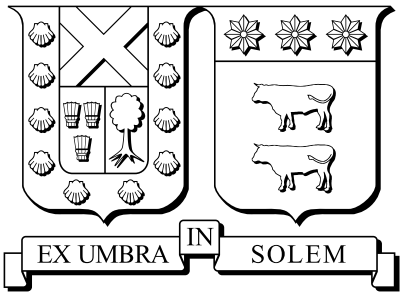
UNIVERSIDAD TECNICA FEDERICO SANTA MARIA
Departamento de Informática
Valparaíso - Chile



Concatenación de representaciones compactas de imágenes en el problema de recuperación de imágenes

Tesis presentada como requerimiento parcial
para optar al grado académico de

Magíster en Ciencias de la Ingeniería Informática

y al título profesional de

Ingeniería Civil Informática

por

Tomás Mardones Latham

Comité de evaluación
Prof. Dr. rer. nat. Héctor Allende Olivares (Profesor Guía, UTFSM)
Dr. Marcelo Mendoza (Profesor Correferente Interno, UTFSM)
Dr. Miguel Carrasco (Profesor Correferente Externo, UAI)

Enero, 2014

UNIVERSIDAD TECNICA FEDERICO SANTA MARIA
Departamento de Informática
Valparaíso - Chile

Título de la tesis

Concatenación de representaciones compactas de imágenes en el problema de recuperación de imágenes

Autor
Tomás Mardones Latham

Tesis presentada como requerimiento parcial para optar al grado académico de **Magíster en Ciencias de la Ingeniería Informática** y al título profesional de **Ingeniería Civil Informática** de la Universidad Técnica Federico Santa María.

Dr. Héctor Allende O.
(Guía) _____

Dr. Marcelo Mendoza
(Evaluador interno) _____

Dr. Miguel Carrasco
(Evaluador externo) _____

Enero, 2014

a Susana

Agradecimientos

Quisiera agradecer de forma sincera al Dr. Héctor Allende, quien siempre me ha tenido paciencia y ha facilitado el camino para poder dedicarme a realizar esta tesis y otros trabajos de investigación. Su guía ha sido de gran ayuda para avanzar y su confianza en mis capacidades un estímulo para desarrollar mis propias habilidades. También quisiera agradecer la ayuda del Dr. Claudio Moraga, quien durante el último año fue de gran ayuda mediante una retroalimentación continua, lo cual me motivaba a seguir adelante.

De forma personal, quisiera agradecer el apoyo de mi familia y amigos, quienes siempre me han apoyado a perseguir estas metas. De forma especial, agradezco a mi esposa, Susana, a quien este trabajo está dedicado, por su amor, apoyo y presión.

También quisiera agradecer el apoyo de las becas que me han sido otorgadas por CONICYT y la Dirección General de Investigación de Postgrado de la Universidad Técnica Federico Santa María.

Abstract

Image related services like Flickr and social networks massification, added to the low cost and increasing quality of digital cameras, have contributed to the exponential growth of the online available images quantity. This has contributed to the need of researching new algorithms capable of working with millions of images. Some applications requiring this class of algorithm are: Image Classification, Image Retrieval, Scene Recognition, Landmark Recognition, Object Recognition and medical and satellital images analysis.

Particularly, when working the problem of retrieving the most relevant images, out of a very big database, given an input image, it is very important to keep the image representation compact. Memory is a limited resource and we must deal with one million images or more. Likewise, reduced response times and acceptable precision are required. In the last decade, the Bag of Features image representation has been the most popular method to solve this difficult task, but in the last few years, its extension called Fisher Vector has show good results at large-scale image retrieval.

A research niche that began to receive attention in the last years is the features combination in image retrieval systems. For most difficult problems in computer science a perfect solution does not exist, but it is common that different solutions tackle distinct parts of the problem. In this work special attention is given to feature sampling techniques used and show how this is relevant to the performance of a image retrieval system. In particular, we present a way to combine Fisher Vectors based on differently sampled descriptors and how does this contribute to enhance the precision maintaining the memory usage. Experiments on different popular image retrieval databases verified this improvement.

Keywords: Image Retrieval, Compact Image Representation, Results fusion, Fisher Vector

Resumen

La masificación de las redes sociales y servicios como Flickr, sumado al bajo costo y creciente calidad de las cámaras digitales, han contribuido al crecimiento exponencial de la cantidad de imágenes disponibles en internet; por lo que en los últimos años se le ha dado mayor énfasis al desarrollo de algoritmos capaces de tratar con millones de imágenes. Algunas de las aplicaciones que requieren este tipo de algoritmos son: clasificación de imágenes, recuperación de imágenes, reconocimiento de escenas, reconocimiento de hitos geográficos, reconocimiento de objetos y análisis de imágenes médicas y satelitales.

Particularmente, el problema de buscar las imágenes más relevantes respecto a una imagen de entrada tiene la necesidad de representar las imágenes de forma compacta, puesto que la memoria es un recurso limitado y se puede estar tratando con más de un millón de imágenes. Igualmente se requieren tiempos de respuesta reducidos y una precisión aceptable en los resultados. Combinar estos requerimientos es una tarea difícil que ha encontrado en los últimos años una respuesta en la representación de la imagen a través de la Bolsa de Características y recientemente en el Vector de Fisher.

Un nicho que recién ha comenzado a ser explotado en los últimos años es el de combinación de tipos de características. En la mayoría de los problemas difíciles de la computación no existe un método capaz de resolver el problema en su totalidad, pero es frecuente que diferentes métodos se especialicen en porciones del problema. En este trabajo se le da especial atención al muestreo de éstas características en la imagen y se muestra que este proceso es relevante en el rendimiento de un sistema de recuperación de imágenes. De forma particular, se expondrá cómo combinar Vectores de Fisher obtenidos mediante descriptores muestreados de formas diferentes y cómo esto contribuye a mejorar la precisión para un mismo uso de memoria. Los experimentos realizados verifican los resultados en las bases de datos más utilizadas del área.

Palabras clave: Recuperación de Imágenes, Representación compacta de imágenes, Combinación de resultados, Vector de Fisher

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

La visión por computador es un campo que incluye métodos para adquirir, procesar, analizar y comprender imágenes y, en general, datos de alta dimensión del mundo real, con el propósito de producir información numérica o simbólica [Stockman y Shapiro, 2001]. Las aplicaciones pueden ir desde sistemas de visión de máquina que, por ejemplo, podrían inspeccionar circuitos para acelerar la línea de producción de una fábrica, hasta la elaboración de sistemas que ayuden a computadores y robots a comprender el mundo alrededor de ellos.

Dentro del campo de visión por computador existe el problema de reconocimiento de instancias. Este trata de reconocer, en una imagen, un objeto 2D o 3D (rígido generalmente) conocido de forma previa, posiblemente visto desde un nuevo punto de vista, contra un fondo atiborrado y/o oclusiones parciales.

A lo largo de los años, muchos algoritmos han sido usados para el reconocimiento de instancias. Los primeros enfoques, mayoritariamente de los años 90 y principios de la década pasada, son resumidos en el estudio de Mundy [Mundy, 2006], donde el enfoque estaba en extraer líneas, contornos o superficies 3D de imágenes, buscando correspondencias con modelos 3D conocidos. Una tendencia posterior, vista en trabajos como [Ferrari et al., 2006; Gordon y Lowe, 2006; Lazebnik et al., 2006; Lowe, 2004; Rothganger et al., 2006; Sivic y Zisserman, 2009], tiende a usar características 2D invariantes al punto de vista. Después de extraer estas características dispersas en la nueva imagen y en la imagen en la base de datos, las características son comparadas contra la base de datos de objetos. Cuando se encuentra un número suficiente de correspondencias de características, se procede a verificar mediante la búsqueda de una transformación geométrica que alínie ambos conjuntos de puntos.

A medida que el número de objetos en la base de datos aumenta, el tiempo que toma encontrar la correspondencia de una imagen de entrada en la base de datos se vuelve prohibitivo. La tarea de encontrar rápidamente correspondencias parciales, es un tema ampliamente tratado en el campo de la recuperación de información (IR). Sivic y Zissermann [Sivic y Zisserman, 2009] fueron los primeros en adaptar técnicas de IR a la búsqueda visual. A partir de este punto, surgió y sigue en constante desarrollo el área de recuperación de imágenes. Este

problema, en el caso particular donde la entrada es una imagen, se denomina recuperación de imágenes basada en contenido (para diferenciarse de enfoques que buscan imágenes utilizando información de texto o de cualquier otro tipo), se puede definir informalmente de la siguiente forma: dada una imagen *común y corriente* de un objeto o localidad, el objetivo es asociar el objeto o localidad a una representación del mismo existente en una base de datos. Cabe mencionar que el término *recuperación*, en este contexto, está asociado a la acción de encontrar elementos de similaridad en una base de datos, de ninguna manera a restaurar.

1.1. Motivación

En los últimos años, por causa de ciertos avances tecnológicos, como la mejora continua y reducción de costos de las cámaras digitales (incluyendo las que se encuentran en los teléfonos celulares), se ha producido un aumento explosivo en el número de imágenes de calidad disponibles libremente en Internet. Por esta razón, se ha planteado la necesidad de crear nuevos métodos en el campo de recuperación de imágenes que puedan lidiar con millones de imágenes.

La aplicación más conocida en que se utiliza recuperación de imágenes es el buscador de imágenes de Google. Si bien, normalmente se realiza la búsqueda basándose en texto, también es posible utilizar solamente imágenes. Relacionado con esto, los algoritmos de búsqueda de imágenes basados en contenido, son útiles para etiquetar imágenes, permitiendo al usuario realizar la típica búsqueda utilizando texto.

Existen también aplicaciones para nichos más específicos. Artfinder es una aplicación para los dispositivos móviles iPhone y iPad, que permite al usuario fotografiar una pintura para reconocerla y entregarle información relevante asociada. Vuforia y Aurasma son plataformas para el desarrollo de aplicaciones que emplean realidad aumentada. Un elemento básico de la realidad aumentada es encontrar un patrón determinado en una imagen y estimar sus coordenadas tridimensionales dentro de la misma, permitiendo la integración de elementos virtuales en esta de forma realista. Hasta hace poco tiempo atrás, el número de patrones que podía utilizar este tipo de sistemas estaba en el orden de decenas. La integración de métodos de recuperación de imágenes para encontrar correspondencias entre la imagen de entrada y las de una base de datos, permitió aumentar el número a más de 1 millón.

1.2. Recuperación de imágenes en grandes bases de datos

Existen dos grupos de métodos que se diferencian en la forma de extraer características de las imágenes. El primer grupo utiliza descripciones globales de la imagen, mientras que el segundo extrae características locales y posteriormente las agrega para obtener una representación compacta de la imagen completa. Este trabajo se enfoca en el segundo tipo.

El problema de recuperación de imágenes es un problema análogo al de los K vecinos más cercanos, donde se buscan las K imágenes más similares a una imagen de entrada. Al llevar este problema a una gran escala se presentan las siguientes dificultades:

- *Tiempo de respuesta*. Muchas de las aplicaciones en el campo de recuperación de información requieren que el tiempo máximo para recibir una respuesta a una consulta sea de pocos segundos. Esto hace imprescindible que la información de los métodos a utilizar se encuentre en memoria de acceso rápido (memoria RAM).

- *Uso de memoria*. Para recuperar una imagen debe existir información suficiente para identificar cada una de forma unívoca. La propia imagen es la mayor fuente de información a la que se puede aspirar, pero su almacenamiento en memoria RAM resulta impracticable. Por ejemplo, si se cuentan con 100 millones de imágenes y se tienen 32GB de memoria RAM a disposición, la representación de cada imagen podría tener un tamaño máximo de 300 bytes aproximadamente sin tener en cuenta otros usos de la memoria.

- *Precisión*. Al buscar cierto objeto o localidad en una base de datos, hay que tener en consideración que la imagen de entrada diferirá de las imágenes en el sistema, por lo que las representaciones compactas deben ser robustas ante cambios de intensidad de la luz y punto de vista, ruido y otras transformaciones.

El método más popular para tratar estos problemas corresponde al llamado Bolsa de Características [Sivic y Zisserman, 2009] (Bag of Features o BoF), el cual está inspirado en el método tradicional, del campo de recuperación de información, Bolsa de Palabras (Bag of Words). La representación del Vector de Fisher (FV), usada extensivamente en esta tesis, puede verse como una extensión de BoF que utiliza estadísticas de mayor orden [Jégou et al., 2012]. Ambos métodos extraen un conjunto de tamaño variable de características de bajo nivel de una imagen y las transforman en una representación de alto nivel de largo fijo (firma o signature de la imagen). El Vector de Fisher, al codificar estadísticas de mayor orden, supera a BoF en cuanto a precisión (para un mismo uso de memoria), memoria (para una misma precisión) y tiempos de respuesta, por lo que BoF se considerará de ahora en adelante solamente para propósitos explicativos.

1.3. Propuesta

En trabajos de Douze y Perronnin [Douze et al., 2011; Perronnin et al., 2010b] se han realizado combinaciones de representaciones con el objeto de incrementar la precisión de un sistema de recuperación de imágenes. En particular, utilizan características muestreadas en la imagen llevada a un espacio de intensidades mediante una heurística para generar una representación, y características muestreadas de forma densa en la imagen en colores para obtener una segunda representación. Al unir estas, la precisión aumenta significativamente, hecho que es atribuido a la información de color que aporta la segunda representación.

Rechazando esta última aseveración, y basándome en el hecho de que los espacios característicos son diferentes según el tipo de muestreo usado, en esta tesis se propone un método para combinar representaciones de imágenes basadas en diferentes técnicas de muestreo, aumentando la precisión, y conservando el uso de memoria y el tiempo de respuesta.

En los experimentos realizados se puede apreciar un aumento de desempeño importante, al comparar el método propuesto con otros enfoques del estado del arte que utilizan una igual cantidad de memoria.

1.4. Contribuciones

Este trabajo aporta a la literatura de recuperación de imágenes, mediante las siguientes contribuciones:

- Un nuevo método capaz de mejorar la precisión de un sistema de recuperación de imágenes, conservando el uso de memoria y el tiempo de respuesta.
- Un estudio comparativo del desempeño del algoritmo propuesto y otros métodos relacionados.

Algunos de los resultados de esta tesis se encuentran publicados en [Mardones et al., 2013].

