Información Tecnológica
Vol. 30(6), 243-254 (2019)
http://dx.doi.org/10.4067/S0718-07642019000600243

Redes Convolucionales Siamesas y Tripletas para la Recuperación de Imágenes Similares en Contenido

Atoany N. Fierro[(1)], Mariko Nakano[(1)]*, Keiji Yanai[(2)] y Héctor M. Pérez[(1)]
(1) Sección de Estudio de Posgrado e Investigación, ESIME UC, Instituto Politécnico Nacional, Ciudad de México, México.
(e-mail: afierror@hotmail.com, mnakano@ipn.mx, hmperezm@ipn.mx)
(2) Escuela de posgrado de Informática e Ingeniería, The University of Electro-Communications, Tokio, Japón.
(e-mail: yanai@cs.uec.ac.jp)

* Autor a quien debe ser dirigida la correspondencia

Recibido Feb. 7, 2019; Aceptado Abr. 10, 2019; Versión final Jun. 5, 2019, Publicado Dic. 2019

Resumen

El objetivo del trabajo presentado fue el desarrollo de un sistema de recuperación de imágenes con base en su contenido, utilizando redes convolucionales siamesas y tripletas. Se utilizaron estas arquitecturas múltiples para generar descriptores visuales, extrayendo información semántica de dos imágenes (siamesa) o tres imágenes (tripleta) a la vez. Posteriormente, se realizó un aprendizaje de similitud, codificando la distancia de estas siamesas o tripletas de descriptores visuales, cuyo almacenamiento no es necesario. Los resultados muestran que los esquemas con base en redes convolucionales extraen mayor cantidad de información semántica. Las arquitecturas múltiples, aparte de extraer información semántica, mejoran la tasa de recuperación de imágenes. Se concluye que las arquitecturas múltiples solucionan los tres retos más importantes de estos sistemas, como lo son la brecha semántica, el aprendizaje de similitud y el espacio de almacenamiento, los cuales no habían sido resueltos en trabajos anteriores.

Palabras clave: redes siamesas; redes tripletas; brecha semántica; redes convolucionales; recuperación de imágenes

Siamese and Triplet Convolutional Neural Networks for the Retrieval of Images with Similar Content

Abstract

The objective of this paper was the development of a content-based image retrieval system, using siamese and triplet convolutional neural networks. These networks were used to generate visual descriptors, extracting semantic information from two images (siamese) or three images (triplet) at the same time. Then, a similarity learning was done, encoding these two or three visual descriptors. In the proposed scheme the storage of descriptors is not required. The experimental results show that the schemes based on convolutional neural networks extract more semantic information. The siamese and triplet architectures, apart from extracting semantic information, improved the image retrieval rate. It is concluded that the proposed scheme solved three of the main challenges in these systems, such as, semantic gap, similarity learning and storage space, which have not been solved in the previous works.

Keywords: siamese network; triplet network; semantic gap; convolutional neural networks; image retrieval

INTRODUCCIÓN

La recuperación de imágenes similares con base en su contenido (CBIR, por sus siglas en inglés) surgió a partir de la necesidad de clasificar, indexar y recuperar imágenes de interés de una gran base de datos sin la necesidad de utilizar metadatos, los cuales han demostrado ser ambiguos y subjetivos (Wang et al. 2011). Los sistemas CBIR dependen crucialmente de dos elementos: i) descriptores visuales y ii) medidas de similitud. Los descriptores visuales son vectores característicos que representan información visual de las imágenes, tales como el color, la forma, la textura o vectores característicos que representan contenido semántico de las imágenes, mientras que la medida de similitud trata de encontrar qué tan similares son las imágenes con base en las distancias entre sus descriptores visuales. Todos los sistemas CBIR procuran reducir el tamaño de sus descriptores para que el espacio de almacenamiento de estos sea mínimo sin sacrificar el funcionamiento de la recuperación.

Desde la última década, se han propuesto una gran cantidad de descriptores visuales (Wang et al. 2011, Tian et al., 2014, Montazar et al., 2015, Calderón et al., 2016, Meng et al., 2017, Pedronette et al., 2017, Song et al., 2018, Meng et al., 2018, Ren et al. 2014, Zhu et al. 2014, Jegou et al. 2012 y Zhang et al. 2015), por citar algunos. Estos descriptores se pueden clasificar en dos categorías dependiendo del tipo de características: globales o locales. Los descriptores globales son vectores característicos con base en colores, formas y texturas que se extraen de la imagen completa, mientras que los descriptores locales extraen características de puntos relevantes que se encuentran en cierta región de la imagen. Wang et al. (2011) combinan linealmente descriptores con base en el color, la forma y la textura para recuperar imágenes relevantes con respecto a una imagen de consulta dada, mientras que Calderon et al. (2016) y Song et al. (2018) desarrollaron descriptores con base en textura para imágenes específicas, tales como imágenes médicas y satelitales. Montazar y Giveki (2015) propusieron un descriptor local con base en SIFT (*Scale Invariant Feature Transform*). El sistema CBIR debe poder desempeñarse bien utilizando una base de datos con un gran número de imágenes sin causar problemas de almacenamiento ni tiempo de ejecución. Las técnicas fundamentadas en *bag of visual features* (BOV), incluyendo *bag of visual words* (BOW) (Ren et al, 2014 y Zhu et al., 2014), VLAD (Jegou et al., 2012 y Zhang et al., 2015) y vectores Fisher (Zhang et al., 2015) se han propuesto para este objetivo, logrando escalabilidad en el sistema.

Los esquemas antes mencionados describen a la imagen utilizando información como colores, texturas, bordes, formas, etc.; es decir, no realizan una descripción semántica del contenido visual de la imagen, es por ello, que se les conoce como descriptores visuales de bajo nivel. En otras palabras, estos descriptores visuales de bajo nivel se ven limitados en la reducción de la brecha semántica que existe entre la información de bajo nivel de la imagen, obtenida por la computadora y la información de alto nivel de los conceptos semánticos que los humanos percibimos. Recientemente, las redes neuronales convolucionales (CNN) junto con el aprendizaje profundo han ofrecido varias soluciones en el campo de la visión por computadora, ya que las redes CNN modelan abstracciones de alto nivel. Las redes CNN emplean arquitecturas compuestas por múltiples transformaciones no lineales (Bangio et al., 2013) lo que le permite a un sistema asociar directamente los valores numéricos de los pixeles de una imagen al concepto semántico que los humanos percibimos de ella, sin la intervención de labores humanas. Las CNN están organizadas jerárquicamente, donde las capas inferiores (cerca de la capa de entrada) extraen información de bajo nivel, mientras que las superiores (cerca de la capa de salida) extraen información semántica de alto nivel. En los sistemas CBIR, las CNN se consideran como posible solución para reducir la brecha semántica y obtener mejores resultados, equivalentes a la capacidad humana.

