Revista Cubana de Ciencias Informáticas
Vol. 12, No. 4, Octubre-Diciembre, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 78- 90
http://rcci.uci.cu

Tipo de artículo: Artículo original
Temática: Inteligencia artificial
Recibido: 03/10/2017 | Aceptado: 05/10/2018

Recuperación de imágenes por contenido usando descriptores generados por Redes Neuronales Convolucionales

Content-based image retrieval using descriptors generated by Convolutional Neural Networks

Sergio Sánchez Santiesteban*

Universidad de las Ciencias Informáticas. Carretera a San Antonio de los Baños, km 2 1/2 La Lisa, La Habana, Cuba.

*Autor para correspondencia: ssantiesteban@estudiantes.uci.cu

Resumen
Los sistemas para la recuperación de imágenes basada en contenido permiten la búsqueda y recuperación de imágenes que son similares a una imagen de consulta dada, empleando rasgos que representan el contenido visual de dichas imágenes. En el presente trabajo se desarrolló un método para la recuperación de imágenes indexadas en bases de datos a partir de su contenido visual, sin necesidad de realizar anotaciones textuales. Se obtuvieron vectores de rasgos a partir de los contenidos visuales mediante técnicas de redes neuronales artificiales con aprendizaje profundo. Se propuso el empleo de redes neuronales convolucionales pre entrenadas para crear los descriptores globales. Se aplicaron técnicas de reducción de la dimensión para incrementar la eficiencia en el procesamiento. Los resultados obtenidos por el método propuesto, sobre bases de datos disponibles públicamente, fueron superiores a los de los métodos tradicionales y comparables con otros basados en aprendizaje profundo, que constituyen el estado del arte en la recuperación de imágenes por contenido. El método propuesto puede ser extendido mediante la adición de etapas posteriores de integración de rasgos con mayor grado de abstracción.

Palabras claves: descriptores globales, recuperación de imágenes, recuperación de información, Redes Neuronales Convolucionales.

Abstract
Content-Based Image Retrieval systems allow to search and retrieve images that are similar to a given query image using features for representing the visual content of the images. In this work it was developed a method to retrieve digital images indexed in databases using its visual content, without textual annotations. Automatic descriptions of the contents were obtained using deep neural networks. Pre-trained Convolutional Neural Networks were proposed to create global descriptors. Dimensionality reduction techniques were applied to increase the efficiency in performance. Results obtained by this method, over two publicly available datasets, were better than performance of traditional methods and comparable to other approaches based on deep learning which are the state of the art in Content-Based Image Retrieval. Proposed method could be extended by the addition of stages of feature integration with a greater degree of abstraction.

Revista Cubana de Ciencias Informáticas
Vol. 12, No. 4, Octubre-Diciembre, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 78- 90
http://rcci.uci.cu

Keywords: *Convolutional Neural Networks, global descriptors, image retrieval, information retrieval.*

Introducción

Las técnicas de recuperación de imágenes basada en contenido (CBIR[1] por sus siglas en inglés) dan solución a un problema de recuperación de información que puede plantearse de la siguiente forma: a partir de una imagen de interés recuperar u obtener imágenes similares de entre las presentes en una gran colección, utilizando solamente características o rasgos extraídos de dichas imágenes (Tunga et al., 2015). Se entiende por imágenes similares aquellas en las que se observa el mismo objeto o escena con variaciones en la perspectiva, las condiciones de iluminación o la escala. Las imágenes almacenadas son pre procesadas y luego se indexan sus correspondientes descriptores. La imagen de consulta también es pre procesada para extraer su descriptor, que luego es comparado con los almacenados aplicando medidas de similitud apropiadas, que permitan la recuperación de aquellas imágenes que sean similares a la imagen de consulta.

Las técnicas de CBIR son utilizadas en varias ramas de las ciencias como son la medicina (Dhara et al., 2017; Anavi et al., 2016), agricultura, seguridad y protección, pronóstico del tiempo, modelado de procesos biológicos, clasificación de imágenes web (Vakhitov et al., 2016), prevención del crimen, procesamiento de imágenes de satélite, entre otras.

