CAPÍTULO 8

SIFT (SCALE INVARIANT FEATURE TRANSFORM)

Enrique ALEGRE[1], Laura FERNÁNEZ-ROBLES[2]

[1] Dpto. de Ingeniería Eléctrica y de Sistemas y Automática, Universidad de León, León, España.
[2] Dpto de Ingeniería Mecánica, Informática y Aeroespacial, Universidad de León, León, España.

SIFT es un método que permite detectar puntos característicos en una imagen y luego describirlos mediante un histograma orientado de gradientes. Y además, lo hace de forma que la localización y la descripción presenta una gran invarianza a la orientación, la posición y la escala. Cada punto característico queda, por lo tanto, definido mediante su vector de características de 128 elementos, y se obtiene la información de su posición en coordenadas de la imagen, la escala a la que se encontró y la orientación dominante de la región alrededor de dicho punto.

En este capítulo se explican los pasos necesarios para obtener descriptores SIFT en una imagen. Se presenta un ejercicio sencillo que sirve para ilustrar numéricamente cómo se obtiene el descriptor a partir de la región que rodea un punto característico. También se comentan las posibilidades de SIFT para realizar reconocimiento de objetos presentes en una imagen. Y, finalmente, se habla brevemente de algunas extensiones del método así como de otros descriptores de imagen relacionados que han surgido posteriormente.

8.1. Introducción

La publicación del primer artículo donde Lowe presentó SIFT (Lowe, 1999) supuso un antes y un después en el campo de la visión por computador. Si bien las publicaciones relacionadas con la detección de características locales se remonta a cuarenta y cinco años atrás, cuando Attneave (1954) observó que la información de la forma se concentra en puntos dominantes que presentan alta

curvatura, Lowe destiló el conocimiento generado durante esos años para presentar un método robusto y eficiente. A partir de ese momento, gran parte de la actividad de los mejores investigadores del campo de la visión se centró en el desarrollo, la mejora o la aplicación de SIFT y en la propuesta de métodos alternativos. Dos de los autores que más han contribuido a este campo, después de Lowe, han sido Kristian Mikolajczyk y Tinna Tuytelaars que en su trabajo, "Local Invariant Feature Detectors: A Survey", (Tuytelaars y Mikolajczyk, 2008), presentan, a lo largo de 104 páginas, una revisión de los distintos métodos y aportaciones que fueron necesarios y en los que se basa el método de SIFT (Scale Invariant Feature Transform).

Si bien, no es fácil entender SIFT sin explicar qué es una característica local, qué es y cómo funciona un detector de esquinas (o corners) y otros conceptos como la detección de blobs, la detección de regiones y la forma de implementar todo ello de manera eficiente, en este capítulo vamos a intentar hacerlo. Nuestro objetivo es presentar de forma completa y concisa las distintas etapas del método de manera que facilite su compresión o sirva, para aquellos que no lo conocen, como un primer acercamiento a esta técnica.

8.2. SIFT (Scale Invariant Feature Transform)

Para obtener un conjunto de descriptores SIFT de una imagen es necesario por un lado obtener los puntos característicos y después, para cada punto de interés, calcular su vector descriptor a partir de la información de los píxeles que lo rodean. SIFT fue propuesto para imágenes en escala de grises, por lo que el vector de características de 128 elementos que define cada píxel, contiene información sobre cómo se distribuyen los niveles de intensidad alrededor de cada punto de interés previamente obtenido.

Por lo tanto, el algoritmo consta de **dos partes** claramente diferenciadas:

a) *Obtención* de los *puntos característicos*

b) *Descripción* de la región alrededor de cada punto de interés

La **obtención de los puntos característicos** o puntos de interés, a los que habitualmente se llama en inglés *keypoints*, consiste en detectar aquellas regiones de la imagen en las que se producen diferencias de gradiente significativas a ambos lados de dicho punto. Si el método solamente hiciera eso, se podría pensar que esta etapa se podría realizar utilizando un detector de esquinas, como por ejemplo el detector de esquinas de Harris (Harris y Stephens, 1988). La propuesta de Lowe (Lowe, 1999) va un poco más allá y propone no detectar únicamente esquinas, sino **blobs**, y hacerlo de manera que esa detección sea consistente cuando el punto característico aparezca a diferentes escalas.

