \quad (11)$$

La verificación cruzada se aplica este caso para aceptar solo coincidencias mutuas. Como resultado, la similitud general entre dos imágenes A y B se calcula la sumando las diferencias entre elementos de estas imágenes como se muestra a continuación:

$$W_{A,B} = \sum_{i,j} w_{i,j}\,. \quad (12)$$

Finalmente, para encontrar la mejor coincidencia entre la imagen de referencia de A respecto a la imagen de consulta B, se consultan todas las imágenes de referencia de la base de datos y se selecciona la que tenga la puntuación de similitud más alta, esto viene dado como sigue:

$$Y(B) = \arg\max_A W_{A,B}\,. \quad (13)$$

3.5. Función de pérdida por puntos de correspondencia

Usar solo las regiones de interés para la representación de una imagen tiene la propiedad de que la representación es más robusta a cambios de vista extremos u oclusiones, por lo que la idea es poder seleccionar estas regiones. Trabajos como [40, 31] utilizan detectores externos, los cuales no se especializan en esta tarea y se enfocan más en la detección de objetos, además de que se requieren pasos extras, ya que estas regiones se introducen en otros modelos RNC para obtener una representación convolucional de estas áreas. Por otro lado, en [10], los autores utilizan las activaciones de los mapas de características para detectar estas regiones.

Este enfoque tiene la ventaja que no requiere pasos extras y las regiones están más enfocadas al reconocimiento de lugares. Sin embargo, tiene el inconveniente que se tiene que seleccionar manualmente la capa convolucional para extraer estas regiones. En este trabajo se propone

aprender a generar estas regiones de interés automáticamente en la última capa junto a su representación, Esto requiere aprender las características que correspondan en dos imágenes, por lo que es necesario usar una función de perdida diseñada para esta tarea.

Basándonos en la idea propuesta por [11], para la identificación de puntos de interés, nuestro método CorVlad-Net se enfoca en regiones de interés. Para aprender estas correspondencias se define una función de perdida que toma N regiones locales de las imágenes I e I', dadas por las áreas a_i y a_i'. Sean $F_i(a_i)$ las características en la región de interés en la imagen I dada por el ´área cubierta por el rectángulo $a_i = (x_i, y_i, z_i + h, y_i + t)$, para pares correspondientes positivas de esta región con I', se espera que estas sean parecidas en el espacio de características, y la función de perdida reduzca la distancia entre estas, para correspondencias negativas con I', se espera aumentar la distancia lo máximo posible o por lo menos sea un margen mínimo m.

Se denotarán las correspondencias positivas entre estas regiones como $c = 1$ y $c = 0$ en caso contrario, estas correspondencias vienen dadas por los datos de entrenamiento previamente etiquetados. La función de pérdida propuesta viene dada como sigue:

$$L = \frac{1}{2N}\sum_{i}^{N}[c_i s + (1 - c_i)\max(0, m - s)], \qquad (14)$$

$$s = \left\|S(a_i, a_i')\right\|^2 = \left(F_I(a_i) - F_I(a_i')\right). \qquad (15)$$

La función de perdida consiste en dos términos. El primer término minimiza la distancia entre pares positivos y el segundo término empuja a los pares negativos a separarse al menos una distancia mínima entre ellos.

3.6. Regiones de interés robustas

Las RNCs pueden manejar cierto grado de variaciones de escala y rotación. Sin embargo, para lidiar con esto y hacer el modelo más robusto, tradicionalmente se ataca el problema explícitamente al aumentar los datos mediante diferentes transformaciones, como el descrito en [15] se utilizan estructuras especiales. En el problema tratado se busca la correspondencia espacial entre regiones de interés entre diferentes imágenes, por lo que es necesario poder comparar regiones de interés de diferentes imágenes sometidas a diferentes cambios de vista, escalado e iluminación y las soluciones clásicas no son suficientes. Esto se logra mediante la normalización de parches mediante el trasformador espacial.

Para lograr esta normalización de regiones de interés, se utilizó el modulo propuesto en [16], el cual se basa en transformadores espaciales, lo que permite que las regiones de interés en las imágenes pueden experimentar una transformación independiente en cada una. Esta característica es especialmente importante para las lograr una buena correspondencia entre imágenes con gran variación en puntos de vista y oclusiones; para cada característica, el transformador espacial aplica una transformación espacial independiente. El módulo de trasformador espacial se ilustra en la Figura 5.

4. Conjuntos de datos de entrenamiento

En esta sección, se describe el conjunto de datos que se utilizó para entrenar a CorVlad-Net. El conjunto de datos usado fue el de Google Landmark Boxes, el cual se basa en el conjunto de datos de monumentos históricos de Google (GLD) [30], y el cual se anotó a mano por ellos. El conjunto de datos contiene 1,2 millones de imágenes de 15,000 puntos de referencia únicos, con una amplia variedad de objetos, incluidos edificios, monumentos, puentes, estatuas y puntos de referencia naturales como montañas, lagos y cascadas. Cada imagen en este conjunto de datos solo representa un punto de referencia. En algunos casos, un punto de referencia puede consistir en un conjunto de edificios: por ejemplo, los horizontes, que son comunes en este conjunto de datos, se consideran como un punto de referencia único.