Los enfoques tradicionales incluyen principalmente la elaboración de descriptores a partir del contenido de la imagen, mediante los llamados rasgos de bajo nivel como son el color (Liu and Yang, 2013), la textura (Lasmar and Berthoumieu, 2014) y la forma (Wang et al., 2015) o una combinación de algunos de estos (Wang et al., 2014). Un aspecto positivo de las técnicas desarrolladas sobre la base de estos enfoques es que no demandan grandes cantidades de datos ni de tiempo para obtener resultados satisfactorios durante las etapas de entrenamiento e inferencia. Por otra parte, pretenden obtener descriptores locales y globales de las imágenes a partir de rasgos elaborados manualmente, que no son genéricos, sino que poseen una marcada dependencia ante las clases representadas en las imágenes. Generalmente restringen sus posibilidades de escalar con éxito hacia colecciones de imágenes con grandes cantidades de clases o categorías.

Recientes métodos combinan rasgos de bajo nivel con otros llamados de alto nivel, que proporcionan una representación más cercana a la percepción humana, permitiendo alcanzar una descripción semántica de las imágenes y lograr mejores resultados en su recuperación. Los principales avances en esta dirección están aparejados al rápido desarrollo de las técnicas de aprendizaje de máquinas y específicamente al aprendizaje profundo o *deep learning* (Chandrasekhar et al., 2017; Yu et al., 2017; Tzelepi and Tefas, 2018). Los modelos neuronales

[1] *Content-Based Image Retrieval*

aprenden descriptores globales elaborados sobre la base de una jerarquía de rasgos y ajustados mediante un proceso de entrenamiento. Estos descriptores son genéricos y robustos ante retos como la variabilidad entre las clases, oclusión o cambios de perspectiva o iluminación. Sin embargo, los vectores de rasgos obtenidos mediante estas técnicas, poseen en la mayoría de los casos, una gran dimensión (2048, 4096 componentes), repercutiendo negativamente en el uso de memoria para su almacenamiento y en la complejidad temporal del proceso de comparación y recuperación.

Varias medidas de disimilitud son utilizadas durante la comparación de los descriptores de las imágenes, las más comunes son: la distancia euclidiana, la distancia de Bhattacharya, la distancia de Mahalanobis, la distancia de Sorensen y la distancia del coseno (Tunga et al., 2015).

Se pretende resolver el problema planteado desarrollando un método donde se utilicen algoritmos de aprendizaje profundo, específicamente las Redes Neuronales Convolucionales, para obtener descriptores globales de las imágenes. Se reducirá la dimensión de estas representaciones vectoriales mediante la aplicación del Análisis de Componentes Principales. Para determinar la similitud entre las imágenes se emplearán medidas robustas, con amplio uso dentro del dominio de la recuperación de imágenes.

Las contribuciones fundamentales del presente trabajo son las siguientes:

- se desarrolla un método para la obtención de descriptores globales aprendidos a partir de las imágenes, mediante Redes Neuronales Convolucionales.
- se introduce el uso del algoritmo de Análisis de Componentes Principales para añadir robustez a los descriptores globales a partir de la reducción de sus dimensiones.
- se propone la utilización de la distancia de Sorensen y la distancia del coseno para el cálculo de la disimilitud entre los descriptores globales de diferentes imágenes.

Materiales y métodos

Las Redes Neuronales Convolucionales (CNN[2] por sus siglas en inglés) (LeCun et al., 1989) son un tipo específico de red neuronal para el procesamiento de datos que posean una topología tipo malla, por ejemplo, las imágenes pueden ser vistas como una malla bidimensional de píxeles. Este tipo de red emplea la operación matemática convolución en lugar de la multiplicación de matrices, para determinar el grado de respuesta de la imagen ante la aplicación de un determinado filtro o núcleo, produciendo como resultado un mapa de características de la imagen convolucionada.