La diferencia entre una esquina, *corner,* y un *blob,* puede resultar, a priori, un poco difusa. Si nos fijamos en su definición, una esquina es un punto, o al menos un área pequeña en una imagen, donde confluyen –al menos- dos bordes. Se define un **blob** como una región de una imagen que se caracteriza porque algunas de sus propiedades se mantienen aproximadamente constantes. Dentro del contexto de la detección de puntos característicos, *keypoints*, la propiedad que se suele considerar constante en un blob es la similitud en su nivel de gris, típicamente medida a partir de la variación del gradiente en esa

región a lo largo de diferentes direcciones. Según esto, claramente un *blob* es una región mayor que un punto y una esquina, *corner*, es "un punto". La confusión se puede producir cuando, en el método de SIFT y en otros métodos similares, para el cálculo de características locales invariantes, se habla de "puntos de interés", *kepoints*, y cuando se estudia y compara el detector de esquinas de Harris con el Laplaciano de la Gaussiana (LoG), o la Diferencia de Gausianas (DoG) como posibles métodos para obtener "los puntos característicos". En resumen, buscamos puntos que pueden ser pequeñas regiones con un nivel de intensidad estable y alrededor de las cuales se produce una variación significativa del gradiente, en más de una dirección.

La detección de esos puntos característicos a diferentes escalas requiere crear un espacio-escala, detectar en él puntos de interés, y eliminar aquellos que se consideren poco estables, como explicaremos en las siguientes secciones.

Para obtener el **descriptor de cada punto característico**, Lowe (Lowe 1999, 2004) propuso calcular un *histograma de direcciones del gradiente local* alrededor del punto de interés. El descriptor que se obtiene se convierte en *invariante a la escala* al normalizar el tamaño del vecindario local al punto de interés en función de ella. Además, es *invariante a la rotación* porque se determina la orientación dominante de los vectores del gradiente en el vecindario del punto característico, y se utiliza dicha información para orientar la rejilla que se usa para calcular el histograma.

Lowe nombró al método SIFT: *Scale Invariant Feature Transform* porque, en base a lo anterior, transforma datos de la imagen en características locales que se obtienen y expresan en coordenadas de la imagen invariantes a la escala.

8.2.1. Etapas del algoritmo

Si bien Lowe presenta por primera vez SIFT en el CVPR'99 (Computer Vision and Pattern Recognition Conference), (Lowe, 1999), es el artículo publicado cinco años después, en la *International Journal of Computer Vision* (Lowe, 2004) donde el método es explicado con un mayor nivel de detalle. En esta última publicación se dice que las etapas para el cálculo del método son las siguientes:

1. **Detección de extremos** en el espacio-escala

   Se buscan puntos de interés en toda la imagen y en todas las escalas consideradas utilizando una diferencia de Gaussianas.

2. **Localización precisa** de puntos característicos

   Para cada uno de los puntos de interés anteriores se ajusta un modelo que permite determinar su localización y escala. Además, se seleccionan los puntos característicos, *keypoints,* eliminando los que están próximos a los bordes o tienen bajo contraste.

3. Asignación de la **orientación**

   A cada punto característico se le asigna una o varias *orientaciones* en función de las *direcciones del gradiente local*. Esta orientación conjuntamente con la ubicación y la escala calculadas anteriormente permiten que el descriptor sea invariante a estas tres situaciones.

4. **Descripción** del punto característico

   Alrededor de cada punto característico se miden los *gradientes locales de la imagen* y se utiliza su histograma para obtener una representación de esa región que es robusta a cambios significativos en la iluminación y a pequeñas distorsiones en la forma.

8.2.2. El espacio escala

La distancia a la que están ubicados los objetos afecta a su percepción. Un objeto que se encuentra lejos presenta un tamaño menor que el mismo objeto cuando está situado más próximo. Este fenómeno produce que el mismo objeto pueda aparecer en una imagen con diferentes dimensiones en función de la distancia al observador. Para obtener una descripción de los objetos que sea robusta ante estas variaciones en tamaño se utiliza el "espacio-escala".

Koenderink (Koenderink, 1984) y Lindeberg (Lindeberg, 1994) mostraron, tras realizar diversas asunciones razonables, que el único kernel posible para el espacio-escala es la función gaussiana. Por lo tanto, el **espacio escala de una imagen**, $I(x,y)$, *consiste en una familia de imágenes derivadas*, $L(x,y,\sigma)$, que se obtienen por la convolución de una gaussiana de desviación típica σ y escala variable, $G(x,y,\sigma)$, con la imagen de entrada $I(x,y)$:

$$L(x,y,\sigma) = G(x,y,\sigma) * I(x,y) \quad , \tag{8.1}$$

donde * denota la operación de convolución en las coordenadas *x* e *y,* y donde

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{8.2}$$

Un ejemplo de dicho espacio escala puede verse en la figura 8.1, donde diferentes valores de la desviación típica de la gaussiana produce sucesivas imágenes con menor nivel de detalle.