1.5. Resumen del capítulo

En este capítulo se introdujo el problema de recuperación de imágenes, exponiendo algunas aplicaciones del mundo real y presentando sus componentes de forma resumida. Se describió brevemente la propuesta y se indicaron las contribuciones realizadas a la literatura.

Este primer capítulo cubrió una introducción del problema y la solución propuesta. El capítulo siguiente se enfocará en describir los conceptos y métodos utilizados ampliamente en el área de recuperación de imágenes, para que el lector pueda comprender de mejor forma temas tratados en la tesis más adelante. En el capítulo 3 se estudiarán los métodos más relevantes actualmente para el problema de recuperación de imágenes. En el capítulo 4 se describirá el método propuesto junto a la metodología seguida para llevar a cabo el estudio experimental. En el capítulo 5 se reportarán y analizarán los resultados del estudio comparativo y se discutirán sus implicaciones. Finalmente, en el capítulo 6 se resumirá el contenido del trabajo y se procederá a discutir de forma general los resultados y realizar las observaciones finales.

Capítulo 2

Conocimiento previo

En este capítulo se describirán el problema de recuperación de imágenes y técnicas, usadas habitualmente en la literatura del área, que se utilizarán a lo largo del trabajo. En su mayoría son métodos usados en visión por computador en la *descripción de imágenes* y *métodos de agrupación* (clustering). Posteriormente se describirá la medida de desempeño utilizada generalmente en los sistemas de recuperación de imágenes.

El problema de la recuperación de imágenes

El problema de recuperación de imágenes consiste en encontrar el conjunto ordenado de las imágenes más similares a una *imagen de entrada*, también conocida como *imagen de la consulta*. Es fácil apreciar que este problema es isomorfo al de la búsqueda del vecino más cercano. Específicamente, si definimos el espacio métrico $M = (D, d)$ definido para un dominio de objetos D y una función de distancia d, una consulta en la base de datos para encontrar los k vecinos más cercanos a la representación de la imagen $i \in D$ se puede definir como:

$$kNN(i) = \{R \subseteq X, |R| = k \wedge \forall x \in R, y \in X - R : d(i, x) \leq d(i, y)\}, \tag{2.1}$$

donde $X \subseteq D$ [Zezula et al., 2006].

Por la alta dimensionalidad y cantidad de imágenes, el problema resultaría intratable sin un pre-procesamiento de los datos. Muchas áreas de investigación de las ciencias de la computación y de la matemática están involucradas en las técnicas desarrolladas para realizar esta búsqueda de la forma más eficiente y efectiva posible, tales como algoritmos de agrupamiento, extracción de características, aprendizaje supervisado, no supervisado, reducción de dimensionalidad, aprendizaje de métricas, recuperación de información, codificación e indexación de datos, compresión y búsqueda de los vecinos más cercanos entre otros. A continuación se describirán algunos de los métodos más utilizados para llegar a representar imágenes de forma compacta.

2.1. Pre-procesamiento de la imagen

En la literatura de recuperación de imágenes, es muy frecuente que los algoritmos trabajen la imagen en un espacio de intensidades, dado que disminuye de forma importante la cantidad de cálculos a realizar (y por lo tanto, los tiempos de respuesta y entrenamiento), manteniendo niveles de precisión relativamente similares. Además, en términos prácticos, facilita la comparación con otros trabajos. A continuación se describe este proceso para una imagen típica en formato RGB.

Sea I_{RGB} una imagen de pixeles de 24 bits (8 por color) bidimensional, captada en el espectro de lo visible, de ancho y altura variables. Se puede caracterizar un vector

$$I_{RGB} \in \mathbb{R}^{ancho*altura*3}. \tag{2.2}$$

A su vez, haciendo abuso de notación, sea $I_{RGB}(i, j, k)$ una función que apunte al color (k) del píxel con ubicación (i, j) en I_{RGB}.

A continuación se pre-procesa I_{RGB} con el objetivo de obtener una imagen I de intensidades normalizada. La imagen de intensidades I_{int} (sin normalizar) se obtiene típicamente como

$$I_{int}(i, j) = I_{RGB}(i, j, rojo) * 0{,}3 + I_{RGB}(i, j, verde) * 0{,}59 + I_{RGB}(i, j, azul) * 0{,}11, \tag{2.3}$$

donde las constantes corresponden a los factores de ponderación de cada componente del color según la sensibilidad del ojo humano al mismo. La normalización se realiza como

$$I(i, j) = \frac{I_{int}(i, j)}{max(I_{int})}, i = 1, ..., ancho, j = 1, ..., altura. \tag{2.4}$$

2.2. Detección de regiones de interés

Una región de interés (ROI) corresponde a un conjunto de pixeles de I, generalmente, encerrados en una región convexa. Los detectores de estas regiones están diseñados para que sean altamente repetitivas, es decir, que sea robusta a transformaciones de escala, luminosidad y punto de vista. Esto último con el objetivo de que entre imágenes similares, se puedan encontrar correspondencias que permitan estimar la similaridad entre dos imágenes.

Uno de los problemas principales que resuelve el uso de detectores de ROI corresponde al problema de apertura. En la figura 2.1 se muestra que es mucho más fácil alinear ROIs con gradientes en al menos dos orientaciones significativamente diferentes, por la cantidad de posibles correspondencias.

Para obtener invariancia ante cambios de escala, se suele procesar la imagen en el **espacio de escalas** [Szeliski, 2011]. En la práctica, esto implica que la imagen se lleva a múltiples resoluciones, donde se extraen características en cada una de estas imágenes. Posteriormente, con el objetivo de no repetir las mismas características en diferentes escalas, en cada región de la imagen se buscan características que sean estables en cuanto a posición *y escala*, por lo que se selecciona una característica por localidad en el espacio de escala.

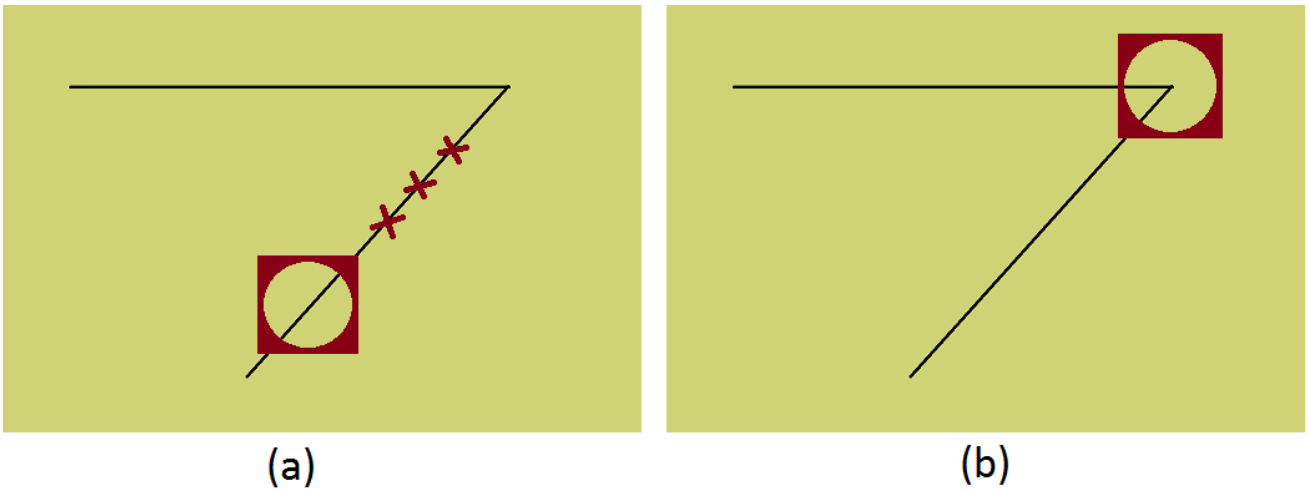


Figura 2.1: (a) Región de apertura y cruces marcan algunas regiones similares en el esquema. (b) Región de apertura única.

2.2.1. Detector Hessiano afín

El detector Hessiano afín es el detector de regiones de interés utilizado en esta tesis [Mikolajczyk et al., 2005]. Primero, en múltiples escalas, se usa un detector basado en la matriz Hessiana,:

$$\begin{aligned} H(\mathbf{x}, \sigma_D) &= \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \\ &= \begin{bmatrix} I_x^2(\mathbf{x}, \sigma_D) & I_x I_y(\mathbf{x}, \sigma_D) \\ I_x I_y(\mathbf{x}, \sigma_D) & I_y^2(\mathbf{x}, \sigma_D) \end{bmatrix} \end{aligned} \tag{2.5}$$

donde $\mathbf{x}$ corresponde a los pixeles que conforman la vecindad de un punto y σ_D corresponde a la escala de un kernel Gaussiano centrado en la $\mathbf{x}$ utilizado para suavizar. Las segundas derivadas utilizadas en esta matriz dan fuertes respuestas ante manchas (blobs) y cordoncillos (ridges). En este contexto, una mancha se refiere a una región de una imagen que difiere en propiedades, tales como el brillo o color, comparado a otras regiones vecinas, mientras que los cordoncillos corresponden a estructuras alargadas, como podrían ser caminos en una imagen satelital o vasos sanguíneos en una imagen biomédica. Un máximo local indica la presencia de una estructura de mancha. La selección de escala se realiza mediante un operador Laplaciano de Gaussianas, el cual entrega la posición y escala en que hay una máxima en la ROI. Dados estos puntos iniciales, se procede a estimar iterativamente regiones elipticas afines [Lindeberg y Garding, 1997]. Los valores propios del segundo momento de la matriz son usados para determinar la forma afin de la elipse. Se busca la transformación que proyecte la forma afín a una con los mismos valores propios. La estimación de la forma afín puede ser aplicada a cualquier punto inicial, siempre el determinante del segundo momento de la matriz sea mayor a cero y la razón señal ruido sea insignificante para este punto. En la figura 2.2 se aprecian algunas regiones de interés detectadas.

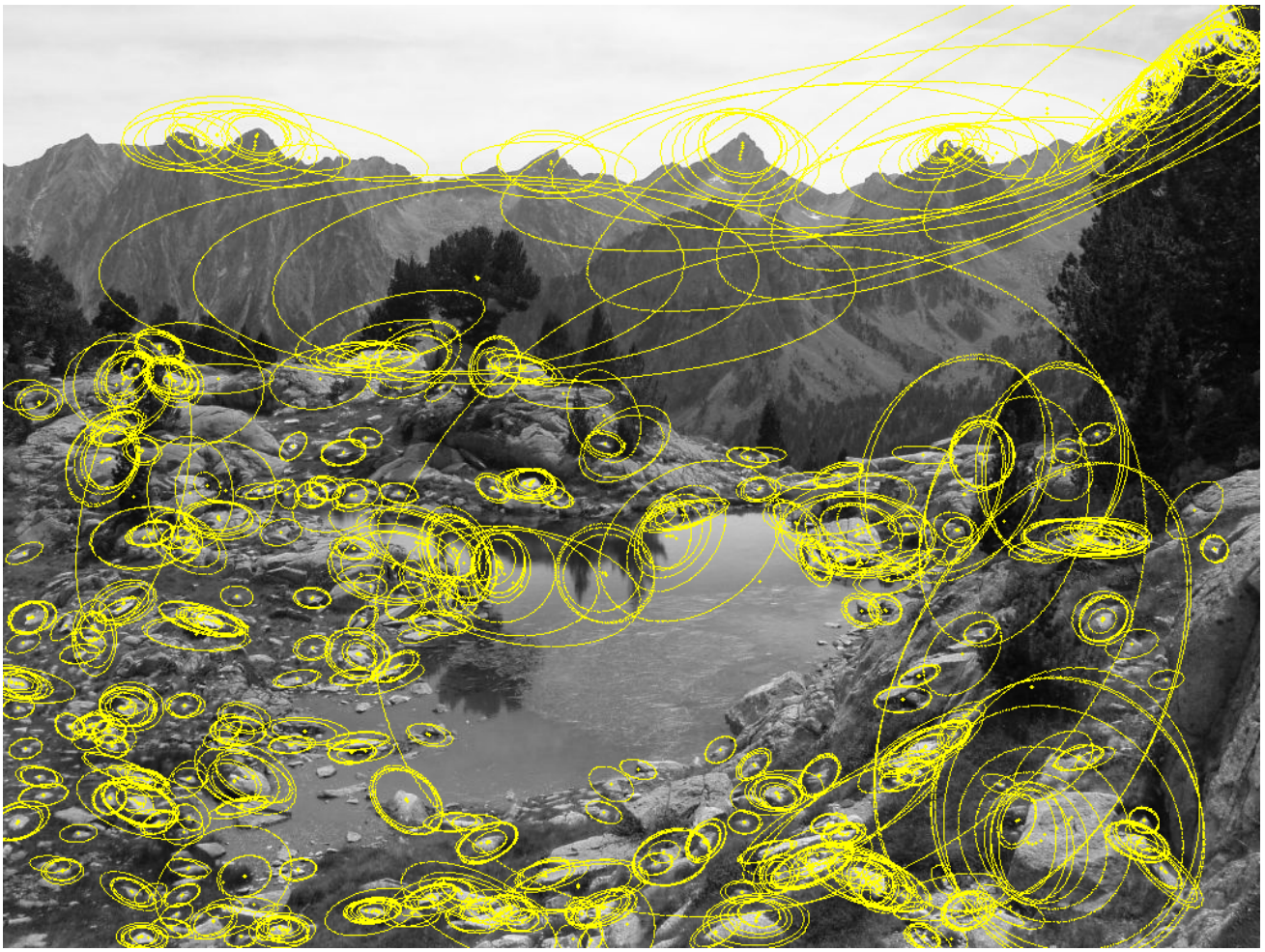


Figura 2.2: Regiones hessianas afín detectadas en una imagen del conjunto de datos Holidays.

2.2.2. Muestreo denso o sistemático de regiones de interés

Una alternativa al uso de detectores de regiones de interés es simplemente seleccionar regiones de interés a lo largo y ancho de toda la imagen, todas equidistanciadas entre sí. Generalmente se utiliza la imagen en varias resoluciones para obtener regiones de interés con diferentes escalas.

Este tipo de muestreo es popular en problemas de clasificación de imágenes y ha demostrado ser una alternativa viable en algunos trabajos recientes de recuperación de imágenes [Delhumeau et al., 2013].