[2] *Convolutional Neural Networks*

Esta operación para el caso específico de las imágenes, dada una imagen I y un núcleo K, los valores obtenidos como respuesta a la aplicación de este núcleo están dados según la ecuación (1):

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i+m, j+n) * K(m,n) \tag{1}$$

Mediante la implementación en las CNN de esta operación, un mapa de características se obtiene con la aplicación repetitiva del núcleo a través de subregiones de la imagen completa. Denotando el $k - ésimo$ mapa de características como h_k, cuyo filtro está determinado por los pesos W_k y el *bias* b_k (2):

$$h_{ij}^k = \tanh((W^k * x)_{ij} + b_k) \tag{2}$$

Los valores de los pesos son compartidos por todas las neuronas de la misma capa, reduciendo así, el número de hyperparámetros a aprender. Los mapas de características obtenidos son sometidos a sucesivas operaciones de submuestreo que reemplazan la salida de la red en determinados puntos por resúmenes de las salidas cercanas. Para esto, se definen ventanas sin solapamiento sobre el mapa, de las que se selecciona solo un valor (*max-pooling*, *average-pooling*) con el objetivo de reducir el número de parámetros a aprender y obtener neuronas más robustas a la posición exacta de los estímulos y que abarquen mayor porción del campo visual (Goodfellow et al., 2016).

El Análisis de Componentes Principales (PCA[3] por sus siglas en inglés) es un algoritmo no supervisado que aprende una representación de los datos con menor dimensión que la entrada. Formalmente el algoritmo PCA aprende una transformación lineal y ortogonal de los datos que proyecta una entrada x hacia una representación z, usualmente con menor dimensión (Goodfellow et al., 2016).

La distancia del coseno es una conocida medida de disimilitud empleada en la comparación de vectores, dados los vectores u y v de dimensión n, la distancia del coseno entre ellos se determina mediante la ecuación (3); en todos los casos se utiliza el producto punto:

$$d(u,v) = \frac{1 - (u^T v)}{||u||||v||} \tag{3}$$

La distancia de Sorensen o distancia de Bray-Curtis es una métrica definida en $\mathbb{R}^n$ para la comparación de

[3] *Principal Components Analysis*

Revista Cubana de Ciencias Informáticas
Vol. 12, No. 4, Octubre-Diciembre, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 78- 90
http://rcci.uci.cu

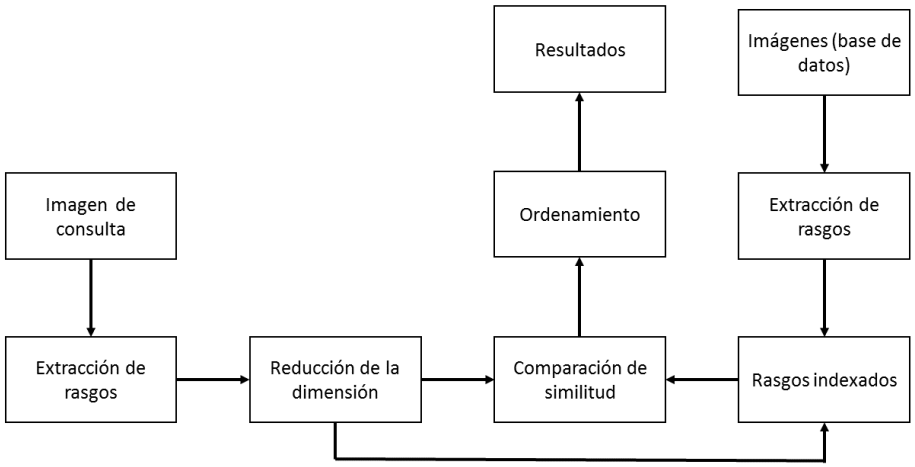

Figura 1. Arquitectura general del método propuesto.

vectores $1 - D$ según la ecuación (4), donde x_i, y_i son las componentes de los vectores dados:

$$d(x, y) = \frac{\sum ||x_i - y_i||}{\sum (x_i + y_i)} \tag{4}$$

Según S.Banuchitra and K.Kungumaraj (2016) el paradigma para la recuperación de imágenes basada en contenido se puede descomponer en las siguientes etapas: adquisición de la imagen, pre procesamiento, extracción de características, comparación de la similitud e imágenes recuperadas como resultado. Algunos sistemas más recientes incluyen además técnicas de retroalimentación. En correspondencia con las etapas anteriores el método que se propone posee la arquitectura general que se observa en la Figura 1, a la que se adiciona una etapa de reducción de la dimensión a las representaciones vectoriales de las imágenes.