Debido a que las redes CNN son métodos para la clasificación de imágenes, investigadores han publicado algunas técnicas para el empleo de estas redes neuronales en los sistemas CBIR (Babenko et al., 2014, Bai et al., 2018 y Tzelepi et al. 2018, Chandrasekhar et al. 2017). La manera de utilizar las redes CNN en los sistemas CBIR es utilizar la salida de una de las últimas capas completamente conectadas (FC6, FC7 o FC8 en el caso de Alexnet) de la red como un vector característico para describir cada una de las imágenes en la base de datos. Estos vectores son almacenados y cuando se requiere hacer una consulta, la red CNN extrae el vector característico de la imagen de consulta y éste es comparado con cada uno de los vectores en la base de datos; los vectores que tengan la distancia más corta corresponderán a las imágenes de la base de datos que serán similares a la imagen de consulta. Los sistemas CBIR con base en redes CNN reducen la brecha semántica, ofreciendo mejores resultados comparado con los sistemas con base en descriptores de bajo nivel. Sin embargo, en general, la longitud de los descriptores de los sistemas con base en CNN es considerablemente grande, por lo que se necesita una gran cantidad de espacio de almacenamiento. Chandrasekhar et al. (2017) propusieron varios métodos de compresión de los descriptores obtenidos de las redes CNN, sin embargo, en general, los resultados muestran que el funcionamiento es directamente proporcional al tamaño del descriptor, ya que la compresión del descriptor sacrifica inevitablemente el funcionamiento del sistema CBIR.

Además, los sistemas CBIR con base en CNN (Babenko et al., 2014, Bai et al., 2018 y Tzelepi et al. 2018, Chandrasekhar et al. 2017) aún presentan problemas serios sobre el tamaño del descriptor y, por ende, el problema de espacio de almacenamiento de estos, ya que en general los descriptores de cada una de las imágenes contienen más de 1K datos reales (4096 datos reales en el uso de FC6 o FC7 de Alexnet). Aunque existen grandes esfuerzos por adaptar las configuraciones de redes CNN para la tarea de CBIR (Bai et al., 2018 y Tzelepi et al. 2018), el concepto de similitud y disimilitud de imágenes no está reflejado directamente en el algoritmo de aprendizaje de redes. Recientemente las arquitecturas de redes neuronales siamesa y tripleta han sido desarrolladas para aprender el concepto de similitud y disimilitud entre imágenes de entrada, y se han utilizado eficientemente para el reconocimiento de rostros y mapeos de regiones similares de imágenes (Melekhov et al. 2016). En este artículo, proponemos el uso de una red siamesa y una red tripleta para realizar tarea de CBIR, en donde se resuelve el problema del almacenamiento de los descriptores, ya que en ambas redes CNN no se requiere extraer los descriptores de cada una de las imágenes. Finalmente, los objetivos de este sistema desarrollado son los siguientes: 1) Reducir la brecha semántica que existe entre el valor numérico de los pixeles y el concepto semántico de la imagen; 2) Aumentar el potencial discriminativo realizando un aprendizaje de similitud y disimilitud de las imágenes en la base de datos; 3) Reducir el espacio de almacenamiento. Los resultados experimentales fueron obtenidos utilizando diferentes bases de datos: i) hemos utilizado una base de datos de referencia como lo es *Inria Holidays*, y j) una base de datos para el reconocimiento de objetos como *Caltech101.* Nuestros resultados muestran que, utilizando un aprendizaje de similitud entre las imágenes, la tasa de recuperación de imágenes similares se incrementa, además, que reducimos el espacio de almacenamiento, ya que no necesitamos almacenar cada uno de los descriptores visuales, resolviendo así los tres retos importantes en los sistemas CBIR, haciendo de este trabajo una gran contribución en este campo de investigación.

El resto del artículo está organizado de la siguiente manera: En la sección *Aprendizaje profundo aplicado a sistemas CBIR* explicaremos de manera detallada el uso de las diferentes variantes de redes CNN en los sistemas CBIR. En la sección *Método Propuesto*, se explican las arquitecturas de redes siamesa y tripleta, junto con el algoritmo de aprendizaje. En la sección de *Experimentos* y *Bases de datos* describiremos las bases de datos utilizadas en este trabajo y explicaremos de manera detallada los experimentos realizados y las métricas de evaluación. En la Sección *Resultados Experimentales* presentaremos los resultados obtenidos por cada una de las bases de datos, así como también las comparaciones entre esquemas. Finalmente, en la Sección *Conclusiones* concluimos este trabajo de investigación.

APRENDIZAJE PROFUNDO APLICADO A SISTEMAS CBIR

Los sistemas CBIR permiten la recuperación de imágenes similares a una imagen de consulta dada. Eso lo realizan creando representaciones numéricas de las imágenes, las cuales describen información como el color, la forma y la textura. Los sistemas CBIR trabajan en dos etapas, la primera es la generación de los vectores característicos llamados descriptores, donde cada una de las imágenes de la base de datos es representada por un vector numérico, dichos vectores son guardados en una base de datos. En la segunda etapa, una imagen de consulta es dada, posteriormente, se representa con un vector característico, el cual, es comparado utilizando una medida de distancia con cada uno de los vectores de la base de datos; los vectores que tengan una distancia corta corresponderán a imágenes similares a la imagen de consulta dada. En la figura 1 se puede observar el diagrama general de un sistema CBIR.

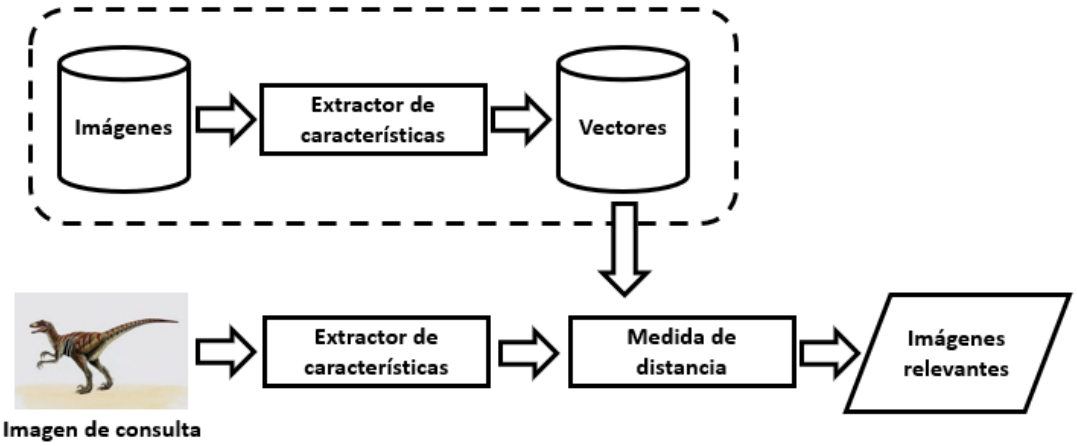

Fig. 1: Diagrama de un sistema CBIR

La extracción de características se realiza utilizando métodos “manuales”, representando la información de color, forma y textura de una manera numérica, sin embargo, existen métodos en donde se utilizan las redes

CNN para la extracción de características. Las redes CNN fueron propuestas por Y. Lecun et al. (1998) pero fue hasta el 2012 que estas redes se hicieron muy populares dentro de la comunidad de visión por computadora gracias al trabajo de Krizhevsky et al. (2012) quienes propusieron la red Alexnet. Inicialmente las redes CNN fueron propuestas para la clasificación de imágenes obteniendo resultados prometedores. A pesar de que las redes CNN han mostrado excelentes resultados en la clasificación de imágenes, aún no se ha hecho un análisis en cómo se pueden desempeñar en los sistemas CBIR.