Para detectar de forma eficiente ubicaciones estables de puntos característicos, Lowe (Lowe 1999) propuso utilizar los valores extremos del espacio escala de la diferencia de Gaussianas de la imagen, $D(x,y,\sigma)$, que se puede calcular como la diferencia de dos escalas consecutivas separadas por un factor multiplicativo constante, *k*:

$$D(x,y,\sigma) = (G(x,y,k\sigma) - G(x,y,\sigma)) * I(x,y) \tag{8.3}$$

$$= L(x,y,k\sigma) - L(x,y,\sigma)$$

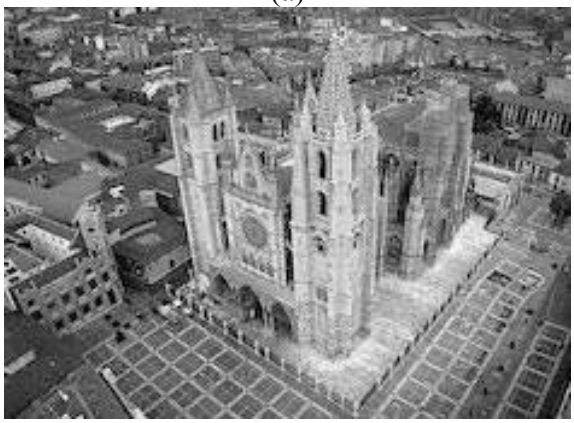
Las principales razones que llevaron a Lowe a escoger esta función, fueron:

- Es *particularmente eficiente de calcular*, ya que las imágenes suavizadas, $L(x,y,\sigma)$, tienen que obtenerse obligatoriamente para la descripción de características en el espacio escala y, por lo tanto, $D(x,y,\sigma)$, puede obtenerse simplemente restando las imágenes anteriores.

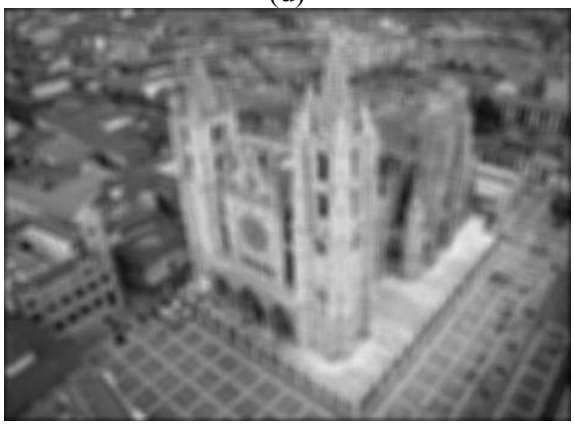
- Es una *buena aproximación al Laplaciano de la Gausiana normalizado a la escala*, $\sigma^2\nabla^2 G$, como estudió Lindeberg (Lindeberg 1994), ya que la normalización del Laplaciano utilizando el factor σ^2 es necesaria para que exista una invarianza real a la escala.

- Además, según demostró Mikolajczyk (Mikolajczyk y Schmid, 2002), los máximos y mínimos de $\sigma^2\nabla^2 G$,producen las características de imagen más estables, al compararlas con otras funciones de detección de esquinas como Harris, el Hessiano o el gradiente.

(a) (b)



(c) (d)



(e) (f)