El objetivo del conjunto de datos es capturar el punto de referencia más destacado en la imagen, el cual se marca mediante rectángulos y a cada imagen solo se le asigna una etiqueta de punto de referencia. Cada rectángulo debe reflejar el objeto

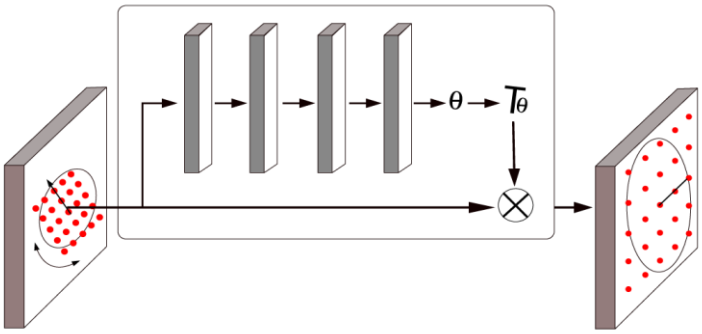

Fig. 5. La arquitectura del módulo del transformador espacial. El mapa de características toma una imagen de entrada que se pasa a una red de localización que regresa los parámetros de transformación θ, o que produce una nueva salida deformada

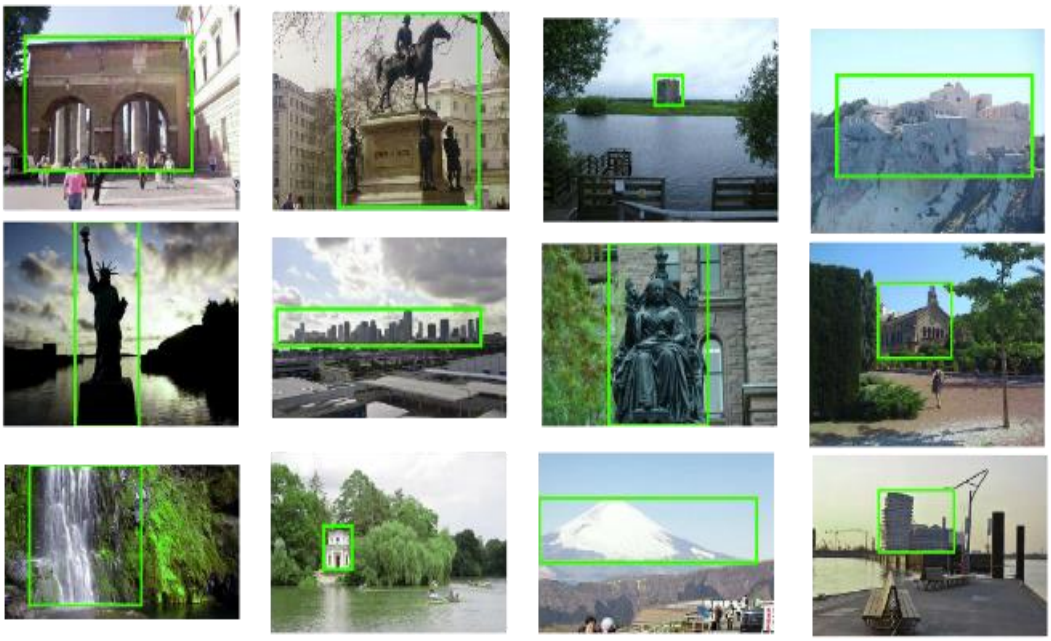

Fig. 6. Ejemplos de imágenes anotadas del conjunto de datos de Google Landmark Boxes. Se dibuja un cuadro alrededor del área de interés más prominente en la imagen. El conjunto de datos contiene una amplia variedad de objetos que van desde lugares creados por el hombre hasta lugares naturales

principal (o conjunto de objetos) que se muestra en cada imagen del conjunto de datos.

En GLD, algunos puntos de referencia están asociados a varios miles de imágenes, mientras que para aproximadamente la mitad de las clases solo se proporcionan 10 o menos imágenes. Por este motivo se separó parte del conjunto de entrenamiento de 1.2M en un conjunto de validación.