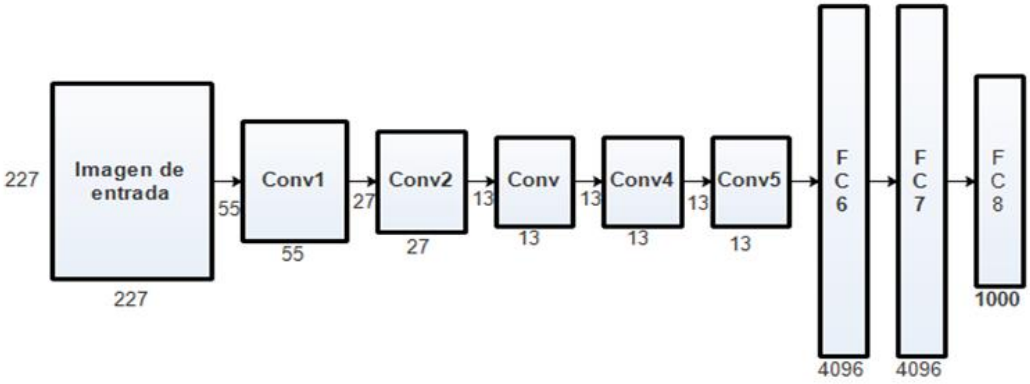

Fig. 2: Arquitectura Alexnet

La arquitectura de Alexnet consta de ocho capas, en donde las primeras cinco capas son de convolución, mientras que las últimas tres son capas totalmente conectadas. En la figura 2 se puede observar la arquitectura Alexnet, así como también las dimensiones de cada una de las capas. Las propuestas de aplicación de las redes CNN en los sistemas CBIR es utilizar las activaciones de las últimas tres capas totalmente conectadas (FC6, FC7, y FC8) de la red Alexnet como descriptores visuales. En este esquema, en (Babenko et al. 2014) se utiliza un modelo pre-entrenado con una base de datos, en este caso, ImageNet (Russakovsky et al., 2015), posteriormente, el modelo es afinado utilizando bases de datos muy similares a ImageNet, para que el tiempo de entrenamiento sea mucho menor, ya que solamente se reentrenan las capas superiores. En caso de que la base de datos objetivo sea muy diferente a la base de datos utilizada en el preentrenamiento, es necesario entrenar capas superiores, medias e inferiores, dependiendo de qué tan diferentes sean las bases de datos. Para obtener las representaciones de las imágenes, directamente introducimos las imágenes en la red y tomamos los valores de la activación de alguna de las últimas tres capas totalmente conectadas como vector característico.

En la figura 3 podemos observar la extracción del vector característico (ya sea FC6, FC7 o FC8) de algún modelo entrenado. Para el afinamiento de una red CNN, el reentrenamiento se realiza utilizando la base de datos objetivo, inicializando todas las capas del nuevo modelo con los parámetros del modelo pre-entrenado usando la base de datos ImageNet, excepto la última capa, la cual se inicializa aleatoriamente y se adapta al número de clases de la nueva base de datos a utilizar. Si la base de datos con la cual se va a trabajar está compuesta por 500 clases u objetos, la última capa de la red CNN deberá tener 500 neuronas de salida, las cuales formaran el vector característico FC8.

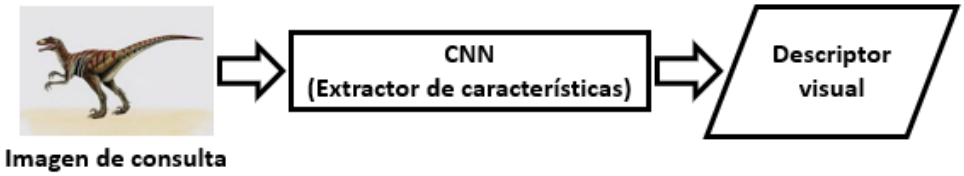

Fig. 3: Extracción de las capas totalmente conectadas (FC) como vectores característicos usando una red pre-entrenada

Para el entrenamiento, este tipo de redes utiliza la función de costo llamada Entropía Cruzada la cual está definida como sigue:

$$C = -\sum_x [y \ln \hat{y} + (1-y) \ln(1-\hat{y})]/n \quad (1)$$

Donde *n* es el número total de imágenes en el conjunto de entrenamiento, la sumatoria es sobre todas las imágenes de entrada *x*, y, (y, ŷ) son la salida deseada y la salida real de cada una de las imágenes *x*, respectivamente.

MÉTODO PROPUESTO

Los esquemas CBIR tradicionales realizan tres procesos importantes, extracción de características, cálculo de distancias y recuperación de imágenes similares a una imagen de consulta dada. Como ya se mencionó anteriormente, estos sistemas presentan tres retos importantes, los cuales son: la brecha semántica, cálculo de similitud y disimilitud, y, el espacio de memoria requerido para almacenar los descriptores visuales. En este trabajo proponemos el uso de arquitecturas CNN siamesas y tripletas para la recuperación de imágenes similares. Para la arquitectura siamesa, hemos utilizado dos redes Alexnet idénticas, las cuales comparten sus pesos sinápticos; la función de costo utilizada en este esquema es la función de costo contrastiva. La arquitectura tripleta es muy similar a la siamesa, hemos utilizado tres redes Alexnet, que al igual que la arquitectura siamesa, comparten pesos sinápticos, sin embargo, la función de costo utilizada se llama función de costo triple.

A diferencia del proceso manual de descripción de contenido visual utilizando características de bajo nivel como color, forma y textura, estas redes CNN nos permiten tener una mejor descripción utilizando información semántica de las imágenes, de esta manera, el método propuesto reduce la brecha semántica. Dentro de la función de costo de estas arquitecturas neuronales se encuentra embebido el cálculo de distancia, por lo que el esquema es capaz de aprender similitudes y disimilitudes entre las diferentes clases de imágenes de la base de datos, lo que le da un poder más discriminativo, además, no es necesario el almacenamiento de los vectores característicos, ya que mediante el cálculo de la activación de las neuronas (propagación hacía adelante) se puede determinar qué imágenes son más similares a la imagen de consulta dada. En la figura 4 y en la figura 5 se pueden observar las arquitecturas siamesa y tripleta de redes CNN, respectivamente.