Extracción de rasgos: durante esta etapa se emplea una CNN estructurada de acuerdo al modelo Inceptionv3 (Szegedy et al., 2016), en el que se presenta una arquitectura neuronal profunda de 42 capas, cuyas mejoras con respecto a arquitecturas anteriores como GoogLeNet (Szegedy et al., 2014) o VGGNet (Simonyan and Zisserman, 2014) están dadas principalmente por la factorización de los núcleos de convolución con la consiguiente reducción espacial de estos filtros, el uso de clasificadores auxiliares al final del entrenamiento y la mejora de las técnicas para la reducción de la cantidad de parámetros a aprender.

Mediante una propagación hacia delante de la imagen, a través de la red, se obtiene el tensor $pool_3$ que, como parte del grafo construido con Tensorflow (Abadi et al., 2016), almacena una representación vectorial de 2048 dimensiones de la imagen. Esta etapa se realiza en dos momentos: primero durante el proceso de indexación de todas las imágenes disponibles, y se guarda junto a cada imagen en la base de datos, y luego para la imagen de consulta.

Reducción de la dimensión: los descriptores vectoriales obtenidos en la etapa anterior poseen 2048 componentes, con el objetivo de reducir este valor se aplica la técnica de PCA. Primero se realiza el cálculo de la matriz de covarianza de las representaciones vectoriales de una muestra de las imágenes disponibles. Luego se calculan los vectores propios y se ordenan por el valor absoluto de su valor propio asociado. Los primeros vectores propios son seleccionados como una nueva base permitiendo que el vector de 2048 componentes, que representa el contenido de la imagen, sea proyectado hacia un espacio de menor dimensión y en el que aún se conserven las capacidades representativa y diferenciadora de dichos descriptores.

Comparación de la similitud: para determinar cuáles imágenes en la base de datos son similares a la imagen de consulta se tiene en cuenta el valor obtenido a partir de la aplicación de una función de disimilitud, como la distancia del coseno o la distancia de Sorensen, sobre los descriptores vectoriales de la imagen de consulta y las indexadas en la base de datos.

Ordenamiento y recuperación: las imágenes similares son ordenadas a partir del valor obtenido durante la etapa anterior, de manera que aquellas que más similares sean a la imagen de consulta son mostradas en los primeros puestos de la lista de candidatas.

Resultados y discusión

El proceso de experimentación se realizó sobre las siguientes bases de datos internacionales, utilizadas por otros autores en trabajos relacionados con la temática.

UKBench: esta es una base de datos del Departamento de Ciencias de la Computación de la Universidad de Kentucky, cuenta con 10200 imágenes en colores, con dimensiones de 640x480 píxeles. Las imágenes se encuentran organizadas en 2550 grupos de 4 imágenes cada uno con capturas del mismo objeto o escena tomadas con radicales cambios de perspectiva, como se puede apreciar en la Figura 2 (Nister and Stewenius, 2006).

Holidays: INRIA Holidays es un conjunto de imágenes sobre escenas o lugares que fueron tomadas por el Instituto Nacional de Investigación en Informática y Automática (INRIA) de Francia, con el propósito de

Figura 2. Grupos de 4 imágenes en UKBench, se puede observar (tercer y cuarto grupos) la similitud entre imágenes pertenecientes a grupos diferentes.

poseer variedad en cuanto a: rotaciones, punto de vista y cambios en la iluminación. La base de datos incluye diversos tipos de escenas: naturales, hechas por el hombre, etc; todas en alta resolución. Holidays contiene 500 grupos de imágenes, cada uno de los cuales representa una escena u objetos distinto. La primera imagen de cada grupo es la imagen de consulta y los resultados correctos son el resto de los miembros del grupo (Jegou et al., 2008).

Los pesos de la red fueron inicializados utilizando un modelo pre entrenado sobre la base de datos ImageNet del 2012. Inception-v3 fue entrenado para el Reto de Reconocimiento Visual a Gran Escala de ImageNet (ILSVRC). Esta es una tarea estándar en visión por computadoras donde los modelos tratan de clasificar imágenes completas en mil clases. Para comparar los modelos se examina sus fallos al predecir la respuesta correcta como una de las primeras cinco clases predichas (tasa de error *top*-5), Inception-v3 alcanza el 3.46 % en el error *top*-5 (Szegedy et al., 2016).