Se seleccionaron al azar cuatro imágenes de entrenamiento y cuatro de validación por punto de referencia. En total, esto produce 58k y 36k imágenes para entrenamiento y validación, respectivamente. En la Figura 6 se muestran

Tabla 1. Comparación entre el método propuesto en este trabajo (CorVlad-Net) y diferentes métodos considerados estado del arte. La etiqueta **Nombre corto** indica el nombre de la arquitectura o un nombre asignado por nosotros para referirnos al trabajo. **Entrenamiento** indica el tipo del conjunto de datos usado en el entrenamiento. **Características** indica el tipo de características usadas. **Tipo de características** indica si son características globales o locales. **Arquitectura** muestra el tipo arquitectura RNC usada. Finalmente, **Codificación** indica cómo se agregaron las características para crear la codificación de la imagen

Trabajo	Nombre corto	Entrenamiento
[12]	FABMAP	
[28]	SEQSLAM	
[8]	AMOSNet HybridNet	SPED [8]
[10]	OnlyLookOnce	IMAGNET
	CorVlad-Net	GLD [30]
Trabajo	**Nombre corto**	**Características**
[12]	FABMAP	SURF
[28]	SEQSLAM	
[8]	AMOSNet HybridNet	Convolucional
[10]	OnlyLookOnce	Convolucional
	CorVlad-Net	Convolucional
Trabajo	**Nombre corto**	**Tipo de características**
[12]	FABMAP	Local
[28]	SEQSLAM	Local
[8]	AMOSNet HybridNet	Global
[10]	OnlyLookOnce	Regiones locales, activaciones
	CorVlad-Net	Regiones locales, activaciones
Trabajo	**Nombre corto**	**Arquitectura**
[12]	FABMAP	
[28]	SEQSLAM	
[8]	AMOSNet HybridNet	VGG 16
[10]	OnlyLookOnce	VGG 16
	CorVlad-Net	Googlenet
Trabajo	**Nombre corto**	**Codificación**
[12]	FABMAP	Bolsa de palabras
[28]	SEQSLAM	
[8]	AMOSNet HybridNet	Pooling multiescala
[10]	OnlyLookOnce	Codificación novedosa, similar a la bolsa de palabras
	CorVlad-Net	VLAD

ejemplos de imágenes del conjunto de datos con sus regiones de marcadas. Estas áreas de correspondencia son las que se utilizan en la función de perdida 14 y 15.

5. Métodos comparados

En esta sección se describen las principales características y ventajas de los métodos con los cuales se compara el trabajo propuesto. En la Tabla 1 se muestra las diferencias entre los diferentes trabajos y CorVlad-Net. Se compran diferentes aspectos como el conjunto de datos utilizados para entrenar los modelos, el tipo de arquitectura de la RNC, el tipo de características, etcétera, Se utiliza el nombre corto en la descripción de cada trabajo.

El método aprende un modelo generativo de apariencia del lugar. Al dividir el problema de aprendizaje en dos partes, se pueden aprender nuevos modelos de lugares en línea a partir de una sola observación de un lugar. La complejidad del algoritmo es lineal en el número de lugares del mapa y es particularmente adecuado para la detección de cierre de bucle en línea.

Fue el primer trabajo que utilizó descriptores de características tipo SURF con codificación BoW. Demostró robustez bajo los cambios de punto de vista al aprovechar las propiedades de variación de SURF. En SEQSLAM en lugar de calcular la ubicación única más probable dada una imagen actual, el enfoque calcula la mejor ubicación de coincidencia candidata dentro de cada secuencia de navegación local. La localización se logra al reconocer secuencias coherentes de estas ”mejores coincidencias locales”.

Este enfoque elimina la necesidad de un rendimiento de coincidencia global por parte de la interfaz de visión; en su lugar, solo debe elegir la mejor coincidencia dentro de una secuencia corta de imágenes. El enfoque es aplicable a los cambios de entorno que hacen que las técnicas tradicionales basadas en características como FAB-MAP sean ineficaces. Ha mostrado un desempeño notable bajo severos cambios de apariencia. Sin embargo, es incapaz de lidiar con cambios simultáneos y la variación del punto de vista. Aprovechando la información secuencial,

SEQSLAM ha mostrado un rendimiento de vanguardia.

En el trabajo donde se propone HybridNet, los autores entrenaron dos arquitecturas para la tarea de reconocimiento de lugar específico y emplearon un método de codificación de características de múltiples escalas para generar características robustas a cambio de condiciones y puntos de vista. Para llevar a cabo el entrenamiento de las redes neuronales, los autores desarrollaron un conjunto de datos masivo de lugares (SPED). HybridNet tiene pesos inicializados por CaffeNet previamente entrenada.

Los resultados mostrados fueron sobresalientes en conjuntos de datos a gran escala y demostró en varios conjuntos de datos desafiantes que las representaciones internas aprendidas de las redes son más sólidas contra las variaciones de apariencia y puntos de vista que las extraídas de redes centradas en objetos.

El último trabajo *onlylookonce*, los autores dan un paso más en la estructura interna de las RNC y proponen características de imagen novedosas basadas en RNC para el reconocimiento de lugares identificando regiones sobresalientes y creando sus representaciones regionales directamente a partir de las activaciones de la capa convolucional.