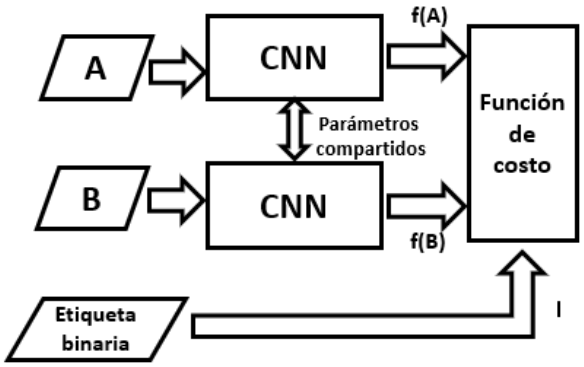

Fig. 4: Arquitectura de una red siamesa. Un par de imágenes como entradas alimentan a cada una de las redes, donde la variable etiqueta nos dice si el par de imágenes es positivo o negativo

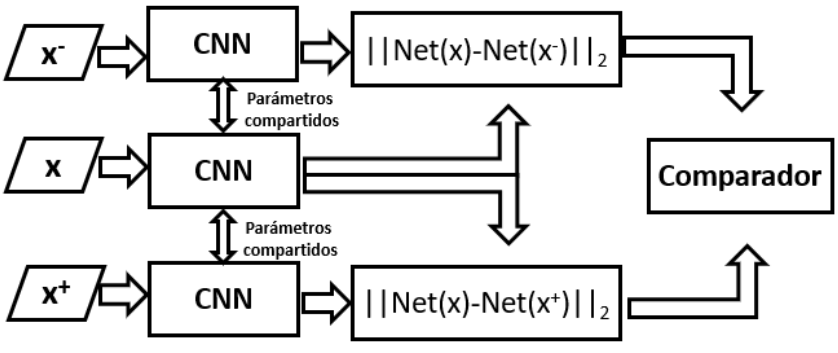

Fig. 5: Estructura de una red tripleta. La entrada a la red son tres imágenes, dos de ellas pertenecen a la misma clase (x y x^+), mientras que la otra pertenece a una clase diferente

La red neuronal siamesa mostrada en la figura 4 consta de un par de imágenes a la entrada de un sistema, el cual está constituido por dos redes idénticas las cuales comparten sus pesos sinápticos. Las salidas de cada una de estas redes son introducidas a una función de costo, donde esta función de costo trata de minimizar alguna métrica de distancia, (L_1, L_2 o coseno) entre las características de un par de imágenes positivo (f(A)) y (f(B)) y maximizarla cuando el par de imágenes sea negativo. A diferencia de las redes CNN convencionales, las redes siamesas utilizan la función de costo *contrastiva* la cual está definida como:

$$L = \{(l/D^2)/2\} + \{(1-l)/2\}\{\max(0, m-D)\}^2 \quad (2)$$

Donde D presenta la distancia obtenida entre las salidas de dos CNN, f(A) y f(B), *l* es una etiqueta binaria la cual indica si los pares de imágenes pertenecen a la misma clase (*l*=1) o si el par de imágenes son de clases diferentes (*l*=0) y m es un margen que indica qué tan similares deben ser las imágenes de una misma clase

Por otro lado, las redes tripletas (figura 5) están compuestas por tres redes idénticas las cuales también comparten sus parámetros. Estas redes entregan dos valores los cuales representan la distancia (L_1, L_2, etc.) entre la representación embebida de dos entradas con una tercera (Hoffer et al., 2015). Supongamos que las tres entradas de la red son las imágenes x, x^+ y x^-, y la representación embebida de la red es *Net(x)*, entonces, los dos valores entregados por la red tripleta utilizando la distancia L_2 son:

$$\text{TripletNet}(x, x^+, x^-) = [||\text{Net}(x)-\text{Net}(x^+)||_2, ||\text{Net}(x)-\text{Net}(x^-)||_2], \in R^2 \quad (3)$$

Donde x y x^+ son imágenes que pertenecen a la misma clase, mientras que x^- pertenece a una clase diferente. En otras palabras, las redes tripletas codifican las distancias de las imágenes x^+ y x^- con respecto a la imagen de referencia x. En la figura 5 se puede observar la estructura de este tipo de redes. Para el entrenamiento, hemos utilizado una función de costo triple, la cual está definida por:

$$\text{Loss}(d_+, d_-) = ||d_+, d_- - 1||_2 \quad (4)$$

Donde:

$$d_+ = (e^{||\text{Net}(x)-\text{Net}(x+)||2})/(e^{||\text{Net}(x)-\text{Net}(x+)||2} + e^{||\text{Net}(x)-\text{Net}(x-)||2}) \quad (5)$$

$$d_- = (e^{||\text{Net}(x)-\text{Net}(x-)||2})/(e^{||\text{Net}(x)-\text{Net}(x+)||2} + e^{||\text{Net}(x)-\text{Net}(x-)||2}) \quad (6)$$

Cabe destacar que el objetivo de las redes tripletas (al igual que las redes siamesas) es obtener una distancia muy grande $e^{||\text{Net}(x)-\text{Net}(x-)||2}$ entre las imágenes x y x^- , y una distancia muy corta $e^{||\text{Net}(x)-\text{Net}(x+)||2}$ entre imágenes x y x^+, por lo tanto:

$$\{e^{||\text{Net}(x)-\text{Net}(x+)||2} / e^{||\text{Net}(x)-\text{Net}(x-)||2}\} \rightarrow 0 \quad (7)$$

En otras palabras, estas arquitecturas múltiples son capaces de realizar un aprendizaje de distancia entre imágenes de la misma clase e imágenes de clases diferentes, lo cual es uno de los retos de los sistemas CBIR propuestos en la literatura.

El sistema CBIR propuesto utiliza el potencial de estas arquitecturas múltiples para reducir la brecha semántica, aumentar el poder discriminativo de imágenes y para reducir el espacio de almacenamiento. El primer esquema propuesto es el mostrado en la figura 6, el cual utiliza una red siamesa; consta de dos etapas, la primera (bloque con líneas punteadas) es la etapa de entrenamiento, en donde hacemos uso de una red Alexnet siamesa, detallada en la sección anterior y mostrada en la figura 4. Como se puede observar de la figura 6, los descriptores de las imágenes de la base de datos no son almacenados, en su lugar, solamente son almacenados los pesos sinápticos de la red entrenada.