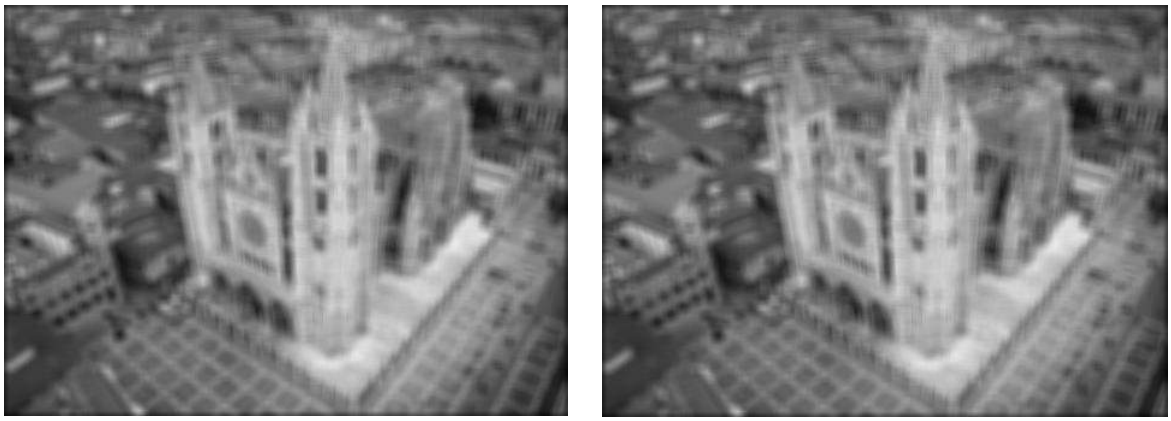


Figura 8.1. Catedral de León (*a*) Imagen original. (*b*) $\sigma = 1$ (*c*) $\sigma = 2$; *d*) $\sigma = 4$; (*e*) $\sigma = 8$; (*f*) $\sigma = 16$;

8.2.3. El espacio escala en SIFT

En SIFT el espacio escala se lleva un paso más allá porque además de construir el espacio escala de cada imagen, mediante la convolución con diferentes gaussianas, se crean varios espacios reduciendo sucesivamente el tamaño de la imagen original. A cada uno de esos espacios escala se llama una octava y se obtiene eliminando una de cada dos filas y columnas sobre la imagen de la octava anterior, reduciendo de esta forma sus dimensiones a la mitad. A este procedimiento se le conoce como la obtención de puntos característicos a partir de los extremos del espacio-escala generado a partir de diferencias de Gaussianas (DoG) dentro de una pirámide de diferencia de Gaussianas.

El concepto de pirámides de paso banda de diferencia de Gaussianas fue propuesto originalmente por Burt y Adelson (Burt y Adelson, 1983) y por Crowley y Stern (Crowley y Stern, 1984). Una pirámide Gaussiana se construye suavizando y reduciendo repetidamente la dimensión de la imagen original. La pirámide de diferencias de Gaussianas se calcula a partir de las diferencias entre los niveles adyacentes de la pirámide de Gaussiana.

Octavas y escalas

Como se ha explicado anteriormente, se parte de la imagen original y se generan imágenes progresivamente más suavizadas mediante la convolución con una gaussiana con desviación típica cada vez mayor. La relación entre las desviaciones típicas viene determinada por la constante *k*, de manera que si a la primera imagen se le aplica una σ, a la segunda se le aplica $k*\sigma$. Se dice que el conjunto de estas *n* imágenes con suavizado progresivo pertenecen a la misma ***octava***. Al dividir el tamaño de la imagen a la mitad, aparece la segunda octava, y así sucesivamente.