2.3. Descripción de regiones de interés

Después de detectar las regiones de interés o puntos clave (*keypoints*), hay que poder encontrar correspondencias entre sí, es decir, determinar cuáles características provienen de lugares correspondientes en imágenes diferentes. En la mayoría de los casos, la apariencia

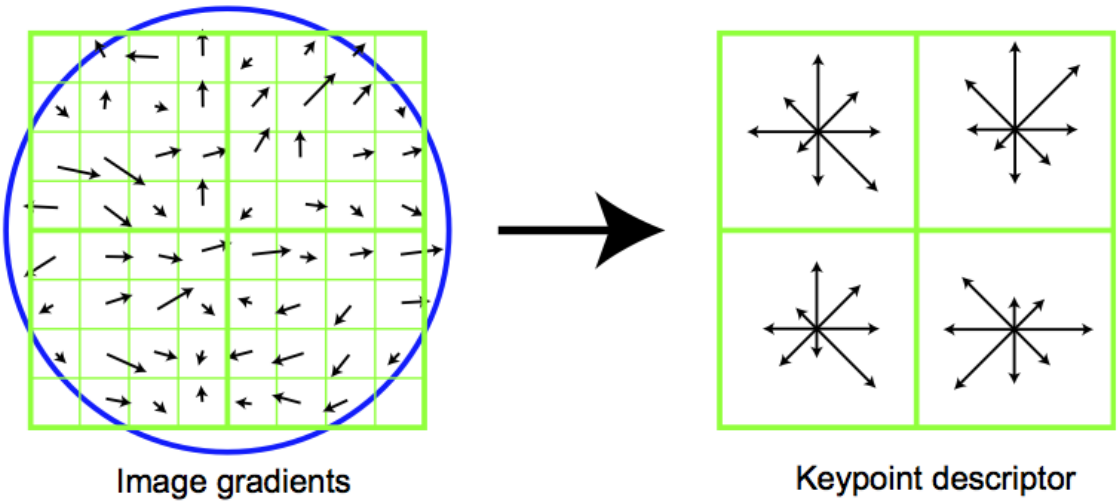


Figura 2.3: A la izquierda: la circunferencia representa el suavizamiento Gaussiano aplicado, las divisiones de la ROI en cuadrantes de 8×8 y las flechas corresponden a los gradientes. A la derecha: se aprecia la agrupación de los cuadrantes y la formación de histogramas las magnitudes de los gradientes agrupados según sus orientaciones. Cada componente de los histogramas corresponde a un número real en el descriptor SIFT. Las imágenes provienen originalmente de Lowe [2004].

local de las características cambian en orientación, escala y transformaciones afines. Generalmente, para evadir estos problemas se utilizan detectores de regiones de interés invariantes ante estas transformaciones, como el descrito anteriormente. Incluso después de compensar por estos efectos, es normal que las apariencias de las regiones de interés varíen de imagen en imagen. Las descripciones de las ROI o *descriptores* tienen la tarea de mantener la invariancia ante estas transformaciones, conservando el poder discriminador entre regiones no correspondientes.

Este trabajo se concentrará en el descriptor SIFT [Lowe, 2004], puesto que indiscutidamente es el más popular en el campo de recuperación de imágenes, facilitando las comparaciones con otros trabajos.

Este método fue inspirado por el trabajo no publicado de Edelman, Intrator y Poggio (1997). En el modelo biológico de la visión, las neuronas responden a un gradiente con una orientación y frecuencia espacial particulares, pero la localización de la gradiente en la retina tiene permitido desplazarse a lo largo de un pequeño campo receptivo, en lugar de estar localizada de forma precisa. Al realizar detallados experimentos de reconocimiento de modelos, tridimensionales de computadora, de objetos y animales, se mostró que al permitir pequeños desplazamientos de los gradientes al buscar correspondencias la clasificación obtuvo muchos mejores resultados bajo rotaciones 3D. Por ejemplo, al rotar en profundidad un conjunto de objetos en 20 grados, la precisión del reconocimiento aumentó de un 35 % para un sistema de correlación de gradientes a un 94 % usando un complejo modelo de celdas.

Para calcular las características SIFT, primero se subdivide la región de interés en 16×16 cuadrantes y se aplica un suavizamiento Gaussiano con un σ correspondiente a 1,5 veces

la escala del punto de interés, centrado en el mismo, obteniendo una región de la imagen suavizada L. A continuación, para cada muestra de la imagen $L(x, y)$ se calculan las gradientes con magnitud $m(x, y)$ y orientación $\theta(x, y)$:

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \tag{2.6}$$

$$\theta(x,y) = tan^{-1}\left(\frac{L(x,y+1) - L(x,y-1)}{L(x+1,y) - L(x-1,y)}\right) \tag{2.7}$$

El objetivo de aplicar este suavizamiento de reducir la influencia de las gradientes más alejadas, puesto que están más expuestas a errores al buscar otras correspondencias. A la izquierda de la Figura 2.3 se puede ver el resultado obtenido hasta este punto.

El paso siguiente consiste en agrupar las gradientes, formando 4×4 nuevos cuadrantes, incluyendo cada uno 16 gradientes, donde en cada uno se construye un histograma de orientaciones de gradientes, sumando el valor del gradiente, ponderado por la Gaussiana, a uno de ocho intervalos del histograma, tal como se gráfica en la Figura 2.3. Cada elemento del histograma corresponde a un componente de una característica SIFT, sumando en total, 16×8 componentes.

El vector de 128 valores no negativos forma al descriptor SIFT. Para reducir los efectos del contraste o ganancia, el vector es normalizado a una longitud unitaria.

2.4. Evaluación de un Sistema de Recuperación de información

El campo de recuperación de información tiene el objetivo de obtener información relevante al realizar una consulta. Este objetivo tiene varios conceptos que conviene precisar para poder crear medidas de desempeño bien formuladas.

El enfoque comúnmente utilizado en la evaluación de un sistema de IR utiliza la noción de documentos relevantes y no relevantes [Manning et al., 2008]. Dada una necesidad de información, un documento de una colección de prueba debe ser clasificado como relevante o no relevante. Esta decisión corresponde al juicio de relevancia con el cual se construye el estándar dorado (gold standard) o verdad fundamental (ground truth).

La relevancia se evalúa según una necesidad de información, no una búsqueda o consulta. Por ejemplo, una necesidad de información puede ser: Ïnformación respecto a si los plátanos son beneficiosos para los deportistas". Una consulta asociada, en el marco de recuperación de texto, podría ser:

$$\text{Plátano } AND \text{ beneficios } AND \text{ deportistas} \tag{2.8}$$

Un documento es relevante si resuelve la necesidad de información planteada, lo cual no pasa simplemente por tener todas las palabras de la consulta. Esta distinción es malentendida a veces en la práctica, porque no siempre la necesidad de información es evidente. En el caso de la evaluación de un sistema de IR, es importante que la consulta esté relacionada de forma evidente con la necesidad de información para evitar ambigüedad. En este trabajo

se trabajarán con decisiones binarias sobre la relevancia, es decir, un documento puede ser relevante o no relevante.

En el caso específico de la recuperación de imágenes basada en contenido, una consulta corresponde a una imagen y la necesidad de información, de forma general, asociada es Ïmágenes de la misma localidad u objeto que los presentes en la imagen de entrada".

2.4.1. Medidas de desempeño

En el campo de recuperación de información, las medidas de desempeño tradicionales como la precisión no logran el objetivo de reflejar el correcto desempeño de un sistema de recuperación de información. Esto se debe a que el número de documentos no relevantes es mucho mayor que el número de documentos relevantes para una necesidad de información específica, existiendo un problema de clases desbalanceadas.

2.4.1.1. Evaluación de resultados no ordenados

Las medidas básicas para medir la efectividad de un sistema IR son la precisión (precision) y exhaustividad (recall). Primero se definirán para el caso en que el sistema, dada una consulta, devuelva un conjunto de documentos no ordenados. Más adelante se extenderá a resultados ordenados.

La precisión (P) es la fracción de los documentos recuperados que son relevantes:

$$Precision = \frac{\#(\text{elementos relevantes recuperados})}{\#(\text{elementos recuperados})}. \tag{2.9}$$

La exhaustividad (R) es la fracción de documentos relevantes que son recuperados:

$$Exhaustividad = \frac{\#(\text{elementos recuperados relevantes})}{\#(\text{elementos relevantes})}. \tag{2.10}$$

La siguiente tabla generaliza un los conceptos en términos de clasificación:

	Relevante	No relevante
Recuperado	verdaderos positivos (tp)	falsos positivos (fp)
No recuperado	falsos negativos (fn)	verdaderos negativos (tn)

Entonces

$$P = tp/tp + fp) \tag{2.11}$$
$$R = tp/(tp + fn) \tag{2.12}$$

Estas medidas son importantes en recuperación de información, puesto que entregan información útil. Éstas se concentran en evaluar la recuperación de verdaderos positivos, preguntando qué porcentaje de documentos relevantes han sido encontrados y cuántos falsos positivos han sido recuperados en el proceso. Tener estas dos medidas también es ventajoso, puesto que hay ocasiones en que una medida es más relevante que la otra. Por ejemplo, al

buscar sitios mediante el buscador de Google, es importante que los resultados de las primeras páginas sean relevantes (alta precisión), pero no interesa revisar exhaustivamente si todos los documentos son relevantes. En el caso de una búsqueda en un disco duro, es más importante la exhaustividad. De todas maneras, ambas medidas se contraponen: es fácil obtener un valor 1 de exhaustividad recuperando todos los documentos, pero la precisión será muy baja. Por otro lado, en un sistema que funcione bien, la precisión generalmente disminuye al aumentar el número de documentos que se está recuperando. En general, se quiere tener un grado de exhaustividad, siempre tolerando hasta cierto porcentaje de falsos positivos.

Una medida que une y pondera la precisión y la exhaustividad es la medida F, que es la media armónica ponderada de la precisión y la exhaustividad:

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R}, \text{ donde } \beta^2 = \frac{1-\ \alpha}{\alpha} \tag{2.13}$$

donde $\alpha \in [0, 1]$ y por lo tanto $\beta^2 \in [0, \text{inf}]$.

2.4.1.2. Evaluación de resultados ordenados

Las medidas de desempeño recién vistas se basan en grupos: no hacen ninguna distinción en el orden en que se entregan los resultados. Los resultados de las búsquedas, en los sistemas de recuperación de información normalmente usados, utilizan resultados ordenados según la relevancia. Para observar el rendimiento en función del número de documentos recuperados se pueden obtener conjuntos compuestos por los k documentos más relevantes. Para cada conjunto, se puede graficar una curva de precisión-exhaustividad. Éstas curvas tienen una forma dentada, puesto que si el (k+1)-ésimo documento recuperado no es relevante, entonces la exhaustividad será la misma, pero la precisión menor. Si es relevante, ambas medidas aumentarán. Una forma estándar para suavizar esta curva corresponde a usar la *precisión interpolada* p_{interp}. Ésta se define para cierto nivel de exhaustividad r como la precisión más alta encontrada en cualquier nivel de exhaustividad $r' \geq r$:

$$p_{interp}(r) = \max_{r' \geq r} p(r') \tag{2.14}$$

La justificación para utilizar esta medida, es que un usuario, la mayoría de las veces, estaría dispuesto a ver unos pocos documentos adicionales si eso incrementase el porcentaje de documentos relevantes recuperados.

La curva de precisión-exhaustividad resulta bastante informativa, pero en algunos casos, para simplificar el proceso de comparación, es deseable resumirla usando unos pocos números o inclusive solamente uno. La medida utilizada mayoritariamente en el campo de recuperación de imágenes corresponde a la *precisión media promedio* (MAP - Mean Average Precision), la cual corresponde a un número que refleja la calidad en todos los niveles de exhaustividad. Entre las diferentes medidas de evaluación, MAP ha mostrado ser especialmente discriminativa y estable [Manning et al., 2008]. Para una necesidad de información, la precisión media corresponde al valor medio de la precisión obtenido al recuperar los documentos hasta el k-ésimo documento relevante; este valor se promedia usando cada necesidad de información. Esto es, sea el conjunto de documentos relevantes para una necesidad de información $q_j \in Q$ $d_1, ..., d_{m_j}$

y R_{jk}el conjunto de resultados de la recuperación ordenados desde el mejor resultado hasta llegar al documento d_k, entonces

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} Precision(R_{jk}). \tag{2.15}$$

Cuando un documento relevante no es recuperado, el valor de la precisión en la ecuación de arriba se toma como 0. Para una necesidad de información individual, la precisión media aproxima el área bajo la curva de precisión-exhaustividad no interpolada, por lo que el MAP sería una estimación del área promedio bajo la curva de precisión-exhaustividad para un conjunto de consultas.

Cabe mencionar que la curva de precisión-exhaustividad es bastante similar a la curva ROC que grafica la taza de verdaderos positivos (la sensibilidad) contra la taza de falsos positivos (1 - especificidad), siendo la sensibilidad otro nombre para la exhaustividad. La taza de falsos positivos está dada por $fp/(fp + tn)$. Una medida de agregación típica de la curva ROC corresponde al área bajo la curva, la cual es análogo al MAP.

2.5. Resumen del capítulo

En este capítulo se explicó de forma precisa la naturaleza del problema a tratar y se revisaron algunas de las técnicas más utilizadas en el campo de recuperación de imágenes con el objetivo de posteriormente exponer adecuadamente la tesis que plantea este documento. Se le dio un especial énfasis a las medidas de evaluación, dado que son de especial importancia en un campo donde el desempeño de los algoritmos se evalúa de forma empírica.

Capítulo 3

Estado del Arte

En los últimos años en el campo de recuperación de imágenes se ha utilizado un reducido número de representaciones de imagen, destacando principalmente el método denominado Bolsa de Características, proveniente del área de recuperación de texto. Sólo en el último par de años ha comenzado a ganar terreno la representación del Vector de Fisher, propuesta el año 2007 por Florent Perronnin [Perronnin y Dance, 2007]. En este capítulo se comenzará describiendo las características de los datos y sus respectivos conjuntos de prueba, utilizados de forma estándar para comparar el desempeño de los algoritmos. Posteriormente se describirá la representación de Bolsa de Características, dada su simplicidad, algunas de sus variantes y se terminará con la exposición de las bases teóricas del Vector de Fisher, su implementación utilizando mezclas de Gaussianas y una reseña de las variantes más relevantes para este trabajo.