La segunda etapa es el proceso de recuperación de imágenes similares, en el cual, una entrada de la red siamesa es la imagen de consulta dada, mientras que la otra entrada son todas las imágenes de la base de datos, las cuales son introducidas una por una. Se calcula qué nivel de similitud tienen cada par de imágenes analizado aplicando la distancia euclidiana a las dos salidas de la red siamesa:

$$\text{Similitud} = ||\text{Net}(DB) - \text{Net}(Q)||_2 \quad (8)$$

Donde Net(DB) y Net(Q) son los vectores característicos obtenidos de la capa FC8 de las imágenes de la base de datos DB y la imagen de consulta Q, respectivamente, y $||.||_2$ es la norma L_2. Una vez obtenido el valor de similitud entre cada par de imagen, se extraen las más similares a la consulta dada. De esta manera, el esquema del sistema CBIR queda simplificado, ya que no se necesita el almacenamiento de los descriptores, además, el extractor de características y la medida de similitud se encuentran embebidos en un solo proceso, una red neuronal siamesa.

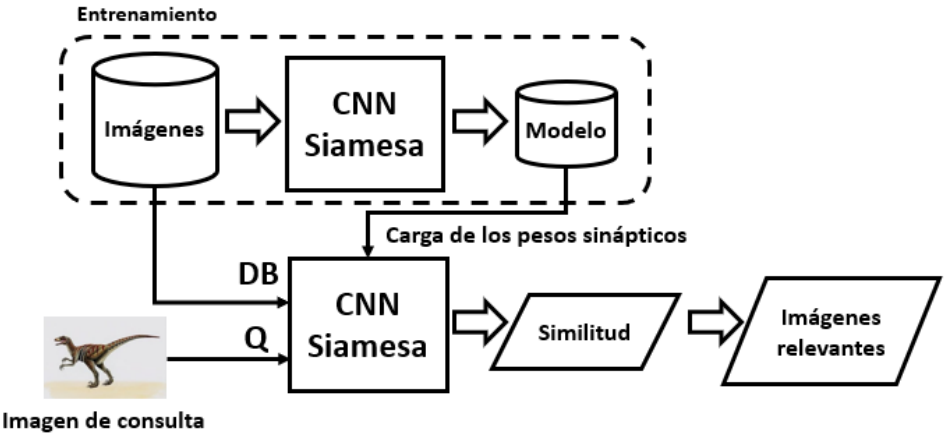

Fig. 6: Sistema CBIR utilizando una red siamesa como extractor de características y medida de similitud

Para el segundo sistema CBIR propuesto hemos utilizado una red tripleta, cuyo entrenamiento se realiza como se detalla en la figura 5. En este esquema no es necesaria una etiqueta binaria, sino que introducimos tres imágenes a la red, dos de la misma clase y una imagen de clase diferente. El entrenamiento es el responsable de aprender qué imágenes del mismo contenido semántico deben tener distancias muy cercanas y qué imágenes de contenido semántico diferente deberían de tener una distancia muy grande. Para la etapa de consulta, utilizamos solamente dos redes (en lugar de tres), ya que a final de cuentas las tres redes comparten los mismos pesos sinápticos. Como resultado tendremos una red siamesa la cual obtendrá el nivel de similitud entre la imagen de consulta y cada una de las imágenes de la base de datos. Al igual que en la arquitectura siamesa, los vectores característicos de las imágenes no son almacenados y el cálculo de similitud, dado por (8), se realiza dentro de la misma arquitectura neuronal, como se ve en la figura 7. Una vez obtenida la similitud entre cada par de imagen, se recuperan las imágenes más similares a la imagen de consulta dada.

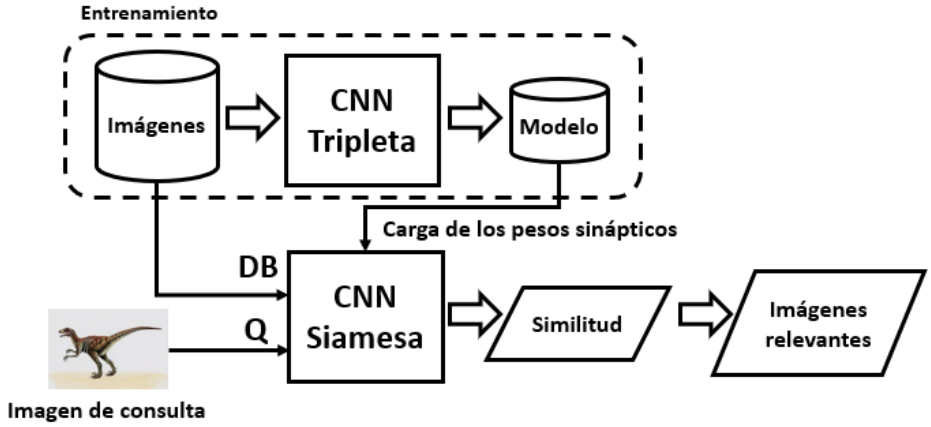

Fig. 7: Sistema CBIR utilizando una red tripleta como extractor de características y medida de similitud

BASES DE DATOS Y EXPERIMENTOS

En este trabajo hemos utilizado dos tipos de bases de datos comúnmente utilizadas en sistemas CBIR: 1) Caltech 101, constituida por 9,146 imágenes divididas en 101 categorías de objetos distintos, donde cada categoría contiene entre 40 y 800 imágenes de dimensiones 300 x 200 pixeles. En este trabajo nos hemos enfocado en el uso de la arquitectura Alexnet, por lo que hemos utilizado 101 neuronas de salida en esta base de datos. La naturaleza de esta base de datos es muy similar a la base de datos ImageNet; 2) Inria Holidays, constituida por 500 grupos de imágenes, donde cada grupo representa una escena u objeto distinto. El número total de imágenes de esta base de datos es 1491, donde 500 de ellas son imágenes de consultas y 991 son imágenes relevantes. Para esta base de datos, hemos utilizado 500 neuronas de salida en la red neuronal. Esta base de datos la hemos utilizado para analizar el funcionamiento de las arquitecturas propuestas (red siamesa y red tripleta) en el caso más crítico existente, ya que, al no haber suficientes imágenes similares, el entrenamiento no se puede completar.

Para la realización de los experimentos, hemos utilizado una GPU *GeForce GTX 1080* y la plataforma de código abierto *Chainer*, la cual es un sistema de aprendizaje profundo con bases en el lenguaje de programación *Python*. Para la comparación de las redes convolucionales en la recuperación de imágenes similares se utilizaron CNN simple (Alexnet) y dos arquitecturas propuestas: CNN Siamesa y CNN Tripleta, de las cuales se extrajeron los siguientes modelos: 1) PT (Modelo pre-entrenado, el cual es el modelo obtenido por el preentrenamiento utilizando la base de datos ImageNet; 2) FT (Modelo afinado), modelo obtenido por el entrenamiento de la red utilizando una nueva base de datos; 3) SN-PT (Modelo siamés pre-entrenado), modelo obtenido por el entrenamiento de una red siamesa; 4) SN-PT-FT (Modelo siamés afinado pre-entrenado), modelo obtenido por el reentrenamiento de la red siamesa utilizando como base el modelo SN-PT; 5) SN-FT (Modelo siamés afinado), modelo obtenido por el reentrenamiento de la red siamesa utilizando como base el modelo PT; 6) TN-PT (Modelo tripleta pre-entrenado), modelo obtenido por el entrenamiento de una red triple; 7) TN-PT-FT (Modelo tripleta afinado pre-entrenado), modelo obtenido por el reentrenamiento de la red triple utilizando como base el modelo TN-PT; 8)TN-FT (Modelo tripleta do), modelo obtenido por el reentrenamiento de la red triple utilizando como base el modelo PT. En el caso de los modelos afinados, las primeras siete capas de la arquitectura Alexnet se han inicializado con los pesos obtenidos en el preentrenamiento (ImageNet), y sólo hemos reentrenado la última capa (FC8), debido a que las bases de datos utilizadas son muy similares a ImageNet.