Inspirado por el éxito de las representaciones de imágenes basadas en regiones para el reconocimiento de lugares y el reciente auge de las técnicas de aprendizaje profundo, el trabajo presenta un método de codificación de características novedoso para crear representaciones de imágenes utilizando las activaciones de capas convolucionales de CNN basadas en el método bolsa de palabras.

El método utiliza una capa convolucional para la extracción de características locales y otra capa convolucional tardía para identificar regiones destacadas, de las cuales se pueden extraer características locales. La presentación de la imagen derivada codifica varias regiones de imagen distintivas que se pueden usar para la comparación cruzada en una etapa de recuperación posterior.

Las comparaciones con técnicas de vanguardia en amplios conjuntos de datos de evaluación comparativa demuestran un rendimiento superior del método propuesto en tareas de reconocimiento de lugar con fuertes variaciones de puntos de vista y condiciones. La RNC usada está entrenada en un conjunto de datos de reconocimiento de objetos, y trabajos anteriores parece indicar que las redes entrenadas para el reconocimiento de lugares ofrecen resultados superiores.

A diferencia de los trabajos mencionados, el enfoque propuesto en este trabajo identifica directamente las regiones de interés en la última capa convolucional. En los trabajos ya mencionados, se debe escoger que capa usar para obtener las características que se usaran para la representación.

Además, las RNCs son solo robustas para algunos tipos de transformaciones como la traslación y en menor medida al escalado, por lo que cambios brucos en la imagen siguen siendo problemáticos para otros trabajos, por lo que, en este trabajo, para aumentar la robustez de las características se implementó un transformador espacial que permite a la RNC aprender a escalar, rotar o aplicar transformaciones arbitrarias a las regiones de interés.

También se propone una función de pérdida específica a la tarea de reconocimiento de lugares por lo que la RNC aprender a generar estas regiones de interés automáticamente, a diferencia de los otros trabajos donde se usan las mismas funciones que se utilizan en el reconocimiento de objetos. Además, la RNC propuesta también se entrena en un conjunto de datos específico a en la tarea de reconocimiento de lugares.

6. Experimentos

En esta sección se presentan los detalles de implementación del método propuesto. También se muestra el rendimiento en la tarea de recuperación de imágenes. La comparación del método propuesto se llevó a cabo contra otros trabajos de última generación considerados estados del arte, para hacer la comparación se utilizaron varios conjuntos de datos de referencia los cuales cuentan con diferentes características como cambios de vista y luz, para poder medir cómo se comporta cada método en cada uno de los conjuntos de datos. Finalmente, se exponen y comparan los resultados obtenidos.

Tabla 2. Conjuntos de datos de referencia para el reconocimiento de lugares

	Conjuntos de datos		
	Berlin A100	Berlin Haleenseetrasse	Garden Point
Ambiente	urbano	urbano, suburbano	urbano
Punto de vista	moderado	muy fuerte	fuerte
Condición	moderado	moderado	fuerte
Datos de prueba	166	224	400

6.1. Conjuntos de datos de prueba

Los conjuntos de datos usados para las pruebas y evaluación el sistema propuesto fueron los siguientes: Gardens Point fue capturado en el campus de QUT con una travesía tomada durante el día en la caminata lateral izquierda y la otra fue registrada en la caminata lateral derecha durante la noche [8]. Los conjuntos Berlin Halenseestrasse y Berli A100 fueron recopilados de la plataforma de geo-etiquetado de origen público llamada Mapillary.

Cada conjunto de datos cubre dos recorridos de la misma ruta cargados por diferentes usuarios. Un recorrido es utilizado como datos de referencia y el otro se usa como datos de pruebas. Los detalles de estos conjuntos de datos y sus diferentes características se muestran en la Tabla 2.

7. Detalles de implementación de los métodos comparados

El método propuesto se implementó usando la arquitectura Googlenet [11], entrenada con el conjunto de datos Google Landmark Boxes [30]. Se empleó este modelo RNC para la extracción de características basadas en la región con un tamaño de imagen de entrada de 256×256. Como se mencionó en la metodología, se utilizó la última capa convolucional para crear el descriptor, el cual se codificó mediante la representación VLAD. Se utilizaron las primeras 64 regiones de interés para la codificación y se entrenó un vocabulario de 3.1k palabras visuales.

Se compararon los resultados obtenidos usando el trasformador espacial y sin él. Se utilizaron también otros métodos de agregación de características como CrossPooling, MaxPooling, SumPool, sin agregación y se compararon los resultados.

Para mostrar el desempeño se hizo la comparación contra otros métodos considerados de vanguardia pero que utilizan enfoques diferentes al nuestros como: FABMAP y SEQSLAM, en donde se utilizó la implementación estándar de FABMAP [12] y para SEQSLAM [12] se utilizó la configuración de tres tramas secuenciales.