3.1. Características de los datos

El dato básico con el cual se trabaja corresponden a imágenes digitales captadas por cámaras digitales ópticas monoculares que utilizan CCD o CMOS para captar el espectro visible de la luz. Las imágenes corresponden a paisajes naturales, entre otros, se pueden encontrar en las bases de datos de prueba paisajes de la naturaleza, construcciones humanas, animales, espacios interiores o exteriores o incluso bajo el agua. Dependiendo de la base de datos, existe un sesgo en el contenido de las imágenes, el cual se entenderá con las descripciones de las mismas en la siguiente sección. Otro tipo de características variables en las bases de datos son la resolución de las imágenes y las cámaras empleadas.

3.2. Bases de datos de prueba

Las bases de datos que se utilizan para estudiar el problema de recuperación de imágenes se componen de dos conjuntos de imágenes no exclusivos. El primero consiste en imágenes de consulta, es decir, las imágenes que recibe el sistema como entrada y el segundo en imágenes relevantes y no relevantes respecto a cada una de las imágenes de entrada. La necesidad de información implícita en las consultas corresponde a "imágenes que correspondan a la misma escena o que contenga los mismos objetos de la imagen de entrada". Las imágenes son etiquetadas manualmente.

La precisión de las técnicas es medida usando MAP [Manning et al., 2008], que fue introducida en la sección 2.4.1.

Holidays corresponde a una base de datos constituida principalmente por imágenes tomadas durante vacaciones por miembros del INRIA [Jégou et al., 2008], pero también incluye imágenes tomadas con el propósito de probar la robustez ante diversas transformaciones, tales como rotaciones, cambios de iluminación y puntos de vista, desenfoque, etc. La base de datos incluye imágenes de escenas variadas (naturales, hechas por el hombre, efectos de agua y fuego, etc) y las imágenes son de alta resolución. La base de datos de imágenes (BDI) contiene 500 grupos de imágenes, donde cada uno representa una escena diferente. La primera imagen de cada grupo corresponde a la imagen de entrada y los resultados correctos de búsqueda son las otras imágenes del grupo.

University of Kentucky Recognition Benchmark (UKB) [Nister y Stewenius, 2006] es una base de datos que contiene imágenes de 2550 objetos y escenas. Cada uno de éstos está representado por cuatro imágenes tomadas bajo diferentes puntos de vista. Para esta base de datos, los autores proponen usar una medida de desempeño llamada KS (Kentucky Score): es el número promedio de imágenes relevantes posicionadas en las primeras cuatro posiciones cuando se realiza una búsqueda en la BDI. Cabe mencionar que existen plataformas de reconocimiento de imágenes con estas características en el campo de la realidad aumentada. En estos sistemas, el usuario sube una imagen a ser reconocida a un servidor y posteriormente el sistema debe ser capaz de reconocer esta imagen desde diferentes puntos de vista, generalmente bajo transformaciones aproximadamente afines.

Oxford Buildings Dataset (o Oxford5k) [Philbin et al., 2007] es conjunto de 5062 imágenes de 11 localidades particulares de Oxford, cada localidad tiene 5 posibles imágenes de entrada asociadas, lo que constituye un total de 55 búsquedas. El desempeño se mide usando MAP. Esta base de datos se diferencia principalmente en dos puntos respecto a UKB y Holidays. Primero, existe un número mucho mayor de imágenes asociadas a una misma consulta. En segundo lugar, las imágenes son predominantemente de edificios, por lo que existe un sesgo en la distribución de los descriptores asociados a estas imágenes. También es pertinente mencionar que existe una variante llamada Oxford105k [Philbin et al., 2007] que incluye las imágenes originales más cien mil imágenes distractoras descargadas de Flickr [Philbin et al., 2007].

En las bases de datos de imágenes disponibles, el número de imágenes no es muy elevado (1491 - 5062), por lo que, para poder probar la capacidad de escalamiento de los algoritmos, se suele agregar un conjunto independiente de imágenes [Jégou et al., 2012]. El conjunto de imágenes independiente utilizado en esta propuesta se conoce como *MIRFLICKR-1M*, el cual

contiene un millón de imágenes extraídas de Flickr [Huiskes et al., 2010].

3.3. Esquema de trabajo de las representaciones de imágenes basadas en agregación de características

Esta tesis se enfocará exclusivamente a representaciones compactas de imágenes basadas en agregación de características de imágenes, por lo que solamente se explicará de forma parcial el funcionamiento de otra clase de técnicas según lo requiera la comprensión del estado del arte.

Las técnicas basadas en agregación de características siguen un esquema de trabajo bastante estándar. Éste se puede resumir en los siguientes cinco pasos (suponiendo una base de datos poblada) [Jégou et al., 2012]:

1. *Recepción y pre-procesamiento de imagen.* Se obtiene una imagen y se transforma en una imagen de intensidades normalizada.

2. *Extracción de características.* Una vez pre-procesada la imagen, se extrae información significativa de ésta, con el objetivo de trabajar con un conjunto reducido de características (descriptores). El resultado es un conjunto de descriptores (vectores de números reales) que caracterizan pequeñas regiones de la imagen.

3. *Agregación de características.* Para problemas de seguimiento o reconocimiento de un número reducido de instancias, basta con usar los descriptores de la etapa anterior para caracterizar la imagen, dado que la memoria no significa un problema, pero al trabajar con miles de imágenes se hace necesario reducir el uso de memoria. Para esto es necesario agregar las características usando una técnica ad-hoc, tales como la Bolsa de Características y el Vector de Fisher.

4. *Búsqueda de los k vecinos más cercanos.* Una vez obtenida la representación de la imagen, se debe realizar la búsqueda en la base de datos para encontrar los k vecinos más cercanos. Para realizar las comparaciones, se normalizan los FV (usando normalización de potencia y L2) y se calcula su producto punto.

5. *Validación.* Finalmente, reducida la lista de candidatos a solo unos pocos, se buscan los mejores resultados por medio de criterios más costosos (no factibles en la etapa anterior) y se vuelve a ordenar la lista de resultados.

Para poblar la base de datos, generalmente se siguen los pasos del 1 al 3 y la representación compacta resultante se inserta en la base de datos.

El objetivo de los primeros dos pasos es obtener descriptores SIFT a partir de la imagen de entrada, tal como fue descrito en la secciones 2.1, 2.2 y 2.3.

A continuación, se describirán las dos técnicas más populares del estado del arte utilizadas para la etapa de agregación de características y sus variantes más relevantes.

3.4. Enfoques basados en Bolsa de Características

El método para agregación de características más popular actualmente, corresponde a la Bolsa de Características (BoF)[Sivic y Zisserman, 2009], el cual es una adaptación del método Bolsa de Palabras, utilizado en el área de recuperación de texto. También se le conoce como Bolsa de Palabras Visuales.

En general, a no ser que se especifique de otra forma, los pasos 1 y 2 del esquema de trabajo se realizan mediante los métodos expuestos en las secciones 2.1, 2.2 y 2.3. A continuación se describirán las implementaciones canónicas de los pasos 3 y 4.

3.4.1. Bolsa de Características

Los descriptores SIFT hacen un buen trabajo representando las imágenes de forma única, pero no es viable guardar los descriptores de cada imagen porque usan una cantidad prohibitiva de memoria. La idea que persigue una BoF es la de construir un conjunto de descriptores ´´estereotipados" llamados típicamente *palabras visuales* con el objetivo de asociar los descriptores de una imagen a una de estas palabras visuales. De esta forma, se representa la imagen mediante estas asociaciones, en lugar de los descriptores.

Las dos primeras preguntas que cabe hacerse es ¿cómo elijo estas palabras visuales? y ¿con qué criterio asocio un descriptor a una palabra visual? La primera pregunta depende de la segunda, por lo que nos enfocaremos en ésta última primero. Un criterio apropiado y aceptado para realizar esta asociación es el de similaridad. Esto implica que cada descriptor se asocia a la palabra visual más similar al mismo. Puesto que los descriptores SIFT (y las palabras visuales, que son descriptores también) son vectores de números reales, la similaridad se puede medir en términos de distancia. En particular se suele usar la distancia euclidiana. También existen otros enfoques que integran otro tipo de información complementaria a la distancia como [Lazebnik et al., 2006; Liu et al., 2008; Morioka y Satoh, 2010] entre otros.

La elección de las palabras visuales es importante para lograr un buen nivel de discriminación entre imágenes. La primera y más utilizada técnica para realizar este proceso, corresponde a k-medias [Lloyd, 1982]. El objetivo de esta técnica es separar N observaciones en K conjuntos, donde cada observación pertenece al *centroide* más cercano. El centroide sirve como una observación prototipo de una agrupación y se calcula iterativamente como la media de las observaciones asignadas al mismo. El algoritmo 1 muestra el funcionamiento básico de esta técnica. Los centroides corresponden a las palabras visuales y los descriptores de una imagen son asociados a la palabra visual más cercana.

Finalmente la Bolsa de Características de una imagen se obtiene creando un histograma de K lotes, donde cada uno muestra la cantidad de descriptores de la imagen que se asignan a una palabra visual específica.

Posterior a la obtención de la representación de la imagen, se procede a la búsqueda del vecino más cercano. Esta tarea no resulta relevante en el contexto de la propuesta, por lo que sólo se describirá brevemente. Como se vio, cada documento es representado por la frecuencia de sus palabras visuales, éstas frecuencias se utilizan para indexar las imágenes permitiendo una búsqueda rápida, pero es importante ponderar las diferentes componentes del vector BoF (el histograma) previamente. El método básico de ponderación usado corresponde a *frecuen-*

Algorithm 1 K-medias

Parámetros de entrada: K número de centroides
Sea D un conjunto de N datos.
Sea C el conjunto de K centroides
$m : D \rightarrow C$ (membresía de un dato en D a un grupo C)
Se inicializan los centroides C_i de manera aleatoria, $i = 1...K$
for cada $j \in \{1...N\}$ **do**
  $m(D_j) = argmin_{i \in \{1...K\}} distancia(D_j, C_i)$
end for
while m haya cambiado **do**
  for cada $i \in 1...K$ **do**
    Recalcular c_i como el centroide de $D|m(D) = i$
  end for
  for cada $j \in 1...N$ **do**
    $m(D_j) = argmin_{i \in \{1...K\}} distancia(D_j, C_i)$
  end for
end while

cia de término - frecuencia inversa de documento o tf-idf [Sivic y Zisserman, 2009]. Este método busca darle una mayor ponderación a las palabras visuales menos frecuentes, dado que normalmente poseen un mayor poder discriminativo, mientras disminuye la ponderación de las palabras visuales más repetidas. Usando esta ponderación, es posible construir un ordenamiento de menor a mayor distancia de las imágenes más similares a la de entrada. La indexación se suele realizar mediante la utilización de archivos invertidos. Esto consiste en asociar a cada palabra visual los índices de las imágenes en la base de datos que la contienen, de esta forma, revisando las palabras visuales existentes en la imagen de entrada, se puede reducir la búsqueda de imágenes a las que comparten características similares.

3.4.2. Avances utilizando Bolsa de Características

Se han realizado diferentes clases de mejoras a la representación que han contribuido a mantener este método vigente y continuo desarrollo. Sin tener en cuenta los trabajos que exploran métodos para aumentar la eficiencia de las búsquedas, existen dos áreas en que se ha concentrado la investigación.

La primera consiste en la incorporación de información espacial [Jégou et al., 2010; Lazebnik et al., 2006; Liu et al., 2008; Morioka y Satoh, 2010]. El problema que ataca esta área consiste en que al representar los descriptores por medio de palabras visuales, toda la información espacial (posición, escala y rotación) que posee cada descriptor es descartada. Las técnicas de ésta área se enfocan en preservar esta información utilizando diferentes estrategias. Van desde la agrupación de descriptores [Liu et al., 2008; Morioka y Satoh, 2010] hasta el aprendizaje de las posiciones absolutas, escalas y ángulos de cada tipo de palabra visual en las imágenes [Jégou et al., 2010; Lazebnik et al., 2006].

La segunda trata la mejora de la asignación de descriptores a palabras visuales [Jégou

et al., 2010; Nister y Stewenius, 2006; Philbin et al., 2008]. Jégou, en Jégou et al. [2010], trata el problema de que utilizar un valor de K muy alto para k-medias (y tener palabras visuales más precisas) resulta muy costoso. Para mitigar este problema, sugiere utilizar k-medias con un K relativamente bajo y agregar una subdivisión de las celdas de Voronoi resultantes agregando a cada descriptor una firma binaria que codifica la posición que utiliza en su respectiva celda. En [Nister y Stewenius, 2006] se propone utilizar k-medias jerárquico, creando una jerarquía de palabras visuales, donde los nodos hojas contienen archivos invertidos con los índices de las imágenes relacionadas, lo cual permite una búsqueda más rápida. En [Philbin et al., 2008], se realiza el cambio de asociar los descriptores a las r palabras visuales más cercanas, evitando el problema de la asignación poco estable de los descriptores ubicados en las fronteras de los diferentes agrupamientos.

Los métodos basados en Bolsa de Características tienen tres ventajas [Jégou et al., 2012]: primero que nada, es una representación que se beneficia de poderosos descriptores locales. En segundo lugar, su representación permite la comparación mediante medidas estándar de distancias. Y por último, al ser una representación de alta dimensionalidad, y siendo los vectores dispersos, es posible usar listas invertidas para una búsqueda eficiente. Sin embargo, hay dos factores que limitan el número de imágenes que pueden ser indexados en la práctica: el tiempo de búsqueda, el cual se vuelve prohibitivo al considerar más de 10 millones de imágenes y el uso de memoria por imagen.

3.5. Enfoques basados en el Vector de Fisher

Los Vectores de Fisher (FVs) corresponden a un método de agregación de descriptores, con algunas características similares a las bolsas de características [Perronnin y Dance, 2007]. Ambos métodos comparten el uso de palabras visuales, pero en el caso del FV se utilizan Gaussianas multivariadas en lugar de los centroides, pero el mayor cambio está en la representación, dado que el FV es mucho más rico en información, puesto que incorpora estadísticas de primer orden, en comparación a la estadística de orden cero que es el histograma de la bolsa de características. Para entrar en los detalles de implementación del Vector de Fisher, primero es necesario exponer el esquema de trabajo del Kernel de Fisher, en el cual se basa este método de agregación.