Como ya se ha mencionado anteriormente, hemos utilizado la estructura de Alexnet. Para la obtención del modelo FT hemos realizado un reentrenamiento de la última capa de la red, además en el entrenamiento se utilizó una tasa de aprendizaje de 0.001 y se entrenó por paquete de 16 imágenes por lote. Este tipo de redes convolucionales son utilizados para la descripción de las imágenes, por lo cual, hemos utilizado la capa FC8 como un descriptor visual y mediante el cálculo de la distancia euclidiana, se calculó la similitud entre imágenes. Para la fase de prueba, se utilizaron imágenes diferentes a las empleadas en el entrenamiento como imágenes de consulta para la búsqueda de imágenes similares.

Para el entrenamiento de la red siamesa se escogieron al azar pares de imágenes positivos y pares negativos y se obtuvieron tres tipos de modelos, SN-PT, SN-PT-FT y SN-FT. Para la obtención de los tres modelos antes mencionados, se realizaron tres entrenamientos (un entrenamiento por modelo) en donde se utilizó una tasa de aprendizaje de 0.001 con un decaimiento de 0.01 por cada 10 épocas y 16 imágenes por lote. Este tipo de redes convolucionales son utilizadas principalmente en el cálculo de similitudes entre imágenes, por lo que hemos utilizado la capa FC8 para encontrar las imágenes similares a una consulta dada, donde estas imágenes de consulta son imágenes diferentes a las utilizadas en el entrenamiento.

El entrenamiento de la red tripleta se realizó de una manera muy similar a la red siamesa; se utilizaron tres imágenes al azar, con una tasa de aprendizaje de 0.001 con decaimiento de 0.01 cada 10 épocas. El tamaño de lote de las imágenes en este caso fue de 16. Al igual que las siamesas, las redes tripletas se utilizan para el cálculo de similitudes, por lo que se utilizó la capa FC8 para extraer las imágenes más similares dada una consulta. Para la evaluación del desempeño de los diferentes tipos de redes sobre diferentes bases de datos hemos utilizado dos de las métricas más utilizadas en esta área de investigación que es la Precisión (P) y la Precisión media promedio (mAP por sus siglas en inglés).

RESULTADOS EXPERIMENTALES

Los resultados obtenidos por los diversos modelos se presentan en la figura 8 y figura 9, donde se puede apreciar el desempeño que tiene cada uno de ellos en las dos métricas utilizadas como evaluación. Como se puede observar en la figura 8, entre más complejo es el modelo (más tiempo de entrenamiento), los resultados serán mejores; por ejemplo, en una red CNN simple, el modelo que obtuvo mejores resultados fue el modelo reentrenado (FT) y para los esquemas de siamesa y de tripleta, el modelo que recuperó más imágenes similares fueron los modelos SN-PT-FT y TN-PT-FT, ya que estos modelos están afinados utilizando la nueva base de datos.

Otro aspecto que debemos observar es que la arquitectura que mostró un mejor desempeño en la recuperación de imágenes similares fue la arquitectura triple, esto se debe a que una red de este tipo tiene más poder discriminativo al tener tres instancias de entrada, por lo que le permite poder conocer las similitudes y disimilitudes de cada una de las imágenes de entrenamiento. Para la base de datos Inria Holidays se observó de la figura 9, que los esquemas de siamesa y de tripleta (que a pesar de tener un poder más discriminativo que una red CNN simple al tener más instancias de entrenamiento) no mostraron un buen desempeño para la recuperación de imágenes similares, debido a la naturaleza de la base de datos. El problema con esta base de datos es que algunas clases sólo contienen una imagen para el entrenamiento lo que dificulta el entrenamiento con un esquema de siamesa o de tripleta, ya que, al presentarse un par de imágenes idéntico, el error de la red tendería a cero, convergiendo exclusivamente para esa clase, perdiendo la capacidad de generalización.

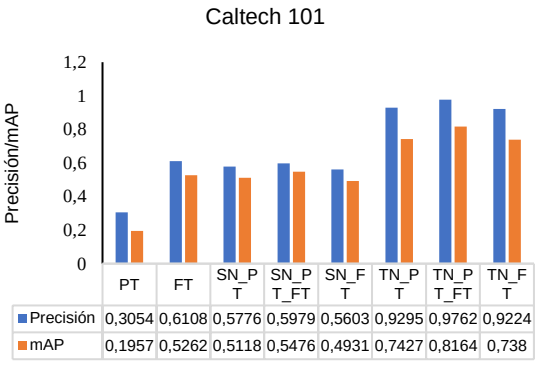

Fig. 8: Resultados de recuperación de imágenes utilizando redes CNN

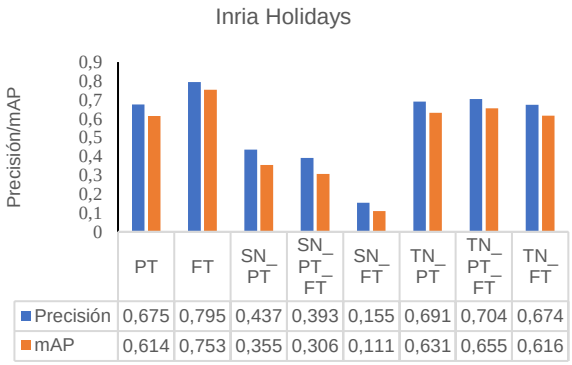

Fig. 9: Resultados de recuperación de imágenes utilizando redes CNN

Los esquemas fundamentados en redes CNN requieren un tiempo considerable para el entrenamiento, ya que se realiza un gran número de operaciones para la obtención de los parámetros óptimos que permiten la convergencia de la red; sin embargo, este proceso se realiza previo a la implementación del sistema CBIR, es decir, cuando se requiere realizar una consulta, los modelos ya están previamente generados, por lo que el tiempo que nos interesa en un sistema CBIR es el de recuperación solamente. En la tabla 1 se puede apreciar que los tiempos para la recuperación de imágenes similares dada una consulta es muy pequeño.