También se comparó contra HybridNet (SPP) [8] y el trabajo de [10], el cual es el enfoque más parecido al nuestros y obtiene resultados similares, al que llamaremos CrossBow.

7.1. Tiempos de procesamiento

El uso de la metodología propuesta de RNC es computacionalmente demandante por lo que, una evaluación de su rendimiento en tiempo de ejecución es especialmente importante.

Para llevar a cabo la medición se escogieron 1000 imágenes y se registró el tiempo promedio de ejecución. Para una sola imagen, una pasada hacia adelante a través de la red Googlenet tarda aproximadamente 214 ms en una GPU NVIDIA RTX 2080 TI y la codificación de la característica toma alrededor de 0,712 s.

En comparación con el trabajo [10] que utiliza una metodología similar, reportan que con una red VGG16 tarda aproximadamente 59,4 ms en una GPU NVIDIA Titan X Pascal y codificar las características aproximadamente 0,349s. Nuestro método es muchísimo más lento que el método propuesto en [10].

Esto se debe a la elección de usar Googlenet e incorporar el trasformador espacial en la arquitectura, una alternativa seria usar una arquitectura más rápida como VGG16, por lo que se tendría que estudiar cómo afecta en los resultados obtenidos.

Tabla 3. Resultados para Berlin A100

Métodos	AUC	Métodos	AUC
CorVlad-Net	**0.742**	CorVlad-Net-No-pooling	0.431
CorVlad-Net-Sin-Tec	**0.621**	**onlylookonce**	**0.753**
CorVlad-Net-Cross-Pooling	0.221	HybridNet	0.036
CorVlad-Net-Max-Pooling	0.124	SEQSLAM	0.151
CorVlad-Net-Sum-Pooling	0.209	FABMAP	0.101

8. Resultados

En las tareas de recuperación de imágenes donde existe un desequilibrio grande entre las clases, lo que significa que las muestras de clase positiva son bastante raras en comparación con las clases negativas, las curvas de Precisión /Recuperación generalmente se emplean como métrica de evaluación [14].

Es por esto que se reporta el área bajo la curva (AUC) para estas en los experimentos.

8.1. Berlin A100

Este conjunto de datos exhibe ambientes dinámicos, aparte de cambios de vista moderados. Los resultados obtenidos se muestran en la Tabla 3. Los resultados muestran que nuestro enfoque CorVlad-Net logra resultados similares a los reportados en [10].

Ya que los cambios no son tan grandes y no se requiere una codificación tan robusta, por lo que los métodos se desempeñan de manera similar. Aun así, los otros enfoques como HyridNet no obtienen resultados convincentes.

Puede deberse a que HybridNet se entrenó usando el conjunto de datos SPED [8], el cual contiene solo cambios mínimos entre los mismos lugares capturados en varias veces al año a diferencia del conjunto de datos que se está probando, el cual contiene cambios lo suficientemente significativos para ser un problema para este método.

Tabla 4. Resultados para Berlin Halenseestrass

Métodos	AUC	Métodos	AUC
CorVlad-Net	**0.812**	CorVlad-Net-No-pooling	0.223
CorVlad-Net-Sin-Tec	**0.503**	**onlylookonce**	**0.523**
CorVlad-Net-Cross-Pooling	0.423	HybridNet	0.073
CorVlad-Net-Max-Pooling	0.086	SEQSLAM	0.021
CorVlad-Net-Sum-Pooling	0.104	FABMAP	0.132

Tabla 5. Resultados para Garden Point

Métodos	AUC	Métodos	AUC
CorVlad-Net	**0.791**	CorVlad-Net-No-pooling	0.510
CorVlad-Net-Sin-Tec	**0.633**	**onlylookonce**	**0.682**
CorVlad-Net-Cross-Pooling	0.413	HybridNet	0.506
CorVlad-Net-Max-Pooling	0.053	SEQSLAM	0.721
CorVlad-Net-Sum-Pooling	0.095	FABMAP	0.034

Para las técnicas de agrupación Sum-Pool, Max-Pool y Cross-Pool, dado que las condiciones y las variaciones del punto de vista son más grandes, en este conjunto de datos, se obtienen resultados pobres. Además, este conjunto de datos cuenta con intervalos de tiempos muy variados entre las imágenes capturadas, por lo que SEQSLAM y FABMAP también obtienen un rendimiento muy inferior.

8.2. Berlin Halenseestrass

Los resultados obtenidos se muestran en la Tabla 4, los resultados de CorVlad-Net para el conjunto de datos superan significativamente a todos los demás métodos de vanguardia. Sorprendente-mente, los resultados para

onlylookonce tienen un rendimiento inferior con un gran margen en comparación de CorVlad-Net.