Como métricas para medir el desempeño del método propuesto, sobre las bases de datos seleccionadas, se emplearon las siguientes:

4xRecall@4: se utiliza cada imagen, por turnos, como imagen de consulta y se reporta el valor promedio de verdaderos positivos obtenidos dentro de las primeras 4 imágenes (*top*-4) recuperadas.

mAP (*Mean Average Precision*): Dado un conjunto de imágenes de consulta esta métrica se define como,

Revista Cubana de Ciencias Informáticas
Vol. 12, No. 4, Octubre-Diciembre, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 78- 90
http://rcci.uci.cu

$$mAP = \frac{\sum_{q=1}^{Q} AvePr(q)}{Q}$$

donde *Average Precision* ($AvePr$) para cada imagen de consulta se define como,

$$AvePr = \frac{\sum_{k=1}^{n}(Pr(k) * rel(k))}{R}$$

donde k es el rango en la secuencia de imágenes recuperadas, n es el número de imágenes recuperadas, $Pr(k)$ es la precisión en k en la lista ($Pr@k$), $rel(k)$ es una función que toma valor 1 si la imagen con rango k es relevante o 0 de otro modo y R es el número de imágenes relevantes (Napoletano, 2018).

Los modelos neuronales se implentaron utilizando Tensorflow y se ejecutaron en una PC con procesador Core i5 3.5 GHz con 4 GB de RAM y con una GPU NVIDIA GeForce GTX 850M. En la tabla 1 se muestra un resumen de las características de las bases de datos utilizadas para la experimentación. La tabla 2 muestra los resultados que se obtuvieron sobre las bases de datos UKBench y Holidays con las imágenes a resolución original, reduciendo la dimensión de los vectores de descripción, aplicando PCA, a 512 y 1024 dimensiones y empleando distintas medidas de disimilitud para la comparación de estos. También se presentan los valores obtenidos por otros métodos clásicos y del estado del arte.

Tabla 1. Características de las bases de datos utilizadas.

Base de datos	**Año**	**Cantidad de imágenes**	**Mínima resolución**
UKBench (Nister and Stewenius, 2006)	2006	10200	640x480
INRIA Holidays (Jegou et al., 2008)	2008	1491	848x480

Se pudo constatar que los mejores resultados se logran cuando se reduce la dimensión de los descriptores originales eliminando información no útil o con cierto *ruido* presente en dichos descriptores. Los nuevos vectores de rasgos permiten un almacenamiento más óptimo al reducirse el consumo de memoria y dan lugar a mejores resultados en el proceso de recuperación. Las distancias del coseno y de Sorensen son las que mejores resultados obtuvieron sobre los vectores originales y al reducir su dimensión también.

Los falsos positivos se obtienen mayoritariamente al recuperarse imágenes que no pertenecen al grupo de resultados correctos, aunque representan el mismo tipo de objeto tomado desde ángulos y condiciones de iluminación

Revista Cubana de Ciencias Informáticas
Vol. 12, No. 4, Octubre-Diciembre, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 78- 90
http://rcci.uci.cu

Tabla 2. Resultados obtenidos sobre las bases de datos UKBench y Holidays.

Método	**UKBench (4xRecall@4)**	**Holidays (mAP)**	**Métrica de distancia**	**Dimensiones Descriptor**
SIFT (Paulin et al., 2017)	3.44	-	-	1024
AlexNet-conv3 (Paulin et al., 2017)	3.74	-	-	-
CKN-mix (Paulin et al., 2017)	3.77	-	-	4096
Res50-NIP (Chandrasekhar et al., 2017)	3.88	-	-	2048
(Zhang et al., 2015)	-	64.4	-	128 bits
(Gong et al., 2014)	-	80.18	-	2048
(Jegou et al., 2012)	-	68.9	-	262k dims.
(Perronnin and Larlus, 2015)	-	84.7	-	4096
(Jégou and Zisserman, 2014)	-	77.1	-	8064
Método propuesto	3.55	70.86	coseno	2048
Método propuesto	3.55	70.88	coseno	1024
Método propuesto	**3.62**	72.12	**Sorensen**	**1024**
Método propuesto	**3.53**	-	**coseno**	**512**
Método propuesto	-	**72.18**	**Sorensen**	**512**