Tabla 1: Tiempos de ejecución

Inria Holidays		Caltech 101	
Referencia	Tiempo (s)	Referencia	Tiempo (s)
PT	1.3	PT	0.12
FT	0.43	FT	0.03
SN_PT	1.4	SN_PT	0.15
SN_PT_FT	0.4	SN_PT_FT	0.05
SN_FT	0.44	SN_FT	0.03
TN_PT	1.1	TN_PT	0.12
TN_PT_FT	0.42	TN_PT_FT	0.03
TN_FT	0.44	TN_FT	0.04

Tabla 2: Comparación entre descriptores

Inria Holidays		Caltech 101	
Referencia	mAP	Referencia	mAP
Jegou et al., 2012	63.4	Ren et al, 2014	38.6
Jegou et al., 2014	72.0	Zhu et al., 2014	43.9
Perronnin et al., 2010	73.5	Tehseen et al., 2019a	65.3
Babenko et al. 2014	71.7	Tehseen et al., 2019b	65.7
Zhang et al., 2015	64.4		
PT	61.4	PT	19.5
FT	75.3	FT	52.6
SN_PT	35.5	SN_PT	51.1
SN_PT_FT	30.6	SN_PT_FT	54.7
SN_FT	11	SN_FT	49.3
TN_PT	63	TN_PT	74.2
TN_PT_FT	65.5	TN_PT_FT	81.6
TN_FT	61.6	TN_FT	73.8

Desde el surgimiento de los sistemas CBIR se han propuesto un gran número de descriptores visuales de bajo nivel. En la tabla 2, en la figura 10 y en la figura 11 se presentan los resultados comparativos utilizando la métrica mAP, obtenidos por estos descriptores de bajo nivel (Ren et al, 2014, Zhu et al., 2014, Jegou et al., 2012, Jegou et al., 2014, Perronnin et al., 2010, Babenko et al., 2014, Tehseen et al., 2019a, Tehseen et al., 2019b y Zhang et al., 2015) y por las características neuronales de alto nivel obtenidas por las redes CNN.

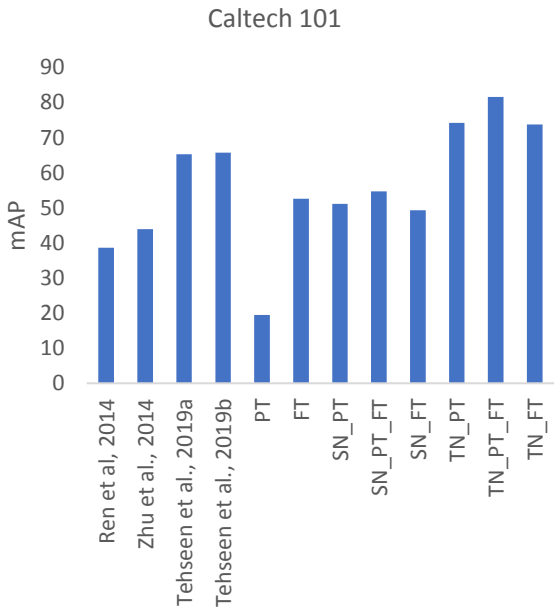

Fig. 10: Comparación de nuestro esquema propuesto con los publicados en la literatura utilizando la base de datos Caltech 101

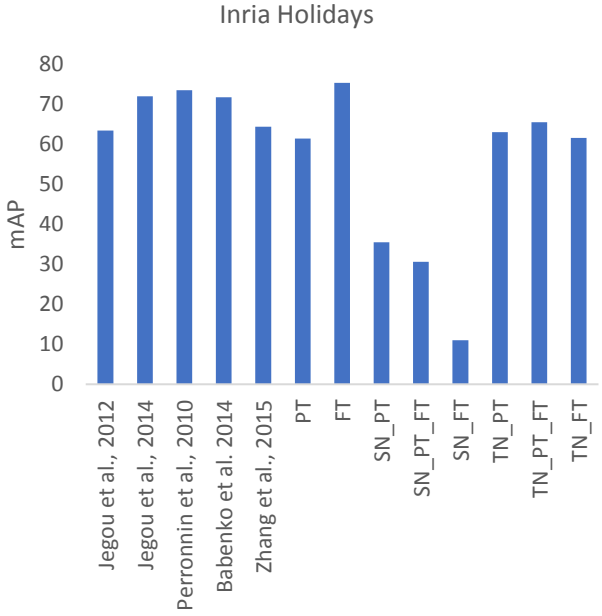

Fig. 11: Comparación de nuestro esquema propuesto con los publicados en la literatura utilizando la base de datos Inria Holidays

Los descriptores fundamentados en la técnica BOV (Ren et al., 2014, Zhu et al., 2014, Jegou et al., 2012 y Zhang et al., 2015) están optimizados para reducir las dimensiones de los descriptores visuales. Como ya se mencionó anteriormente, las redes neuronales CNN extraen características semánticas, imitando la manera en cómo los seres humanos distinguimos las similitudes entre ellas, y como resultado, se obtiene una mejor descripción de las imágenes, mejorando la recuperación de las imágenes similares.

Tabla 3: Complejidad computacional

Esquema	M	K o N	KM o NM
Ren et al, 2014	192	400	76,800
Zhu et al., 2014	128	300	38,400
Jegou et al., 2012	128	1024	131,072
Zhang et al., 2015	8	16,384	131,072
FT _Inria Holiday	500	1,491	745,500
FT_Caltech	101	9,146	923,746

Con respecto a la comparación del tiempo de ejecución, los esquemas fundamentados en CNN se ejecutan en GPUs paralelizando masivamente las operaciones requeridas; sin embargo, los descriptores propuestos por Ren et al, (2014), Zhu et al., (2014), Jegou et al., (2012) y Zhang et al., (2015), no están diseñados para la implementación de cálculos paralelos. Considerando esta diferencia, la comparación se realiza con base al número de operaciones requeridas para la recuperación de imágenes similares dada una imagen de consulta, como se puede observar en la tabla 3. Cabe mencionar que en los descriptores fundamentados en BOV (Ren et al, 2014, Zhu et al., 2014, Jegou et al., 2012 y Zhang et al., 2015), el diccionario generado, que consta de *K* vocabularios visuales los cuales a su vez conforman una palabra visual (descriptor) de dimensión *M,* determinan el número de operaciones, el cual viene dado por O(KM), independientemente del tamaño de la base de datos.

Por otro lado, los esquemas que tienen base en las redes CNN, el número de operaciones está dado por O(NM), donde *N* es el número total de imágenes contenidas en la base de datos y *M* es la dimensión del descriptor. Como ya se mencionó anteriormente, las operaciones en los esquemas con base a las redes CNN se realizan de manera paralela, por lo tanto, aunque el número de operaciones son mayores que los esquemas basados en BOV (Ren et al., 2014, Zhu et al., 2014, Jegou et al., 2012 y Zhang et al., 2015), los tiempos de ejecución son totalmente aceptables para el uso práctico de un sistema CBIR como se muestra en la tabla 1.