El conjunto de datos contiene cambios extremos en el punto de vista, que es el problema que CorVlad-Net se diseñó a resolver, por lo que los resultados son esperados y superiores a todos los demás trabajos. El uso del trasformador espacial es de gran ayuda ya que logra una representación más robusta en este conjunto de datos con cambios extremos de vista.

Como se mencionó, el segundo mejor resultado es Cross-Bow, que toma un enfoque parecido al de nosotros, pero falla en ofrecer una mayor robustez a los cambios extremos presente en este conjunto de datos. Por otro lado, los métodos clásicos FABMAP y SEQSLAM, debido a la naturaleza del conjunto de datos, se desempeñan pobremente.

Entre los mejores métodos de agregación usados esta CorVlad-Net-Cross-Pooling, el cual muestra resultados parecidos a Cross-Bow, ya que usa en método de codificación similar.

8.3. Garden Point

Los recorridos en este conjunto de datos muestran fuertes variaciones de luz y fuerte coherencia temporal entre las imágenes. Los resultados obtenidos se muestras en la Tabla 5 CorVlad-Net logra un rendimiento similar pero superior que Cross-BoW.

Por lo que se sigue mostrando una gran robustez en diferentes conjuntos de datos y condiciones. El método SEQSLAM reporta resultados de vanguardia debido a que los conjuntos de datos presentan una muy fuerte coherencia secuencial que es en lo que mejor se desempeña este método.

Sin embargo, los métodos basados en regiones de interés como CorVlad-Net y Cross-BoW, obtienen resultados similares, pero superiores en otras pruebas por lo que resultan más robustos en más situaciones. Los enfoques que incluyen Sum-Pool, Max-Pool y FABMAP tuvieron un rendimiento bastante bajo aun para este tipo de condiciones.

9. Conclusiones

Obtener un rendimiento y precisión de vanguardia en la tarea de reconocimiento de lugares de manera visual en presencia de grandes cambios de puntos de vista, iluminación y oclusiones es altamente deseable. Sin embargo, resulta un problema aun desafiante.

Inspirado por el auge y éxito de las técnicas de aprendizaje profundo en los últimos años, en diferentes tareas de visión por computadora y reconocimiento de patrones, este trabajo afronta esta problemática presentando una nueva metodología y arquitectura RNC, denominada CorVlad-Net. Esta arquitectura produjo resultados superiores para diferentes conjuntos de datos, superando a otros métodos del estado del arte.

Con los resultados obtenidos se pudo concluir que el método propuesto que utiliza regiones de interés de imagen obtiene resultados de vanguardia, sobre todo con conjunto de datos en presencia de puntos de vista extremos y oclusiones. Además, el trasformador espacial que se integró en la arquitectura del modelo RNC ayuda significativamente a la normalización de estas regiones de interés, lo cual logra una representación aún más robusta de la codificación de la imagen.

Otra ventaja fue el uso de la función de perdida de correspondencia propuesta, la cual permite entrenar el modelo de manera completa, de principio a fin. No se necesitan pasos intermedios; directamente se aprende a identificar las regiones de interés que mejor caracterizan la imagen en la última capa convolucional, a diferencia de otros métodos donde se requieren pasos adicionales. Por último, la función de perdida propuesta permite entrenar con el nuevo conjunto de datos Google Landmark Boxes, el cual fue creado específicamente para la tarea de reconocimiento de lugares, lo cual ayuda a obtener aún mejores resultados que otros métodos entrenados con conjuntos de datos de objetos. La combinación de todos estos métodos permitió alcanzar resultados de vanguardia en conjuntos, en comparación a otros métodos considerados estado del arte.

Para trabajos futuros se propone analizar el rendimiento del marco conceptual propuesto con otros conjuntos de datos más desafiantes.

Además, como el método sigue dependiendo de la selección de ciertos parámetros, como el número de regiones interés que se utilizan en la codificación, se propone explorar la posibilidad de seleccionar de estas regiones de interés de características en forma automática mediante otras técnicas de búsqueda como algoritmos genéticos, técnicas de enjambre, etcétera.

Agradecimientos

E. Zamora y H. Sossa desean agradecer el apoyo proporcionado por CIC-IPN para llevar a cabo esta investigación. Este trabajo fue apoyado económicamente por los proyectos SIP-IPN: 20200651, 20180180, 20190166, 20190007 y 20200886, así como los proyectos CONACYT 65 (Fronteras de la Ciencia) y 6005 (FORDECYT-CONACYT). Omar E. Lugo Sánchez agradece a CONACYT por la beca otorgada para continuar sus estudios de doctorado.