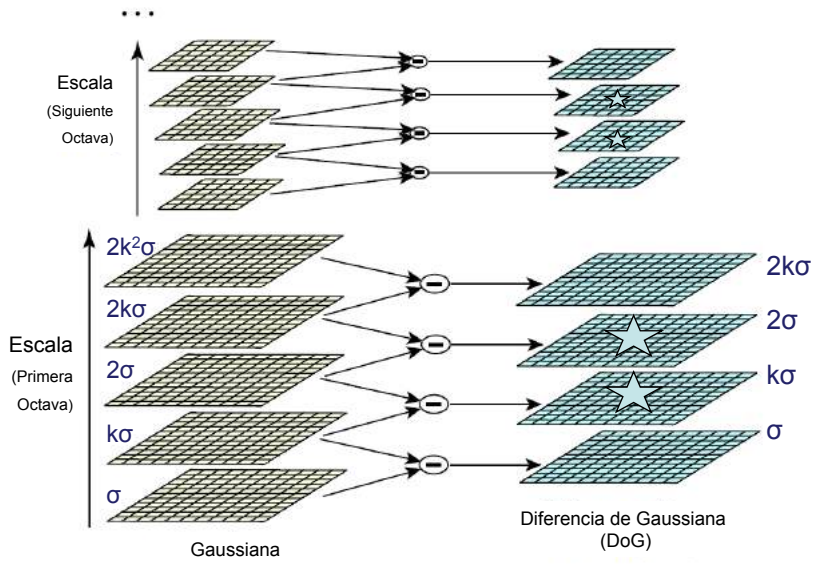


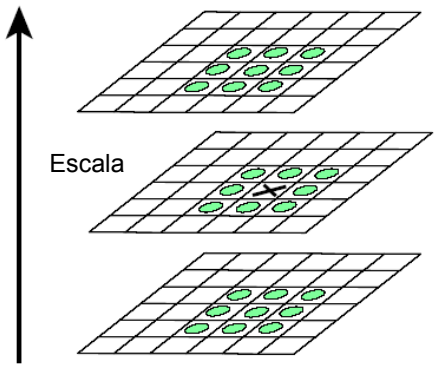
Figura 8.2. En la izquierda puede verse el espacio escala donde se representan dos octavas, la primera y la segunda con 5 escalas cada una de ellas. En la parte derecha (columna) del dibujo se ve cómo se obtiene el espacio escala de diferencias de gaussianas. Las estrellas indican las dos únicas imágenes, de cada octava, en las que puede buscarse máximos o mínimos, al ser las únicas que tienen una imagen vecina superior y otra inferior

El número de octavas y escalas dependen de la implementación que se desee realizar y suelen estar condicionadas por el tamaño y el detalle de la imagen original. Pero Lowe indicó en su trabajo (Lowe 2004) que 4 octavas y 5 escalas son lo ideal para el resto del proceso de detección de características invariantes.

Un detalle importante es que el método asume que la imagen original, $I(x,y)$ ha sido previamente suavizada con el objetivo de capturar de forma burda el efecto de que el tamaño de píxel en un CCD es finito. Esto se asume considerando la imagen de entrada como $I(x,y,\sigma_n)$, donde σ_n es un suavizado nominal, tomando $\sigma = \sigma_n = 0.5$, u otro valor similar, como el primer valor de la escala que se calcula.

8.2.4. Detección de extremos locales

Para detectar los máximos y mínimos locales de la función $D(x,y,\sigma)$ (ver ecuación 8.3), cada punto de la muestra se compara con sus ocho vecinos de la imagen en la que él se encuentra y también con los nueve vecinos de la imagen que se encuentra a una escala superior e inferior (ver figura 8.3). En dicha imagen pueden verse los 25 vecinos marcados con un círculo en verde. El punto se selecciona como un posible extremo solo si el valor de *D* es mayor que todos sus 25 vecinos o bien es menor que todos ellos. Esta comprobación tiene un coste relativamente bajo ya que la mayoría de los candidatos son descartados después de unas pocas comparaciones.



Espacio escala de DoG

Figura 8.3. Los círculos en verde indican aquellos píxeles que son vecinos de un punto y que son evaluados para determinar si el píxel marcado con "X" es un máximo o un mínimo local.

8.2.5. Localización precisa de los puntos de interés

Inicialmente (Lowe, 1999), los puntos en SIFT se ubicaban en la localización y la escala en la que se detectaban los extremos del espacio escala. Pero posteriormente, Brown (Brown y Lowe, 2002) desarrolló un método para ajustar una función cuadrática 3D a los puntos de una muestra local de manera que se pudiera determinar la localización interpolada del máximo. Sus experimentos mostraron que este ajuste proporciona una mejora sustancial en la búsqueda de correspondencias y la estabilidad ya que se eliminan puntos que tienen bajo contraste y por lo tanto son sensibles al ruido, o puntos que no están bien localizados al encontrarse a lo largo de un borde.

La propuesta de Brown utiliza la expansión de Taylor hasta los términos cuadráticos de la función espacio-escala, $D(x,y,\sigma)$, desplazada de forma que el origen esté en el punto de la muestra. La ubicación del extremo se determina calculando las derivadas de dicha función con respecto a x e igualando a cero. Brown sugirió aproximar el Hessiano y la derivada de D utilizando diferencias entre los vecinos del punto de la muestra. El valor de la función en el extremo, $D(\hat{x})$, es útil para rechazar extremos inestables con bajo contraste en la siguiente función:

$$D(\hat{x}) = D + \tfrac{1}{2}\frac{\partial D^T}{\partial x}\hat{x}\,, \tag{8.4}$$

En su artículo sobre SIFT (Lowe, 2004), Lowe propuso descartar todas los extremos de $|D(\hat{x})|$ con un valor menor de 0.03, asumiendo que los niveles de gris de la imagen están en el rango [0, 1].

Supresión de la respuesta de los puntos de interés a lo largo de los bordes

Los puntos característicos detectados a lo largo de los bordes son puntos menos estables ya que al estar su ubicación pobremente determinada pueden ser similares a pequeñas cantidades de ruido y, por ese motivo, SIFT propone su eliminación.