3.5.1. Esquema de trabajo del Kernel de Fisher

Sea X un conjunto de N descriptores extraídos de una imagen. Supondremos que el proceso de generación de X puede ser modelado por una función de densidad de probabilidad u_λ, independiente de la imagen, con parámetros λ (se abusará de la notación, representando λ los parámetros y al mismo tiempo sus estimaciones). Jaakola y Hausler [Jaakkola y Haussler, 1999] propusieron describir X a través del vector

$$G_\lambda^X = \frac{1}{N}\nabla_\lambda \log u_\lambda(X). \tag{3.1}$$

El gradiente de la log-verosimilitud describe la contribución de los parámetros al proceso de generación de X. Su dimensionalidad solamente depende del número de parámetros en λ

y la dimensionalidad de los elementos de X.

En el problema de recuperación de imágenes, el objetivo es encontrar las imágenes más cercanas entre sí, por lo que es importante contar con un método eficiente para realizar esta operación. Para esto, sea Y un segundo conjunto de descriptores extraídos de una imagen cualquiera y G_λ^Y su representación agregada. Un kernel sugerido en [Jaakkola y Haussler, 1999] que permite realizar la comparación entre G_λ^X y G_λ^Y corresponde al kernel de Fisher:

$$K(X,Y) = G_\lambda^X F_\lambda^{-1} G_\lambda^Y, \tag{3.2}$$

donde F_λ es la matriz de información de Fisher de u_λ:

$$F_\lambda = E_{x\sim u_\lambda}[\nabla_\lambda \log u_\lambda(x) \nabla_\lambda \log u_\lambda(x)^T]. \tag{3.3}$$

Al ser F_λ^{-1} simétrica y positiva definida, se puede realizar una descomposición de Cholesky $F_\lambda^{-1} = L_\lambda^T L_\lambda$. De esta forma, podemos reescribir $K(X,Y)$ como:

$$K(X,Y) = (\mathcal{G}_\lambda^X)(\mathcal{G}_\lambda^Y)^T, \tag{3.4}$$

$$K(X,Y) = G_\lambda^X L_\lambda^T L_\lambda G_\lambda^Y \tag{3.5}$$

$$= (\mathcal{G}_\lambda^X)(\mathcal{G}_\lambda^Y)^T, \tag{3.6}$$

donde

$$\mathcal{G}_\lambda^X = L_\lambda G_\lambda^X. \tag{3.7}$$

La ventaja de esta nueva representación $\mathcal{G}_\lambda^X$ es que al almacenar el vector de esta forma, resulta mucho más eficiente realizar las comparaciones entre diferentes imágenes. Para esto generalmente se utiliza la norma L_2 o un simple producto punto.

A $\mathcal{G}_\lambda^X$ se le conoce como el Vector de Fisher (FV) de X.

3.5.2. Vector de Fisher basado en Mezclas de Gaussianas Multivariadas

Siguiendo el trabajo de Perronnin y Dance [Perronnin y Dance, 2007] se elige que u_λ sea una mezcla de Gaussianas multivariadas (GMM): $u_\lambda(x) = \sum_{i=1}^K w_i u_i(x)$. El conjunto de parámetros λ corresponden a $w_i, \mu_i, \sigma_i, i = 1, ..., K$, donde w_i, μ_i y σ_i son respectivamente el peso, el vector de medias y la matriz de varianza (asumida diagonal) de la Gaussiana u_i. Por otra parte, la dimensión de cada uno de los parámetros depende de D (dimensión de los descriptores), siendo $w_i \in \mathbb{R}$, $\mu_i \in \mathbb{R}^D$ y $\sigma_i \in \mathbb{R}^D$. Los parámetros del modelo u_λ se obtienen durante una fase de entrenamiento offline, usando descriptores de muchas imágenes y el método de estimación de máxima verosimilitud. F_λ se obtiene mediante la aproximación cerrada diagonal de [Perronnin y Dance, 2007], normalizando el gradiente por $L_\lambda = F_\lambda^{-1/2}$, lo cual se traduce en un blanqueamiento de las dimensiones. Cabe mencionar que para la representación a utilizar, sólo se estiman los gradientes respecto a la log-verosimilitud del

vector de medias, dado que para un mismo tamaño del vector, el desempeño es similar si se usan o no las componentes de la varianza y los pesos.

Sea $\gamma_n(i) \in \mathbb{R}$ la asignación suave del descriptor x_n a la i-ésima Gaussiana:

$$\gamma_n(i) = \frac{w_i u_i(x_n)}{\sum_{j=1}^{K} w_j u_j(x_n)}, \tag{3.8}$$

$$\text{con} \sum w_i = 1, w_i \in [0,1], i = 1, ..., K, n = 1, ..., N. \tag{3.9}$$

Sea $\mathcal{G}_i^X$ el gradiente D-dimensional respecto a la log-verosimilitud de la media μ_i de la Gaussiana i. Suponiendo que los x_n son generados de manera iid por u_λ se puede obtener sin grandes dificultades:

$$\mathcal{G}_i^X = \frac{1}{N\sqrt{w_i}} \sum_{n=1}^{N} \gamma_n(i) \sigma_i^{-1}(x_n - \mu_i). \tag{3.10}$$

El vector final $\mathcal{G}_\lambda^X$ se obtiene concatenando los vectores $\mathcal{G}_i^X$ para i=1,...,K, siendo entonces de dimensión final KD. Los valores de K, normalmente, van del rango de $K = 16$ a $K = 256$.

Después de obtener el vector, se procede a normalizarlo, primero utilizando una normalización de potencia

$$f(z) = sign(z)|z|^\alpha, \tag{3.11}$$

con $0 \leq \alpha \leq 1$. A continuación se normaliza usando la norma L_2. En [Jégou et al., 2012] se argumenta que la normalización de potencia ayuda a reducir el impacto de múltiples correspondencias y explosiones visuales. Una explosión visual es la propiedad de que un elemento visual dado aparezca más veces en una imagen de lo que un modelo estadístico independiente podría predecir, lo que corrompe las medidas de similaridad visual [Jégou et al., 2009].

Adicionalmente, se recomienda en [Jégou et al., 2012] reducir la dimensionalidad de los descriptores de la imagen mediante Análisis de componentes principales (PCA) con el objetivo de descorrelacionar los datos, puesto que de esa forma el modelo GMM con matrices de covarianza diagonales puede ajustarse mejor. PCA construye una transformación lineal que proyecta los datos a un nuevo sistema de coordenadas, de tal forma que la varianza de mayor tamaño se encuentre a lo largo del primer eje, la segunda más grande en el segundo eje y así sucesivamente.

Los Vectores de Fisher son vectores densos de gran dimensionalidad, pero se ha mostrado en numerosos trabajos que su dimensión se puede reducir y aún conservar gran parte de la información [Jégou et al., 2012; Perronnin et al., 2010a]. Varios autores han enfocado sus trabajos [Gong et al., 2012; Gordo et al., 2012; Jégou et al., 2011] en crear y aplicar métodos para reducir la dimensionalidad o indexación en el campo de recuperación de imágenes utilizando Vectores de Fisher. Inicialmente, en este trabajo, se utilizará PCA para reducir la dimensionalidad de los FVs y se omitirá el proceso de indexación, puesto que los tiempos de respuesta siguen siendo reducidos y resulta más fácil de comparar con otros métodos al haber menos procesos involucrados.

3.5.3. Estado del arte de métodos relevantes para la propuesta que utilizan el Vector de Fisher

En esta sección se revisan algunos de los trabajos que tienen mayor relación con la propuesta.

En Douze et al. [2011] se adapta el uso de atributos aprendidos mediante clasificadores al problema de recuperación de imágenes. En particular, se combinan dos tipos de representaciones de imágenes: la primera un Vector de Fisher y la segunda un vector de puntajes de atributos aprendidos mediante una máquina de soporte vectorial Bishop [2006] por cada atributo. Estos atributos corresponden a 2659 conceptos y objetos obtenidos de la base de datos "Large Scale Concept Ontology for Multimedia"Naphade et al. [2006]. Las imágenes asociadas a estos conceptos (disponibles en la misma base de datos) se utilizan para entrenar las SVM. Específicamente, los datos que se extraen de las imágenes, para usarlos como entrada en las SVM, corresponden a descriptores de auto-similaridad Shechtman y Irani [2007], histogramas de gradientes orientados a diferentes escalas, un descriptor GIST (uno de los descriptores global más conocidos), a color Oliva y Torralba [2001] y una bolsa de características. Para combinar el vector de puntajes de atributos con el Vector de Fisher, los autores se preocupan de que ambos vectores tengan una misma magnitud promedio para poder ser comparados mediante la distancia L2. En el caso del FV se aplica la normalización de potencia ($\alpha = 0{,}5$). Para la normalización del vector de atributos, se aprende la media de la distribución del vector usando el conjunto de datos de entrenamiento. Esta media se utiliza para centrar los datos y estimar un escalar por el cual se dividen los vectores de atributos para ser normalizados. Después de esto ambos vectores de concatenan.

Los resultados en el conjunto de datos Holidays reportados son de un MAP de 59.5 % y 55.5 % para un FV de 4096 dimensiones (64 Gaussianas) y el vector de puntajes de atributos de 2659 dimensiones, respectivamente, mientras que la combinación de ambos consigue un MAP de 64.5 %. Adicionalmente, si se pondera el valor del FV por 2.3 (valor obtenido experimentando con el conjunto de datos) el MAP aumenta a 69.9 %.

Gordo, Rodríguez-Serrano y Perronnin, en [Gordo et al., 2012], exploran diferentes métodos para aprovechar la información de etiquetas de bases de datos de categorías de objetos en el problema de recuperación de imágenes. Un aspecto relevante de este trabajo es el uso de descriptores SIFT de 128 dimensiones y descriptores de colores de 96 dimensiones, ambos muestreados densamente, no utilizando un detector de regiones de interés, contrario a lo empleado usualmente en la literatura. Esta decisión tiene un carácter más bien empírico, puesto que fueron observados mejores resultados al usar la base de datos Holidays y peores en UKB. Cabe notar que esto ahorra el tiempo de cómputo que toma utilizar el detector que es sustancial. La dimensionalidad de las características SIFT y de colores es reducida a 64 dimensiones utilizando PCA. Posteriormente, por cada tipo de descriptor, se crea un FV de 2048 dimensiones y se procede a concatenarlos. La distancia se calcula mediante producto punto. Con esto se logra un MAP de 77.4 % en Holidays y 3.19 en UKB. El aporte principal de este trabajo es la prueba de diferentes métodos de reducción de dimensionalidad supervisados, como aprendizaje de métricas, atributos, análisis canónico de correlación (CCA) y su propuesta de *subespecie conjunto y aprendizaje de clasificador* (JSCL). Las etiquetas para el entrenamiento de estos métodos se extraen del conjunto de datos ImageNet Large Scale

Visual Recognition Challenge 2010, que contiene mil clases y más de un millón de imágenes. Usando JSCL y CCA se logran obtener mejoras considerables en los resultados después de reducir la dimensionalidad de los Vectores de Fisher, respecto a PCA.

Uno de los métodos que se utilizaron para extraer información de las bases de datos de imágenes categorizadas fueron los atributos propuestos por [Douze et al., 2011]. Al usar Vectores de Fisher con la información de color y muestreados densamente, no se obtuvo mejoras significativas al incorporar los atributos para una misma dimensionalidad. Los autores mencionan que la diferencia se debe a que en el trabajo original [Douze et al., 2011] los atributos tenían información de la imagen que los FV no poseen. Para probar este punto realizan pruebas con Vectores de Fisher sin información de color y los concatenan a un vector de atributos obtenido utilizando información de color. De esta forma sí se logró una mejora significativa en la precisión. Algo que no se tomó en consideración es el método de muestreo de descriptores, lo cual se pondrá de manifiesto en la propuesta.

3.6. Versión no probabilística del Vector de Fisher: VLAD

VLAD (Vector of Locally Aggregated Descriptors) o Vector de Descriptores Agregados Localmente, es una alternativa no probabilística del Vector de Fisher [Jégou et al., 2012]. Resulta importante conocerla, puesto que hay numerosos trabajos que la utilizan al ser fácil de implementar y rápido de entrenar y tiene fuertes nexos con el FV. Tal como con la Bolsa de Características, tiene un conjunto de palabras visuales $\{\mu_1, \ldots, \mu_K\}$ que es aprendido utilizando k-medias. Cada descriptor local x_n es asociado a su palabra visual más cercana $NN(x_n)$. Por cada palabra visual μ_i, las diferencias $x_n - \mu_i$ de los vectores x_n asignados a μ_i son acumuladas:

$$v_i = \sum_{x_n : NN(x_n) = i} x_n - \mu_i \tag{3.12}$$

VLAD corresponde a la concatenación de los K vectores D-dimensionales v_i. En [Jégou et al., 2012] se detallan las condiciones bajo las cuales la representación del Vector de Fisher converge a VLAD.

3.7. Resumen del capítulo

En este capítulo se ha revisado el estado del arte de los métodos de representaciones compactas de imágenes. Primero se dieron a conocer las bases de datos de prueba que se utilizan comúnmente con el objetivo de describir el ambiente en el cual se desenvuelven los métodos de representaciones compactas de imágenes. Posteriormente el enfoque estuvo en la descripción de la representación Bolsa de Características y alguna de sus variantes, puesto que es de las más populares en el estado del arte, es simple y tiene algunas similaridades con los Vectores de Fisher. A continuación se detallaron las bases y forma de construir Vectores de Fisher mediante la utilización de mezclas de Gaussianas en el esquema de trabajo del Kernel de Fisher. Finalmente se dieron a conocer algunos trabajos relevantes del estado del arte que tienen alguna relación con lo que se propone en este documento.

En el siguiente capítulo se procederá a describir la propuesta, consistente en un método simple de combinación de Vectores de Fisher que tiene la capacidad de mejorar la precisión, manteniendo un uso casi idéntico de memoria.

Capítulo 4

Metodología

En este capítulo se dará a conocer la metodología seguida para diseñar el método propuesto aplicable al problema de recuperación de imágenes basado en contenido. Primero se iniciará con una descripción de carácter intuitiva para dar paso a una propuesta más formal.