CONCLUSIONES

De acuerdo a los resultados obtenidos en este estudio, se puede concluir que el sistema propuesto resuelve tres de los más importantes retos en los sistemas CBIR de manera exitosa: 1) el esquema propuesto realiza una mejor descripción de las imágenes debido a su gran poder de extracción de información semántica de alto nivel, reduciendo así la brecha semántica que existe entre la información numérica de los pixeles y los conceptos semánticos que los humanos percibimos; 2) uno de los grandes aportes del sistema propuesto es la realización de un aprendizaje de similitud, el cual se lleva a cabo codificando los vectores característicos de las imágenes; imágenes similares tendrán una distancia más pequeña entre ellas, mientras que imágenes no similares tendrán distancia más grande. En otras palabras, estas arquitecturas múltiples permiten capturar la similitud y disimilitud entre imágenes de interés; 3) otra de las grandes contribuciones de este trabajo es que el sistema propuesto no necesita almacenar los descriptores visuales de cada imagen, solo basta con almacenar el modelo entrenado (pesos sinápticos) para realizar la recuperación de imágenes; el espacio de almacenamiento del modelo no depende del tamaño de la base de datos. Estas tres contribuciones arriba mencionadas hacen que el sistema propuesto sea una considerable contribución en el campo de los sistemas CBIR. Una limitación de los sistemas propuestos es que requiere una suficiente cantidad de imágenes similares para su entrenamiento, ya que, si no cumple con este requisito, como es el caso de la base de datos Inria Holidays, los sistemas propuestos no pueden ofrecer resultados satisfactorios.

REFERENCIAS

Babenko A.; A. Slesarev y otros dos autores, *Neural Codes for Image Retrieval*; European Conference on Computer Vision, 584-599, Zurich, Suiza, 6-12 septiembre (2014)

Bai C.; L. Huang y otros tres autores, *Optimization of Deep Convolutional Neural Network for Large Scale Image Retrieval*; Neurocomputing, 303, 60-67 (2018)

Bangio Y.; A. Courville y V. Pascal, *Representation Learning: A Review and New Perspectives*; IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798-1828, (2013)

Calderón G.; A. Fierro y otros dos autores, *Efecto de la Transformada Motif en Desarrollo de Descriptores de Textura para la Extracción de Imágenes;* Información Tecnológica, 27(3), 199-214 (2016)

Chandrasekhar V.; J. Lin y otros cinco autores, *Compression of Deep Neural Networks for Image Instance Retrieval*; Data Compression Conference, 300-309, Snowbird, Estados Unidos de América, 4-7 abril (2017).

Guimaraes. D. y R. Torres, *Unsupervised Rank Diffusion for Content-Based Image Retrieval*; Neurocomputing, 260, 478-489 (2017)

Hoffer E. y N. Ailon, *Deep Metric Learning Using Triplet Network*; International Workshop on Similarity-Based Pattern Recognition, 84-92, Copenhagen, Dinamarca (2015)

Jegou H. y A. Zisserman, *Triangulation embedding and democratic aggregation for image search*; IEEE International Conference on Computer Vision and Pattern Recognition, 3310-3317, Columbus, OH, USA (2014)

Jegou H.; F. Perronnin y otros cuatro autores, *Aggregating Local Image Descriptors into Compact Codes*; IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(9), 1704-1716 (2012)

Krizhevsky A.; I. Sutskever y G. Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*; International Conference on Neural Information Processing Systems, 1097-1105, Nevada, Estados Unidos de America, 3-8 diciembre (2012)

Lecun Y.; L. Bottou, y otros dos autores, *Gradient-Based Learning Applied to Document Recognition;* Proceedings of the IEEE, 86(11), 2278-2324 (1998)

Melekhov I.; J. Kannala y E. Rahtu, *Siamese Network Features for Image Matching*; International Conference on Pattern Recognition, Cancún, México, 4-8 diciembre (2016)

Meng F.; D. Shan y otros cuatro autores, *Merged Region-Based Image Retrieval;* Visual Communication and Image Representation, 55, 572-585 (2018)

Meng J.; Y. Jiang y otros dos autores, *Support Top Irrelevant Machine: Learning Similarity Measures to Maximize Top Precision for Image Retrieval*; Neural Computing & Applications, 28, 1145-1154 (2017)

Montazar G. y D. Giveki, *Content Based Image Retrieval Systems Using Clustered Scale Invariant Feature Transforms*; Optik, 126(18), 1695-1699 (2015)

Perronnin F.; Y. Liu y otros dos autores, *Large-Scale Image Retrieval with Compressed Fisher Vectors*; IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 3384-3391, San Francisco, CA, USA (2010)

Ren Y.; J. Benois y A. Burgeau, *A Comparative Study of Irregular Pyramid Matching in Bag-of-Bags of Words Model for Image Retrieval*; International Conference on Image and Signal Processing, 539-548, Cherburgo, Francia, 30 junio – 2 julio (2014)

Russakovsky O.; J. Deng y otros diez autores, *ImageNet Large Scale Visual Recognition Challenge*; Int. J. of Computer Vision, 115(3), 211-252 (2015)

Song W.; Y. Zhang y otros cinco autores, *Taking Advantage of Multi-Regions-Based Diagonal Texture Structure Descriptor for Image Retrieval*; Expert Systems with Applications, 96, 347-357 (2018)

Tehseen K; S. Ali y otros dos autores, *Convolution, Approximation and Spatial Information Based on Object and Color Signatures for Content Based Image Retrieval*; International Conference on Computer and Information Sciences, Sakaka, Saudi Arabia, Saudi Arabia (2019a)

Tehseen K.; S. Ummesafi y A. Iqbal, *Content Based Image Retrieval Using Image Features Information Fusion;* Information Fusion, 51, 76-99 (2019b)

Tian X.; L. Jiao y otros dos autores, *Feature Integration of EODH and Color-SIFT: Application to Image Retrieval Based on Codebook*; Signal Processing: Image Communication*,* 29(4), 530-545 (2014)

Tzelepi M. y A. Tefas, *Deep Convolutional Image Retrieval: A General Framework;* Signal Processing: Image Communication, 63, 30-43 (2018)

Wang X.; Y. Yu y H. Yang, *An Effective Image Retrieval Ccheme Using Color, Texture and Shape Features*; Computer Standards & Interfaces, 33(1), 59-68 (2011)

Zhang T.; G. Qi y otros dos autores, *Sparse Composite Quantization*; IEEE International Conference on Computer Vision and Pattern Recognition, 4548-4556, Boston, Estados Unidos de América, 7-12 junio (2015)

Zhu L.; H. Jin y otros dos autores, *Weighting Scheme for Image Retrieval Based on Bag-of-Visual-Words*; IET Image Processing, 8(9), 509-518 (2014)