Referencias

1. **Krizhevsky, A., Sutskever, I., & Hinton, G. (2012).** Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, pp. 1097–1105.
2. **Arandjelovic, R. & Zisserman, A. (2013).** All about VLAD. *IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society*, pp. 1578–1585.
3. **Arandjelovic, R. & Zisserman, A. (2014).** Dislocation: Scalable descriptor distinctiveness for location recognition. *12th Asian Conference on Computer Vision*, pp. 188–204.
4. **Arroyo, R., Fernández-Alcantarilla, P., Bergasa, L., & Romera, E. (2016).** Fusion and binarization of CNN features for robust topological localization across seasons. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4656–4663. DOI: 10.1109/IROS.2016.7759685.
5. **Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008).** Surf: Speeded up robust features. *Computer Vision and Image Understanding,* Vol. 110, pp. 346–359.
6. **Cao, S. & Snavely, N. (2013).** Graph-based discriminative learning for location recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 700–707. DOI:10.1109/CVPR. 2013.96.
7. **Chen, D.M., Baatz, G., Köser, K., Tsai, S.S., Vedantham, R., Pylvänäinen, T., Roimela, K., Chen, X., Bach, J., Pollefeys, M., Girod, B., & Grzeszczuk, R. (2011).** City-scale landmark identification on mobile devices. CVPR. IEEE Computer Society, pp. 737–744. DOI:10.1109/ CVPR.2011.5995610.
8. **Chen, Z., Jacobson, A., Sunderhauf, N., Upcroft, B., Liu, L., Shen, C., Reid, I., & Milford, M. (2017).** *Deep learning features at scale for visual place recognition*. pp. 3223–3230. arXiv:1701.05105 [cs.CV]
9. **Chen, Z., Lam, O., Jacobson, A., & Milford, M. (2014).** *Convolutional neural network-based place recognition*. arXiv.org>cs>arXiv:1411.1509.
10. **Chen, Z., Maffra, F., Sa, I., & Chli, M. (2017).** *Only look once, mining distinctive landmarks from convnet for visual place recognition*. pp. 9–16. DOI: 10.1109/IROS.2017.8202131
11. **Choy, C., Gwak, J., Savarese, S., & Chandraker, M. (2016).** *Universal correspondence network*. arXiv.org>cs>arXiv:1606.03558.
12. **Cummins, M. & Newman, P. (2008).** Fab-map: Probabilistic localization and mapping in the space of appearance. *I. J. Robotic Res.,* Vol. 27, pp. 647– 665.
13. **Gronát, P., Obozinski, G., Sivic, J., & Pajdla, T. (2013).** Learning and calibrating per-location classifiers for visual place recognition. *CVPR, IEEE Computer Society*, pp. 907–914. DOI: 10.1109/CVPR.2013.122.
14. **Hanley, J. & Mcneil, B. (1982).** The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology,* Vol. 143, pp. 29–36. DOI: 10.1148/radiology.143.1.7063747.
15. **He, K., Zhang, X., Ren, S., & Sun, J. (2015).** *Spatial pyramid pooling in deep convolutional networks for visual recognition.* arXiv.org>cs> arXiv:1406.4229.
16. **Jaderberg, M., Simonyan, K., Zisserman, A., & Kavukcuoglu, K. (2015).** *Spatial transformer networks*. arXiv.org>cs>arXiv:1506.02025.
17. **Jegou, H., Douze, M., & Schmid, C. (2010).** Product quantization for nearest neighbor search. *IEEE Transactions on PAMI*, Vol. 33, No. 1, pp. 117–128.
18. **Jegou, H., Douze, M., Schmid, C., & Perez, P. (2010).** Aggregating local descriptors into a compact image representation. *IEEE Computer Society on Computer Vision and Pattern Recognition*, pp. 3304–3311. DOI: 10.1109/CVPR. 2010.5540039.
19. **Jegou, H., Perronnin, F., Douze, M., Sanchez, J., Perez, P., & Schmid, C. (2011).** Aggregating local image descriptors into compact codes. *IEEE*