4.1. Método propuesto

Los descriptores SIFT extraídos de regiones encontradas por detectores de regiones de interés y aquellos extraídos mediante un muestreo denso obedecen a diferentes procesos de generación. En particular, en este capítulo se considerará el detector hessiano afín descrito en la sección 2.2.1 y el muestreo denso o sistemático de la sección 2.2.2. En primera instancia esto podría no parecer evidente, puesto que todos son descriptores SIFT, pero la mayoría de los detectores de puntos de interés centran al descriptor en una zona de alto contraste como una esquina y lo rota según un criterio con el objetivo de que sea invariante ante cambios de rotación, posición y escala como se vio en el capítulo 2. Estas características hacen que los descriptores incapaces de ubicarse en regiones planas como el cielo o de diferenciar esquinas con el mismo aspecto, pero con rotaciones diferentes en la misma imagen. Por esto, las estadísticas relacionadas a cada grupo de descriptores no son iguales, por lo que al ajustar una mezcla de Gaussianas, sus parámetros, y sus Vectores de Fisher, serán diferentes.

Cada uno de estos dos métodos de muestreo de descriptores tiene fortalezas y debilidades. Los detectores de puntos de interés funcionan muy bien bajo rotaciones, a diferencia del muestreo denso, pero este último funciona mejor en otros escenarios donde los detectores no logran encontrar una gran cantidad de puntos.

Para tomar ventaja de ambos métodos de muestreo, dentro del esquema de trabajo del Vector de Fisher, una posibilidad consiste en generar una nueva GMM uniendo las GMMs aprendidas previamente. Al ser las Gaussianas independientes entre sí, realizar esta unión sería un trabajo fácil: simplemente habría que introducir cada Gaussiana con sus respectivos parámetros y re-normalizar los pesos al finalizar. Hay dos problemas relacionados con esta

Figura 4.1: Comparación entre los resultados obtenidos en la recuperación de imágenes con diferentes conjuntos de descriptores. Los resultados "sparse´´ se refieren a los obtenidos mediante un detector de regiones de interés Hessiano afín. La primera fila muestra una imagen usada como consulta y las filas restantes las imágenes que usan los tres primeros lugares para las diferentes representaciones de imágenes. La imagen es original de Mardones et al. [2013].

idea: el primero es que la técnica de muestreo (y por tanto la distribución) ligada a un grupo específico de descriptores se conoce a priori, pero al usar la GMM unida cada descriptor es tratado de la misma forma, perdiéndose la información que decía como éste fue muestreado. Un segundo problema es que al calcular los FVs, los distintos tipos de descriptores se compararán con las distribuciones de sus pares, agregando ruido no deseado. Una solución eficiente, aunque informal, corresponde a utilizar los pesos originales de las Gaussianas para cada grupo de descriptores, usando solamente los parámetros que modelan a esos descriptores, poniendo artificialmente un peso cero a las otras Gaussianas. El vector resultante, después de eliminar los ceros obtenidos de las Gaussianas ignoradas, es idéntico al FV que se obtendría mediante la concatenación de los dos Vectores de Fisher originales basados en los diferentes métodos de muestreo. De esta forma, la concatenación de FVs es el método propuesto en este trabajo para combinar Vectores de Fisher.

Al usar PCA u otra técnica de reducción de dimensionalidad en los Vectores de Fisher concatenados, se debe aplicar a cada FV de manera individual para poder beneficiarse del conocimiento de la distribución base de cada FV.

Una ventaja importante del método propuesto al utilizar descriptores muestreados de formas diferentes es que los FVs concatenados proveen información adicional en un mismo espacio en memoria (en el próximo capítulo se profundizará en esto), mientras que solamente hay un costo computacional fijo asociado a la extracción de características adicionales de la imagen de entrada y el proceso previo de ajustar los parámetros de una GMM extra.

La Figura 4.1 muestra los resultados de un par de búsquedas para FVs de forma individual

y su contraparte concatenada. Se puede apreciar que los Vectores de Fisher concatenados son capaces de encontrar las imágenes relevantes, incluso cuando una de sus partes falla.

4.1.1. Formalización de la propuesta

Sea $u_1 = \sum_{k=1}^{K} w_{1,k}\mathcal{N}(X_1|\mu_{1,k}, \sigma_{1,k}^2)$ una mezcla de K Gaussianas que representa la distribución de los descriptores X_1, los cuales son muestreados mediante el método M_1 y son D-dimensionales. Por otro lado, sea $u_2 = \sum_{k=1}^{K} w_{2,k}\mathcal{N}(X_2|\mu_{2,k}, \sigma_{2,k}^2)$ una mezcla de K Gaussianas que representa la distribución de los descriptores X_2, los cuales son muestreados mediante el método M_2 y son D-dimensionales. Que el número de Gaussianas y la dimensionalidad de los descriptores sea la misma tiene el objetivo de simplificar lo que se desarrollará a continuación, sin pérdida de generalidad. Las covarianzas de las Gaussianas de u_1 y u_2 serán diagonales siguiendo lo realizado con los Vectores de Fisher.

Suponiendo que las mezclas de Gaussianas que modelan la generación de descriptores X_1 son sustancialmente diferentes de las de X_2, lo siguiente se debería cumplir.

$$P(X \in X_1) = \sum_{k=1}^{K} w_{2,k}\mathcal{N}(X|\mu_{2,k}, \sigma_{2,k}^2) \approx 0 \tag{4.1}$$

$$P(X \in X_2) = \sum_{k=1}^{K} w_{1,k}\mathcal{N}(X|\mu_{1,k}, \sigma_{1,k}^2) \approx 0 \tag{4.2}$$

Puesto que las Gaussianas son independientes entre sí, es posible unir ambas mezclas de Gaussianas con el objetivo de que representen datos de X_1 y X_2:

$$u = \sum_{k=1}^{2K} w_k\mathcal{N}(X|\mu_k, \sigma_k^2), \tag{4.3}$$

donde $w_k = [w_{1,1}, \cdots, w_{1,K}, w_{2,1}, \cdots, w_{2,K}]^T$, $\mu_k = [\mu_{1,1}, \cdots, \mu_{1,K}, \mu_{2,1}, \cdots, \mu_{2,K}]^T$ y $\sigma_k = [\sigma_{1,1}, \cdots, \sigma_{1,K}, \sigma_{2,1}, \cdots, \sigma_{2,K}]^T$.

Si formamos un FV con esta nueva mezcla de Gaussianas, usando un conjunto X de N datos que contiene datos provenientes de X_1 y X_2, el aporte de la i-ésima Gaussiana es:

$$\mathcal{G}_i^X = \frac{1}{N\sqrt{w_i}} \sum_{n=1}^{N} \gamma_n(i)\sigma_i^{-1}(x_n - \mu_i). \tag{4.4}$$

donde

$$\gamma_n(i) = \frac{w_i\mathcal{N}(x_n|\mu_i, \sigma_i^2)}{\sum_{j=1}^{2K} w_j\mathcal{N}(x_n|\mu_j, \sigma_j^2)}, \tag{4.5}$$

$$con \sum w_i = 1, w_i \in [0,1], i = 1, ..., 2K, n = 1, ..., N. \tag{4.6}$$

Si $x_n \in X_1$ e $i \in 1, \cdots, K$, haciendo uso del supuesto de la ecuación 4.1, la ec.4.5 se reduce a:

$$\gamma_n(i) = \frac{w_i \mathcal{N}(x_n|\mu_i, \sigma_i^2)}{\sum_{j=1}^{K} w_j \mathcal{N}(x_n|\mu_j, \sigma_j^2)} \tag{4.7}$$

Si $x_n \in X_1$, e $i \in k+1, \cdots, 2K$, $\gamma_n(i) \approx 0$.

Esto implica que los términos desde el $KD+1$ al $2KD$ del FV son igual a cero si los descriptores provienen de X_1. Es decir, si los descriptores pertenecientes a X_1 sólo aportan valor entre los términos 1 y KD del FV. De forma análoga, los descriptores que pertenecen a X_2 solamente entregan valor desde los términos $KD+1$ al $2KD$ del Vector de Fisher.

Por lo tanto, si se agrupan los descriptores pertenecientes a X_1 en B_1 y los de X_2 en B_2 y solamente se consideran las Gaussianas que aportan valor en cada grupo, el FV resultante se puede escribir como $\mathcal{G}_\lambda^X = [\mathcal{G}_{\lambda_1}^{B_1} \mathcal{G}_{\lambda_2}^{B_2}]^T$, donde λ corresponde a los parámetros de u_1 y u_2, λ_1 y λ_2 a los parámetros de u_1 y u_2 respectivamente. Este vector es equivalente al que se puede obtener mediante la concatenación de los FVs calculados de forma independiente con las mezclas de Gaussianas iniciales y sus respectivos descriptores.

4.2. Resumen del capítulo

En este capítulo se describió el método propuesto para enfrentar el problema de recuperación de imágenes utilizando la concatenación de Vectores de Fisher. La explicación intuitiva inicial fue seguida con un enfoque formal en el cual se realizó un supuesto que da una posible explicación sobre lo que sucede al concatenar estos vectores utilizando diferentes características.

Capítulo 5

Experimentos y resultados

Este capítulo empieza por describir la elección de bases de datos y medidas de desempeño a utilizar, analizando estas decisiones. A continuación se describirán y justificarán una serie de experimentos que tienen el objetivo de demostrar empíricamente que la propuesta obtiene mejores resultados y que sus supuestos son adecuados. Finalizará realizando una comparación con los resultados de otros trabajos del estado del arte que funcionan con restricciones similares.

5.1. Conjuntos de datos de prueba y medidas de desempeño

Los conjuntos de datos de prueba a utilizar corresponden a Holidays [Jégou et al., 2008] y University of Kentucky Recognition Benchmark (UKB) [Nister y Stewenius, 2006]. Las descripciones de estas bases de datos fueron dadas en la sección 3.2. Con Holidays se utiliza la medida MAP (precisión media promedio) 2.4.1.2 para medir el desempeño de los métodos de recuperación de imágenes, mientras que UKB utiliza una medida propia llamada Puntaje Kentucky (KS) que fue descrita junto a la base de datos.

Ambos conjuntos de datos comparten algunas características tales como el uso de imágenes naturales y pequeño número de imágenes relevantes por consulta. La mayor diferencia está en que UKB se enfoca exclusivamente en detectar objetos vistos en diferentes perspectivas, mientras que Holidays tiene un enfoque más variado, conteniendo fotos de paisajes naturales, hechos por el hombre, como también animales, objetos y monumentos. Una falencia que podría considerarse importante dependiendo de la aplicación es el reducido número de imágenes relevantes por consulta, puesto que en muchos casos de uso reales, el número de éstas puede ser mucho mayor.

5.2. Diseño de Experimentos

Los experimentos realizados tienen dos objetivos: el primero consiste en demostrar de manera empírica que el uso de Vectores de Fisher concatenados, basados en descriptores muestreados de maneras complementarias, logran un aumento de precisión para un uso de memoria idéntico por imagen, respecto a su contraparte sin concatenar. El segundo objetivo es dar evidencia de que los supuestos realizados en la propuesta son razonables.

5.2.1. Sistemas de recuperación de imágenes base

En este trabajo se considerarán dos variantes de sistemas de recuperación de imágenes que producen Vectores de Fisher. El primero consiste en un sistema que utiliza un detector de regiones de interés hessiano afín (sección 2.2.1) y descriptores SIFT (sección 2.3). Los descriptores SIFT son de 128 dimensiones y son reducidos a 64 mediante PCA como se sugiere en Jégou et al. [2012] para ser un poco más robusto ante el ruido en las imágenes. Para ajustar la mezcla de Gaussianas que se utiliza para calcular el Vector de Fisher, se utilizan dos millones de descriptores de la base de datos MIR-FLICKR25K [Huiskes y Lew, 2008]. Se realizan experimentos utilizando 64 y 256 Gaussianas con el objetivo de ver la influencia de este parámetro de los resultados en los distintos escenarios. Para disminuir el uso de memoria de los Vectores de Fisher con el propósito de mostrar cómo se comportaría a gran escala, se utiliza PCA. Para estimar la matriz de vectores propios se utilizan 25000 FV extraídos de las imágenes de MIR-FLICKR25K.

El segundo sistema de recuperación de imágenes solamente se diferencia en el detector de regiones de interés por un muestreo denso (sección 2.2.2) con un espaciado de ocho pixeles entre el centro de cada región de interés, sólo a una escala.

Los resultados obtenidos por ambos sistemas se pueden ver en la parte superior de las tablas 5.1 y 5.2.

5.2.2. Sistemas de recuperación de imágenes propuesto

El sistema de recuperación de imágenes propuesto consiste en la concatenación de los Vectores de Fisher de los sistemas base. Cabe mencionar que si se quiere reducir la dimensionalidad a una dimensión D, para realizar una comparación justa respecto a los métodos base, cada componente del Vector de Fisher concatenado ve su dimensionalidad reducida a D/2.

En las tablas 5.1 y 5.2 se puede apreciar que los resultados del método propuesto son variados dependiendo de la base de datos. En Holidays los resultados son muy buenos, logrando un MAP superior a lo que puede lograr cada método por si solo. En UKB, el KS disminuye al utilizar el método propuesto.