similares. Dichas imágenes son recuperadas primero que los miembros del grupo de la imagen de consulta, que varían drásticamente la perspectiva de captura. Es necesario reconocer que las imágenes recuperadas, desde el punto de vista semántico, son realmente similares a la de consulta y el error cometido está estrechamente relacionado a la escasa variabilidad entre determinados grupos de imágenes presentes en la base de datos sobre la que se experimenta. La incapacidad de los descriptores obtenidos para ser discriminatorio ante las situaciones mencionadas está dada por el hecho de que la red ha sido entrenada para realizar tareas de clasificación y por tanto los rasgos aprendidos tienden a ser robustos ante la variabilidad entre las clases y pierden especificidad a las diferencias existentes entre imágenes, que al ser clasificadas se encontrarían en la misma clase o categoría. Alternativas que complementan estas limitaciones se describen en (Chandrasekhar et al., 2017) donde, en lugar de utilizar los vectores de activación de las últimas capas de la red neuronal entrenada, elaboran descriptores que integran dichos vectores con técnicas estadísticas y aplican transformaciones a la imagen de entrada para hacerlos más invariantes. Sin embargo, las alternativas mencionadas incrementan la dimensión de los descriptores, el volumen de operaciones necesarias por cada imagen a procesar y la complejidad temporal de estas, demandando mayor poder de cómputo y espacio en memoria.

Los resultados obtenidos, que mejoran significativamente los logrados por algoritmos tradicionales como los descriptores SIFT, son alentadores teniendo en cuenta que se ha utilizado un modelo pre entrenado, sin ajustar

para la base de datos evaluada, y que las dimensiones de los descriptores vectoriales globales son menores que las empleadas por otros métodos. Utilizando descriptores hasta 8 veces más pequeños (4096 vs 512) se lograron resultados inferiores a los del estado del arte solo por un margen mínimo. El espacio en memoria que ocupan los descriptores globales de las imágenes indexadas es un tema importante al tratarse de cantidades del orden de los miles o millones de imágenes. El volumen de operaciones por imagen demandadas por nuestro método, así como la baja dimensión de los descriptores a procesar y almacenar, lo hacen factible para entornos donde se necesite un balance entre el consumo de tiempo y recursos del sistema de recuperación y la obtención de resultados similares en eficacia a los del estado del arte.

Se pretende continuar mejorando el desempeño del método mediante la utilización de otros modelos neuronales más recientes y aplicando transformaciones de rotación y escalado a la imagen de entrada, siguiendo un enfoque similar al abordado en (Chandrasekhar et al., 2017) pero optimizando el costo computacional.

El método propuesto puede ser empleado en el desarrollo de componentes para sistemas de recuperación de información, específicamente recuperación de imágenes basada en contenido, humanizando la tarea de recuperar imágenes de interés almacenadas en grandes colecciones de datos.

Conclusiones

Se desarrolló un método para la recuperación de imágenes basada en contenido que abordó todas las etapas de este proceso, haciendo uso de los recientes avances en las redes neuronales artificiales con aprendizaje profundo.

Se lograron resultados comparables a los reportados por métodos del estado del arte y mejores que los obtenidos mediante técnicas tradicionales.

El desempeño, sobre bases de datos internacionales, validó la eficacia del método, convirtiéndolo en un punto de partida para trabajos futuros en esta área.

Se pretende continuar desarrollando el método con la adición de etapas para el pre procesamiento e integración de los descriptores obtenidos, en otros con mayor grado de abstracción.

Referencias

Martin Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, and others. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

Yaron Anavi, Ilya Kogan, Elad Gelbart, Ofer Geva, and Hayit Greenspan. Visualizing and enhancing a deep learning framework using patients age and gender for chest x-ray image retrieval. In *SPIE Medical Imaging*, pages 978510–978510. International Society for Optics and Photonics, 2016.

Vijay Chandrasekhar, Jie Lin, Qianli Liao, Olivier Morere, Antoine Veillard, Lingyu Duan, and Tomaso Poggio. Compression of Deep Neural Networks for Image Instance Retrieval. *arXiv preprint arXiv:1701.04923*, 2017.