Transactions on PAMI, Vol. 34, No. 9, pp. 1704–1716. DOI :10.1109/TPAMI.2011.235

20. **Knopp, J., Sivic, J., & Pajdla, T. (2010).** Avoiding confusing features in place recognition. *Computer Vision ECCV´10*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 748–761.
21. **Liu, L., Shen, C., & Hengel, A. (2015).** *The treasure beneath convolutional layers: Cross-convolutional layer pooling for image classification*. arXiv.org>cs>arXiv:1411.7466.
22. **Liu, L., Shen, C., & Hengel, A. (2016).** Cross-convolutional layer pooling for image recognition. *IEEE Transactions on PAMI*, Vol. 39, No. 11, pp. 2305–2313.
23. **Lowe, D. (2004).** Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, Vol. 60, pp. 91–110.
24. **Lowry, S., Sünderhauf, N., Newman, P., Leonard, J., Cox, D., Corke, P., & Milford, M. (2015).** Visual place recognition: A survey. IEEE Transactions on Robotics, Vol. 32, No. 1, pp. 1–19. DOI: 10.1109/TRO.2015.2496823.
25. **Mcmanus, C., Churchill, W., Maddern, W., Stewart, A., & Newman, P. (2014).** Shady dealings: Robust, long-term visual localisation using illumination invariance. *International Conference on Robotics and Automation*, pp. 901–906. DOI: 10.1109/ICRA.2014.6906961.
26. **Mcmanus, C., Upcroft, B., & Newmann, P. (2014).** Scene signatures: Localised and point-less features for localisation. *Robotics: Science and Systems*, pp. 1–9.
27. **Middelberg, S., Sattler, T., Untzelmann, O., & Kobbelt, L. (2014).** Scalable 6-dof localization on mobile devices. *ECCV ´14*. LNCS 8690, pp. 268– 283.
28. **Milford, M. & Wyeth, G. (2012).** Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. *International Conference on Robotics and Automation*, pp. 1643–1649. DOI: DOI: 10.1109/ICRA.2012.6224623.
29. **Ng, J., Yang, F., & Davis, L. (2015).** Exploiting local features from deep networks for image retrieval. arXiv.org>cs>arXiv:1504.05133.
30. **Noh, H., Araujo, A., Sim, J., Weyand, T., & Han, B. (2017).** *Large-scale image retrieval with attentive deep local features*. arXiv.org>cs>arXiv:1612.06321.
31. **Panphattarasap, P. & Calway, A. (2017).** Visual place recognition using landmark distribution descriptors. *ACCV´16.* LNCS 10114, pp. 487–502. arXiv:1608.04274 [cs.CV]
32. **Perronnin, F., Liu, Y., Sánchez, J., & Poirier, H. (2010).** *Large-scale image retrieval with compressed fisher vectors*. **CVPR Perronin, F., Sánchez, J., Liu, Y. (2010).** Large-Scale, pp. 3384–3391.
33. **Philbin, J., Chum, O., Isard, M., Sivic, J., & Zisserman, A. (2007).** Object retrieval with large vocabularies and fast spatial matching. *IEEE CVPR, IEEE Computer Society*, pp. 1–8. DOI: 10.1109/CVPR.2007.383172.
34. **Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011).** Orb: an efficient alternative to sift or surf. *ICCV´11*, pp. 2564–2571. DOI:10.1109/ICCV.2011.6126544
35. **Schindler, G., Brown, M., & Szeliski, R. (2007).** City-scale location recognition. *IEEE CVPR*, pp. 1- 7.
36. **Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & Lecun, Y. (2013).** *Overfeat: Integrated recognition, localization and detection using convolutional networks.* arXiv.org>cs>arXiv:1312.6229
37. **Simonyan, K. & Zisserman, A. (2014).** *Very deep convolutional networks for large-scale image recognition.* arXiv.org>cs>arXiv:1409.1556.
38. **Sivic, J. & Zisserman, A. (2003).** Video google: A text retrieval approach to object matching in videos. *ICCV, IEEE Computer Society*, pp. 1470–1477. DOI: 10.1109/ICCV.2003.1238663.
39. **Sünderhauf, N., Shirazi, S., Jacobson, A., Pepperell, E., Dayoub, F., Upcroft, B., & Milford, M. (2015).** Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free. *Robotics: Science and Systems*, pp. 1-10. DOI: 10.15607/RSS.2015.XI.022
40. **Sunderhauf, N., Shirazi, S., Dayoub, F., Upcroft, B., & Milford, M. (2015).** *On the performance of convnet features for place recognition.* arXiv.org>cs>arXiv:1501.04158.
41. **Tanaka, K. (2016).** *Self-localization from images with small overlap.* arXiv.org>cs>arXiv:1603.00993.
42. **Tolias, G., Sicre, R., & Jégou, H. (2015).** *Particular object retrieval with integral max-pooling of CNN activations.* arXiv.org>cs>arXiv:1511.05879.
43. **Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., & Pajdla, T. (2017).** 24/7 place recognition by view synthesis. *IEEE Transactions on PAMI*, Vol. 40, No. 2, pp. 257–271.
44. **Torii, A., Sivic, J., Pajdla, T., & Okutomi, M. (2015).** Visual place recognition with repetitive structures. *IEEE Transactions on PAMI*, Vol. 37, No. 11, pp. 2346–2359.
45. **Tuytelaars, T. & Mikolajczyk, K. (2008).** K.: Local invariant feature detectors: A survey. FnT Comp. *Graphics and Vision*, pp. 177–280.

46. Yandex, A. & Lempitsky, V. (2015). Aggregating local deep features for image retrieval. *IEEE ICCV*, pp. 1269–1277.

47. **Zheng, L., Yang, Y., & Tian, Q. (2016).** Sift meets CNN: A decade survey of instance retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40, No. 5, pp. 1224–1244.

Article received on 21/01/2020; accepted on 18/09/2020.
Corresponding author is Humberto Sossa.