Esto último tiene una explicación relativamente simple: UKB es una base de datos cuyas consultas consisten exclusivamente en imágenes con transformaciones afines, por lo que utilizar un detector robusto ante transformaciones afines es una mucho mejor opción que utilizar un muestreo denso. De todas formas, esto no significa que el muestreo denso no proporcione información. Siguiendo el ejemplo de Douze et al. [2011], se procedió a ponderar los Vectores

Cuadro 5.1: Resultados en Holidays

Tipo de muestreo		K	D	MAP en Holidays					
denso	H.A.			D' = D	D' = 2048	D' = 512	D' = 128	D' = 64	D' = 32
X		64	4096	60.9	61.4	60.8	58.6	55.3	51.2
	X	64	4096	56.6	58.8	60.3	54.9	52.0	47.4
X	X	64	8192	69.8	**70.1**	**68.1**	**65.0**	**61.3**	**56.2**
X		256	16384	62.3	61.9	61.3	59.3	57.4	52.3
	X	256	16384	62.3	64.6	59.4	53.9	50.8	47.6
X	X	256	32768	**72.8**	69.8	67.6	64.6	60.2	54.2

Cuadro 5.2: Resultados en UKB

Tipo de muestreo		K	D	KS en UKB					
denso	H.A.			D' = D	D' = 2048	D' = 512	D' = 128	D' = 64	D' = 32
X		64	4096	2.24	2.26	2.24	2.15	2.06	1.96
	X	64	4096	3.28	**3.33**	**3.33**	**3.15**	**3.03**	2.82
X	X	64	8192	3.18	3.15	2.98	2.76	2.63	2.44
X		256	16384	2.29	2.33	2.27	2.23	2.20	2.09
	X	256	16384	**3.38**	3.33	3.09	3.08	3.01	**2.84**
X	X	256	32768	3.29	3.07	2.89	2.76	2.67	2.46

de Fisher para darle más importancia a la descriptores muestreados con el detector Hessiano afín. En la tabla 5.3 se aprecia cómo al aumentar la ponderación el rendimiento de los vectores concatenados supera al original por un margen significativo en las dimensionalidades más altas. En las menores la diferencia disminuye, lo cual se debe a que la descripción basada en regiones de interés llega a un umbral respecto a la reducción de dimensionalidad donde comienza a perder una gran cantidad de información relevante. Esto podría ser mitigado asignando una cantidad de memoria una superior al FV que utiliza regiones de interés, respecto a su contraparte.

Para verificar que los supuestos hechos en la formulación de la propuesta son razonables, se realizaron tres experimentos que consistieron en extraer los descriptores muestreados de ambas maneras y usarlos en conjunto sin diferenciar su procedencia. En el primer experimento (Método 1 en la tabla 5.4) se extraen ambas clases de descriptores de conjunto de imágenes de entrenamiento y sus dimensionalidades son reducidas a 64 dimensiones empleando una matriz de proyección aprendida usando los dos tipos de descriptores. Éstos últimos se utilizan para estimar la mezcla de Gaussianas. Los datos de prueba son reducidos utilizando la misma matriz de proyección. El resto del proceso es idéntico al base. El segundo experimento (Método 2) se diferencia del primero en que la reducción de dimensionalidad de los descriptores se realiza por separado dependiendo de su origen y posteriormente se tratan de igual manera. El último experimento (Método 3) es idéntico al primero, con la particularidad de que no se aplica PCA a los descriptores, usando 128 dimensiones. El objetivo de este último es mostrar la influencia que PCA ejerce. Los resultados se exhiben en la tabla 5.4. Es posible

Cuadro 5.3: Resultados en UKB con diferentes ponderaciones

K	D	KS en UKB								
		W = 1	W = 1.2	W = 1.4	W = 1.6	W = 1.8	W = 2.0	W = 2.2	W = 2.4	W = 2.6
64	8192	3.17	3.30	3.37	3.40	3.41	**3.41**	3.41	3.41	3.40
64	2048	3.15	3.29	3.37	3.43	3.47	3.48	**3.48**	3.48	3.47
64	512	2.98	3.13	3.24	3.31	3.36	3.38	3.39	**3.40**	3.38
64	128	2.76	2.89	3.00	3.06	3.11	3.14	3.16	3.17	**3.18**
64	64	2.63	2.74	2.81	2.88	2.93	2.96	2.99	3.00	**3.02**
64	32	2.44	2.52	2.58	2.62	2.65	2.67	2.68	2.69	**2.71**
256	32768	3.29	3.41	3.47	3.49	3.51	**3.51**	3.50	3.50	3.49
256	2048	3.07	3.22	3.30	3.36	3.39	3.40	**3.41**	3.39	3.38
256	512	2.90	3.03	3.15	3.23	3.27	3.30	3.31	**3.32**	3.31
256	128	2.76	2.87	2.96	3.04	3.09	3.13	3.16	3.18	**3.18**
256	64	2.67	2.76	2.84	2.90	2.94	2.98	3.02	3.04	**3.05**
256	32	2.46	2.54	2.60	2.65	2.68	2.71	2.73	2.75	**2.75**

Cuadro 5.4: Resultados mezclando descriptores

Método	K	D	MAP en Holidays
Método 1	64	4096	64.5
Método 2	64	4096	57.0
Método 3	64	8192	59.0
Propuesta	32	4096	**67.8**

apreciar que en ambos casos la precisión de la propuesta es superior, reafirmando la validez de los supuestos, pero también existe una diferencia sustancial entre los resultados de los experimentos realizados. En el primer experimento, se cree que al reducir la dimensionalidad de forma conjunta, todos los descriptores son llevados a un subespacio común que puede ser descrito de mejor forma por un número limitado de Gaussianas, en relación al segundo experimento y el tercer experimento confirma esta idea.

Otra pregunta que surge al realizar estos experimentos corresponde a la selección del número de Gaussianas, teniendo en cuenta que después de una etapa de reducción de dimensionalidad pueden llegar a usar un mismo espacio en memoria. Para esto se realizan pruebas adicionales en Holidays con 32 y 128 Gaussianas. En la tabla 5.5 se muestran los resultados. En general se cumple que al aumentar el número de Gaussianas la precisión aumenta [Perronnin y Dance, 2007], pero al reducir la dimensionalidad esto no necesariamente se cumple. A menor uso de memoria final por imagen, es sería recomendable utilizar un número menor de Gaussianas y vice versa.

De forma adicional se han realizado experimentos para verificar que las ventajas del método propuesto se mantengan a medida que el número de imágenes en la base de datos aumenta. Las imágenes adicionales corresponden al segundo subconjunto de MIRFLICKR-1M. Los resultados obtenidos por el método propuesto son bastante prometedores, puesto que siempre las bajas registradas de los vectores concatenados, siempre fueron inferiores a la suma de las bajas registradas por los FVs por separado. En las figuras 5.1 y 5.2 se pueden

Cuadro 5.5: Resultados en Holidays con diferente cantidad de Gaussianas

K	D	MAP en Holidays					
		D' = D	D' = 2048	D' = 512	D' = 128	D' = 64	D' = 32
32	4096	67.8	67.7	67.2	63.2	58.4	54.1
64	8192	69.8	70.1	68.1	65.0	61.3	**56.2**
128	16384	72.3	**71.9**	**69.2**	**65.2**	**61.5**	55.4
256	32768	**72.8**	69.8	67.6	64.6	60.2	54.2

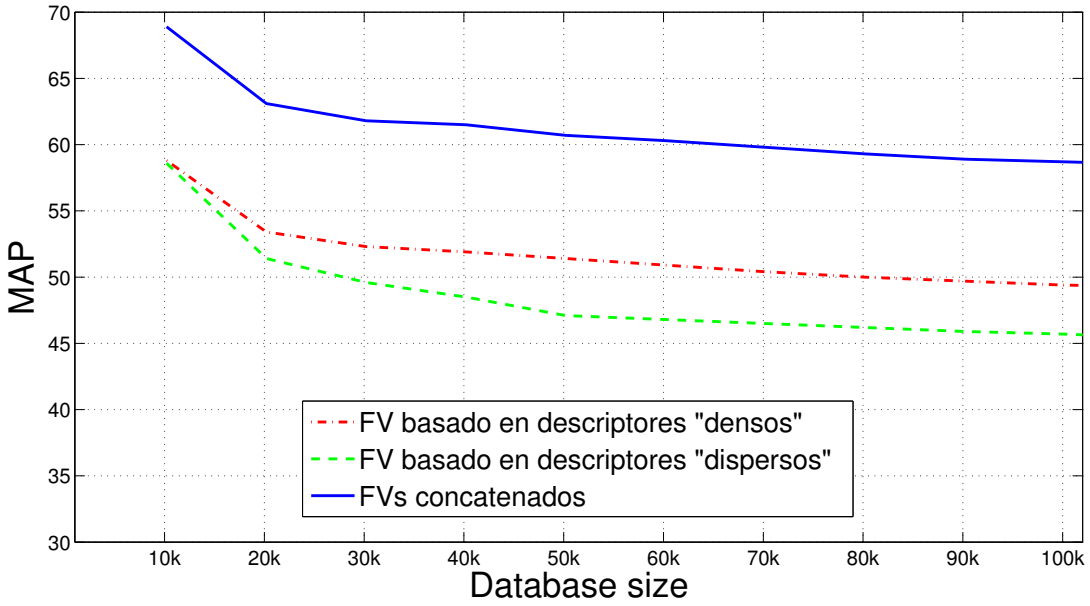


Figura 5.1: MAP de la propuesta en Holidays al aumentar el número de imágenes. Los FVs concatenados son reducidos mediante PCA a 512 dimensiones (256 y 256) y se compara con FVs con descriptores muestreados de ambas formas, con 512 dimensiones cada uno.

ver los resultados obtenidos para el conjunto de datos Holidays y UKB respectivamente.

5.3. Comparaciones con otros trabajos del área

Es importante realizar comparaciones con otros trabajos del estado del arte para cuantificar el aporte generado con la propuesta. Es relevante notar que en muchos casos, la propuesta es compatible con otros trabajos, puesto que se integra como un método de ensamblado, agregando un paso al flujo de trabajo normal.

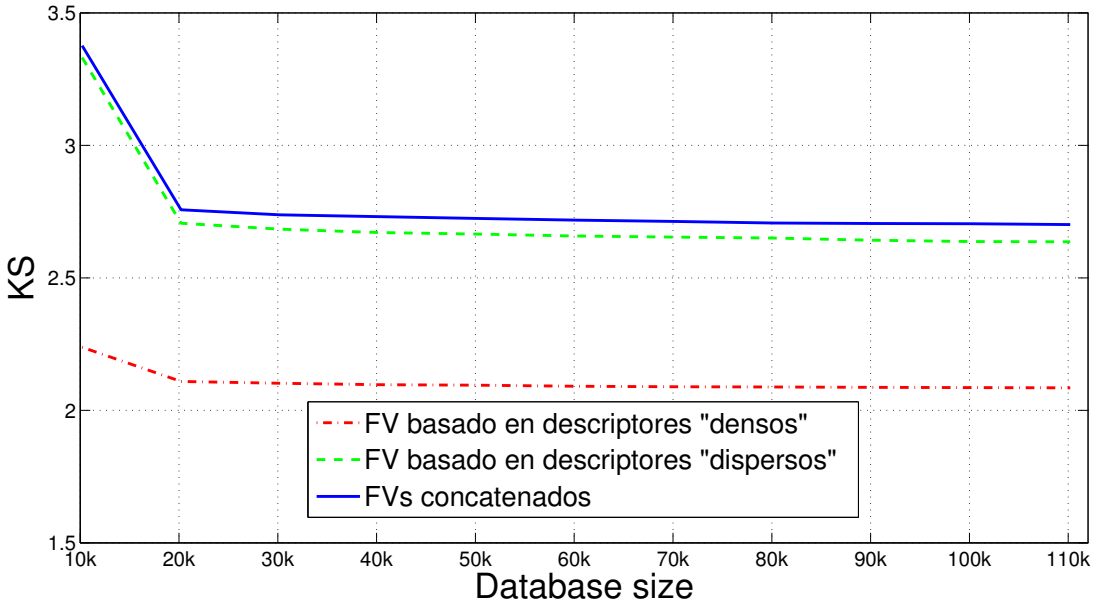


Figura 5.2: KS de la propuesta en UKB al aumentar el número de imágenes. Los FVs concatenados son reducidos mediante PCA a 512 dimensiones (256 y 256) y se compara con FVs con descriptores muestreados de ambas formas, con 512 dimensiones cada uno.

5.3.1. Criterios de selección de trabajos a ser comparados

Para realizar comparaciones justas, es importante que las condiciones de trabajo sean similares. Los criterios a considerar son los siguientes:

1. Sin uso de memoria adicional por imagen en la base de datos. Esto descarta por ejemplo, métodos que incluyen información espacial de las imágenes y otros que realizan consultas extendidas (las imágenes de la base de datos guardan información de sus vecinos). En algunos casos se podría discutir que se podría reducir la dimensionalidad de la representación para quedar no emplear espacio adicional; en estos casos se considerará el enfoque que fue dado en el documento donde se expone el método.

2. Sin conocimiento previo de las imágenes a utilizar. Existen algunos métodos que se aplican a situaciones donde se conocen de antemano las imágenes a utilizar y se ajustan solamente a este conjunto.

3. Solamente uso de la información de intensidad de la imagen. Con esto se descartan resultados obtenidos utilizando información de color, por ejemplo. De todas formas, si resulta interesante, se harán excepciones, pero se mencionará que el método no se ajusta al criterio.

Cuadro 5.6: Comparaciones en Holidays

Método	D	RdD	MAP en Holidays D' = D	D = 512	D = 128	D = 32
FV + Atributos [1]	6755	PCA	69.9	68.2	63.3	54.0
FV (SIFT + color) [2]	4096	JSCL		78.9	76.4	67.7
FV, K = 64 [3]	4096	PCA	59.5	61.0	56.5	48.0
VLAD [3]	4096	PCA	55.6		55.7	
BoF K = 200.000 [3]	200000		54.0			
VLAD + Unión de Vocs [4]		PCA			61.4	
VLAD + Voc adapt +innorm [5]	32768	PCA	64.6		62.5	
VLAD + RootSIFT + RN + LCS [6]	8192		65.8			
Propuesta, K = 64	8192	PCA	69.8	**68.1**	**65.0**	**56.2**
Propuesta, K = 256	32768	PCA	**72.8**	67.6	64.6	54.2

5.3.2. Comparación con Trabajos externos

El primer trabajo a utilizar para realizar las comparaciones corresponde a [Jégou et al., 2012]. Este trabajo fue elegido como base de comparación externa, puesto que la implementación base de nuestra propuesta se basa en la que se expone en este trabajo. Se emplea un Vector de Fisher basado en un detector de regiones de interés Hessiano afín y el resto de los pasos seguidos son idénticos. Las pequeñas diferencias de rendimiento se pueden explicar por la utilización de diferentes bases de datos de entrenamiento.

Otro trabajo bastante relacionado con la propuesta corresponde al de Douze [Douze et al., 2011], descrito en la sección 3.5.3. Hay que tener en cuenta que los resultados de ese trabajo incluyen información de color, pero se tomó en cuenta por ser uno de los pocos trabajos que combina diferentes tipos de características en el problema de recuperación de imágenes basado en contenido. Gordo [Gordo et al., 2012], en su implementación también concatena Vectores de Fisher, pero ambos usando un muestreo denso, diferenciándose en el tipo de información que extraen de la imagen (descriptores SIFT y estadísticas de colores [Perronnin et al., 2010b].