Ashis Kumar Dhara, Sudipta Mukhopadhyay, Anirvan Dutta, Mandeep Garg, and Niranjan Khandelwal. Content-Based Image Retrieval System for Pulmonary Nodules: Assisting Radiologists in Self-Learning and Diagnosis of Lung Cancer. *Journal of Digital Imaging*, 30(1):63–77, February 2017. ISSN 0897-1889, 1618-727X. doi: 10.1007/s10278-016-9904-y. URL https://link.springer.com/article/10.1007/s10278-016-9904-y.

Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *European conference on computer vision*, pages 392–407. Springer, 2014.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Hervé Jégou and Andrew Zisserman. Triangulation embedding and democratic aggregation for image search. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3310–3317, 2014.

Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. *Computer Vision–ECCV 2008*, pages 304–317, 2008.

Herve Jegou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Perez, and Cordelia Schmid. Aggregating local image descriptors into compact codes. *IEEE transactions on pattern analysis and machine intelligence*, 34(9):1704–1716, 2012.

Nour-Eddine Lasmar and Yannick Berthoumieu. Gaussian copula multivariate modeling for texture image retrieval using wavelet transforms. *IEEE Transactions on Image Processing*, 23(5):2246–2261, 2014.

Yann LeCun et al. Generalization and network design strategies. *Connectionism in perspective*, pages 143–155, 1989.

Guang-Hai Liu and Jing-Yu Yang. Content-based image retrieval using color difference histogram. *Pattern Recognition*, 46(1):188–198, 2013.

Paolo Napoletano. Visual descriptors for content-based retrieval of remote-sensing images. *International Journal of Remote Sensing*, 39(5):1–34, 2018.

D. Nister and H. Stewenius. Scalable Recognition with a Vocabulary Tree. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2161–2168, June 2006. oral presentation.

Mattis Paulin et al. Convolutional patch representations for image retrieval: An unsupervised approach. *International Journal of Computer Vision*, pages 165–166, 2017.

Florent Perronnin and Diane Larlus. Fisher vectors meet neural networks: A hybrid classification architecture. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3743–3752, 2015.

S.Banuchitra and K.Kungumaraj. A Comprehensive Survey of Content Based Image Retrieval Techniques. *International Journal Of Engineering And Computer Science(IJECS)*, 5, August 2016. doi: 10.18535/ijecs/v5i8.26. URL https://www.ijecs.in.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. *CoRR*, abs/1409.4842, 2014. URL http://arxiv.org/abs/1409.4842.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

Satish Tunga, D Jayadevappa, and C Gururaj. A comparative study of content based image retrieval trends and approaches. *International Journal of Image Processing (IJIP)*, 9(3):127, 2015.

Maria Tzelepi and Anastasios Tefas. Deep convolutional learning for content based image retrieval. *Neurocomputing*, 275:2467–2478, 2018.

Alexander Vakhitov, Andrey Kuzmin, and Victor Lempitsky. Internet-Based Image Retrieval Using End-to-End Trained Deep Distributions. *arXiv preprint arXiv:1612.07697*, 2016.

Xiang-Yang Wang, Bei-Bei Zhang, and Hong-Ying Yang. Content-based image retrieval by integrating color and texture features. *Multimedia Tools and Applications*, 68(3):545–569, 2014.

Xinjian Wang, Guangchun Luo, and Ke Qin. A composite descriptor for shape image retrieval. In *International Conference on Automation, Mechanical Control and Computational Engineering*, pages 759–764, 2015.

Jiachen Yang, Bin Jiang, Baihua Li, Kun Tian, and Zhihan Lv. A fast image retrieval method designed for network big data. *IEEE Transactions on Industrial Informatics*, 2017.

Wei Yu, Kuiyuan Yang, Hongxun Yao, Xiaoshuai Sun, and Pengfei Xu. Exploiting the complementary strengths of multi-layer CNN features for image retrieval. *Neurocomputing*, 237:235–241, 2017.

Ting Zhang, Guo-Jun Qi, Jinhui Tang, and Jingdong Wang. Sparse composite quantization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4548–4556, 2015.