En [Jégou y Chum, 2012], se describe cómo podría aplicarse en VLAD unas mejoras en el proceso de reducción de dimensionalidad, considerando el uso de PCA. Resulta interesante la comparación, puesto que son métodos complementarios. En [Arandjelović y Zisserman, 2013] utiliza VLAD y propone un método para mantener un vocabulario adaptativo capaz de incorporar patrones de la base de datos diferentes a los encontrado en el conjunto de entrenamiento. También expone cómo realizar una nueva normalización en VLAD con el objetivo de restarle importancia a las características altamente repetitivas en las imágenes. Siguiendo esta línea, en [Delhumeau et al., 2013] implementa varias mejoras incrementales a VLAD, procesando los descriptores SIFT de manera distinta y realizando una normalización de los residuos, las cuales pueden ser aplicadas a los Vectores de Fisher con algunas modificaciones pequeñas.

Cuadro 5.7: Comparaciones en UKB

Método	D	RdD	KS en UKB			
			D' = D	D = 512	D = 128	D = 32
FV (SIFT + color) [2]	4096	JSCL		3.36	3.31	3.04
FV, K = 64 [3]	4096	PCA	3.35		3.33	
VLAD [3]	4096	PCA	3.28		3.35	
BoF K = 200.000 [3]	200000		2.81			
VLAD + Unión de Vocs [4]		PCA			**3.36**	
Propuesta, K = 64, W = 2	8192	PCA	3.41	**3.38**	3.14	2.67
Propuesta, K = 256, W = 2	32768	PCA	**3.51**	3.30	3.13	2.71

En las tablas 5.6 y 5.7 se pueden ver los resultados comparativos entre los trabajos mencionados y la propuesta, bajo las dimensionalidades reportadas en cada trabajo. En general, los resultados son favorables para la propuesta, sin contar los métodos que utilizan información de color, logra mejores resultados en ambas bases de datos, excepto cuando la descripción es de baja dimensionalidad en UKB. De forma adicional, es muy importante tener en consideración que la mayoría de los métodos en la comparación pueden incorporar nuestra propuesta a su esquema de trabajo y mejorar sus resultados.

5.4. Resumen del capítulo

En este capítulo se ha discutido sobre los experimentos a realizar, se ha apoyado empíricamente el valor de los supuestos realizados en la propuesta y se han realizado numerosos experimentos en dos conjuntos de datos de prueba, de forma independiente en primera instancia, para después realizar comparaciones de nuestro método con otros métodos del estado del arte, con resultados favorables.

[1] Douze et al. [2011]
[2] Gordo et al. [2012]
[3] Jégou et al. [2012]
[4] Jégou y Chum [2012]
[5] Arandjelović y Zisserman [2013]
[6] Delhumeau et al. [2013]

Capítulo 6

Conclusiones

En este último capítulo se presenta un resumen de los resultados obtenidos en esta tesis junto a comentarios finales. Primero se realizará un recuento de los resultados obtenidos y se realizarán algunas observaciones. Esto llevará a una discusión sobre las posibles direcciones de investigación en trabajos futuros.

6.1. Resultados

Lo primero que se pudo observar al emplear el método propuesto, en el conjunto de datos Holidays, fue una precisión significativamente superior respecto a sus partes. En UKB el desempeño de los vectores concatenados fue levemente inferior, pero se logró explicar por las características propias del conjunto de datos, teniendo los detectores de regiones de interés un impacto mucho mayor. Una forma simple de tratar esta peculiaridad es dando una mayor ponderación al FV basado en detectores de ROI.

Al agregar cien mil imágenes a los conjuntos de datos, las conclusiones respecto a la mejora significativa del vector concatenado respecto a sus partes (para un mismo uso de memoria) se mantienen. También se realizaron experimentos que apoyaron el supuesto realizado en el planteamiento formal de la propuesta.

La realización de pruebas utilizando diferente número de Gaussianas permitió ver un patrón que indicaba que mientras mayor fuese el número de Gaussianas se requiere una mayor cantidad de memoria para conservar intactos los resultados. Esto implica que si no existen limitaciones de memoria, usar un FV con una gran número de Gaussianas sin reducción de dimensionalidad entregaría la máxima precisión, pero si solo se dispone de una cantidad de memoria limitada por FV y se requiere reducir la dimensión del FV inicial, mientras menor sea el espacio disponible, menor debería ser el número de Gaussianas empleado. El rendimiento del método propuesto supera generalmente por un margen significativo al resto de los algoritmos que utilizan solamente un tipo de descriptor y/o método de selección de regiones de interés.

Los resultados son favorables en general, la combinación de descriptores muestreados de formas diferentes ha probado ser una forma poderosa de incrementar el rendimiento del sistema, simple de implementar y con un costo fijo. Adicionalmente, es posible integrar este método a la mayoría de los trabajos del estado del arte, lo cual lo hace flexible y un aporte valioso.

6.2. Trabajo Futuro

El campo de combinación de métodos de recuperación de imágenes se ha explorado de forma bastante tangencial hasta ahora, por lo que queda muchísimo trabajo por hacer. Algunas de las direcciones a seguir de forma más inmediata, corresponden a la investigación de formas de incorporar información sobre la distribución de distancias, para que los aportes de cada Vector de Fisher sea realmente equitativo. La correlación de las distancias entre diferentes Vectores de Fisher puede entregar información adicional que permitiría combinar de mejor forma los resultados. También la utilización de diferentes detectores de regiones de interés, descriptores y características de las imágenes podrían resultar en mejoras sustanciales. Otra línea de investigación relevante para este tipo de técnicas es la de caracterización de imágenes con el propósito de seleccionar un conjunto de características idóneas al momento de representar las imágenes de una base de datos, para así evitar problemas como los que surgieron con el conjunto de datos UKB.

Otro punto que requiere trabajo es la selección del número de Gaussianas asociadas a cada clase de Vector de Fisher. Por ejemplo, la variabilidad de una mezcla de Gaussianas podría ser indicadora de si un determinado número de Gaussianas es adecuado para una clase de característica. Este punto también podría estar ligado a la inclusión de los componentes excluidos del FV (gradiente de la log-verosimilitud respecto a los pesos y varianzas) por las sugerencias realizadas en algunos de los primeros trabajos que emplearon el FV para recuperación de imágenes [Jégou et al., 2012; Perronnin et al., 2010a].

Bibliografía

Arandjelović, R. y Zisserman, A. All about VLAD. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738.

Delhumeau, J., Gosselin, P.-H., Jégou, H., y Pérez, P. Revisiting the vlad image representation. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 653–656, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2404-5. doi: 10.1145/2502081.2502171.

Douze, M., Ramisa, A., y Schmid, C. Combining attributes and Fisher vectors for efficient image retrieval. In *IEEE Conference on Computer Vision & Pattern Recognition*, Colorado Springs, United States, June 2011.

Ferrari, V., Tuytelaars, T., y Gool, L. Simultaneous object recognition and segmentation from single or multiple model views. *Int. J. Comput. Vision*, 67(2):159–188, April 2006. ISSN 0920-5691. doi: 10.1007/s11263-005-3964-7.

Gong, Y., Lazebnik, S., Gordo, A., y Perronnin, F. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(PrePrints):1, 2012. ISSN 0162-8828. doi: http://doi.ieeecomputersociety.org/10.1109/TPAMI.2012.193.

Gordo, A., Rodriguez-Serrano, J. A., Perronnin, F., y Valveny, E. Leveraging category-level labels for instance-level image retrieval. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3045–3052, 2012. doi: 10.1109/CVPR.2012.6248035.

Gordon, I. y Lowe, D. What and where: 3d object recognition with accurate pose. In Ponce, J., Hebert, M., Schmid, C., y Zisserman, A., editors, *Toward Category-Level Object Recognition*, volume 4170 of *Lecture Notes in Computer Science*, pages 67–82. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-68794-8. doi: 10.1007/11957959_4.

Huiskes, M. J. y Lew, M. S. The mir flickr retrieval evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*, New York, NY, USA, 2008. ACM.

Huiskes, M. J., Thomee, B., y Lew, M. S. New trends and ideas in visual concept detection: The mir flickr retrieval evaluation initiative. In *MIR '10: Proceedings of the 2010 ACM International Conference on Multimedia Information Retrieval*, pages 527–536, New York, NY, USA, 2010. ACM.

Jaakkola, T. S. y Haussler, D. Exploiting generative models in discriminative classifiers. In *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II*, pages 487–493, Cambridge, MA, USA, 1999. MIT Press. ISBN 0-262-11245-0.

Jégou, H. y Chum, O. Negative evidences and co-occurrences in image retrieval: the benefit of PCA and whitening. In *ECCV - European Conference on Computer Vision*, Firenze, Italy, October 2012.

Jégou, H., Douze, M., y Schmid, C. Hamming embedding and weak geometric consistency for large scale image search. In David Forsyth, A. Z. Philip Torr, editor, *European Conference on Computer Vision*, volume I of *LNCS*, pages 304–317. Springer, oct 2008.

Jégou, H., Douze, M., y Schmid, C. On the burstiness of visual elements. In *Conference on Computer Vision & Pattern Recognition*, jun 2009.

Jégou, H., Douze, M., y Schmid, C. Improving bag-of-features for large scale image search. *International Journal of Computer Vision*, 87(3):316–336, feb 2010.

Jégou, H., Douze, M., y Schmid, C. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 33(1):117–128, jan 2011. to appear.

Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., y Schmid, C. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, September 2012.

Lazebnik, S., Schmid, C., y Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 0-7695-2597-0. doi: 10.1109/CVPR.2006.68.

Lindeberg, T. y Garding, J. Shape-adapted smoothing in estimation of 3-d shape cues from affine deformations of local 2-d brightness structure. *Image and Vision Computing*, 15(6): 415 – 434, 1997. ISSN 0262-8856. doi: http://dx.doi.org/10.1016/S0262-8856(97)01144-X.

Liu, D., Hua, G., Viola, P. A., y Chen, T. Integrated feature selection and higher-order spatial feature extraction for object categorization. In *Proceedings of the 2008 IEEE Conference*

on Computer Vision and Pattern Recognition, pages 1–8, 2008. doi: 10.1109/CVPR.2008.4587403.

Lloyd, S. Least squares quantization in pcm. *Information Theory, IEEE Transactions on*, 28(2):129–137, 1982. ISSN 0018-9448. doi: 10.1109/TIT.1982.1056489.

Lowe, D. G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, November 2004. ISSN 0920-5691. doi: 10.1023/B:VISI.0000029664.99615.94.

Manning, C. D., Raghavan, P., y Schütze, H. *Introduction to Information Retrieval*. Cambridge University Press, New York, 2008. ISBN 9780521865715 0521865719.

Mardones, T., Allende, H., y Moraga, C. Combining descriptors obtained through different sampling techniques in image retrieval. In *Chilean Conference on Pattern Recognition*, Temuco, Chile, November 2013.

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., y Van Gool, L. J. A comparison of affine region detectors. *Int. J. Comput. Vision*, 65(1-2):43–72, November 2005. ISSN 0920-5691. doi: 10.1007/s11263-005-3848-x.

Morioka, N. y Satoh, S. Building compact local pairwise codebook with joint feature space clustering. In *Proceedings of the 11th European Conference on Computer Vision*, pages 692–705, Berlin, Heidelberg, 2010. ISBN 3-642-15548-0, 978-3-642-15548-2.

Mundy, J. L. Object recognition in the geometric era: A retrospective. In Ponce, J., Hebert, M., Schmid, C., y Zisserman, A., editors, *Toward Category-Level Object Recognition*, volume 4170 of *Lecture Notes in Computer Science*, pages 3–28. Springer, 2006. ISBN 3-540-68794-7.

Naphade, M., Smith, J. R., Tesic, J., Chang, S.-F., Hsu, W., Kennedy, L., Hauptmann, A., y Curtis, J. Large-scale concept ontology for multimedia. *MultiMedia, IEEE*, 13(3):86–91, 2006. ISSN 1070-986X. doi: 10.1109/MMUL.2006.63.

Nister, D. y Stewenius, H. Scalable recognition with a vocabulary tree. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2161–2168, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 0-7695-2597-0. doi: 10.1109/CVPR.2006.264.

Oliva, A. y Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2001.

Perronnin, F. y Dance, C. R. Fisher kernels on visual vocabularies for image categorization. In *CVPR*. IEEE Computer Society, 2007.

Perronnin, F., Liu, Y., Sánchez, J., y Poirier, H. Large-scale image retrieval with compressed fisher vectors. In *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3384–3391, 2010a. doi: 10.1109/CVPR.2010.5540009.

Perronnin, F., Sánchez, J., y Mensink, T. Improving the fisher kernel for large-scale image classification. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, pages 143–156, Berlin, Heidelberg, 2010b. Springer-Verlag. ISBN 3-642-15560-X, 978-3-642-15560-4.

Philbin, J., Chum, O., Isard, M., Sivic, J., y Zisserman, A. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

Philbin, J., Chum, O., Isard, M., Sivic, J., y Zisserman, A. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, 2008. doi: 10.1109/CVPR.2008.4587635.

Rothganger, F., Lazebnik, S., Schmid, C., y Ponce, J. 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *Int. J. Comput. Vision*, 66(3):231–259, March 2006. ISSN 0920-5691. doi: 10.1007/s11263-005-3674-1.

Shechtman, E. y Irani, M. Matching local self-similarities across images and videos. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, 2007. doi: 10.1109/CVPR.2007.383198.

Sivic, J. y Zisserman, A. Efficient visual search of videos cast as text retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(4):591–606, April 2009. ISSN 0162-8828. doi: 10.1109/TPAMI.2008.111.

Stockman, G. y Shapiro, L. G. *Computer Vision*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition, 2001. ISBN 0130307963.

Szeliski, R. *Computer Vision: Algorithms and Applications*. Springer-Verlag London Limited, 2011. ISBN 9781848829343.

Zezula, P., Amato, G., Dohnal, V., y Batko, M. *Similarity Search - The Metric Space Approach*, volume 32 of *Advances in Database Systems*. Kluwer, 2006. ISBN 978-0-387-29146-8.