Al ser la diferencia de Gaussianas una función equivalente al Laplaciano de la Gausiana, es de entender que la respuesta que tiene a lo largo de los bordes sea muy fuerte. Un punto ubicado en un borde y que esté pobremente definido en la función de diferencia de Gaussianas presentará una gran curvatura a lo largo del borde pero una pequeña curvatura en la dirección principal. La curvatura en un punto, a una escala determinada se puede obtener mediante el Hessiano de 2x2, ***H****:*

$$H = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{yx} & D_{yy} \end{bmatrix}, \tag{8.5}$$

Donde D_{xx} es la derivada segunda en *x*, D_{xy} es la derivada segunda en *x* e *y*. Las derivadas se estiman a partir de las diferencias de los vecinos del punto de muestreo.

Aunque se conoce que los valores propios de ***H*** son proporcionales a las curvas principales, D. Lowe utilizó la aproximación propuesta por Harris (Harris y Stephens, 1988) para calcular esquinas. De esta manera, utilizando la traza y el determinante del Hessiano, se evita calcular explícitamente los valores propios y estudiando la relación entre ellos se eliminan los puntos de bajo contraste. Siguiendo la nomenclatura de Lowe (Lowe, 2004), llamemos α al valor propio de mayor magnitud y β al de menor. La suma y el producto de ambos puede obtenerse de la siguiente manera:

$$\begin{aligned} Tr(H) &= D_{xx} + D_{yy} = \alpha + \beta, \\ Det(H) &= D_{xx}D_{yy} - (D_{xy})^2 = \alpha\beta \end{aligned} \tag{8.6}$$

En los casos poco frecuentes en los que el determinante sea negativo el punto también se descarta como posible extremo ya que las curvaturas tienen diferentes signos.

Si se considera *r* la relación entre ambos valores propios, de manera que $\alpha = r\,\beta$, se puede expresar esta relación en función de la traza y el determinante del Hessiano, de la siguiente manera:

$$\frac{Tr(H)^2}{Det(H)} = \frac{(\alpha+\beta)^2}{\alpha\beta} = \frac{(r\beta+\beta)^2}{r\beta^2} = \frac{(r+1)^2}{r}, \tag{8.7}$$

Cuando ambos valores propios son iguales el resultado de $(r+1)^2/r$ será mínima y se irá incrementando al hacerlo *r*. Por lo tanto, para comprobar que la relación entre las curvaturas principales es inferior a un umbral determinado, *r*, Lowe propone evaluar la siguiente expresión, que se calcula de forma muy eficiente, y conservar los puntos que la cumplan:

$$\frac{Tr(H)^2}{Det(H)} < \frac{(r+1)^2}{r}, \tag{8.8}$$

La propuesta en SIFT es utilizar un valor de $r=10$ para eliminar puntos clave con una relación entre las curvaturas principales con valores mayores de 10.

8.2.6. Asignación de la orientación

Con el objetivo de obtener invarianza a la rotación, a cada descriptor se le asigna una orientación en función de las propiedades locales de la imagen.

Para realizar este cálculo de forma que sea invariante a la escala se selecciona la imagen Gaussiana suavizada, *L,* con la escala más próxima a la del punto característico del que se va a obtener su descriptor. Para cada muestra de imagen a la escala especificada, $L(x,y)$, se calcula la magnitud del gradiente, $m(x,y)$, y su orientación, $\theta(x,y)$, utilizando diferencias entre los píxeles vecinos.

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + ((L(x,y+1) - (L(x,y-1))^2}\,,$$
$$\theta(x,y) = \tan^{-1}\frac{L(x,y+1) - L(x,y-1))}{L(x+1,y) - L(x-1,y))} \tag{8.9}$$

A partir de las orientaciones obtenidas de la región alrededor del punto característico, se calcula un histograma de 36 bins conteniendo cada bin 10º hasta cubrir los 360º. Cada uno de los valores que se añaden al histograma se ponderan por la magnitud del gradiente utilizando una ventana circular Gaussiana con una σ que es 1,5 veces el tamaño de la escala a la que se localizó el punto característico.

Las direcciones dominantes de los gradientes locales se corresponderán con los picos del histograma. Se detecta el pico más alto y para él y todos aquellos con un valor superior al 80% de ese pico más alto, se creará un punto característico en la misma posición pero con las correspondientes orientaciones dominantes. Aunque los puntos con múltiples orientaciones son únicamente alrededor del 15%, éstos contribuyen mucho a que la búsqueda de correspondencias sea estable. El último paso consiste en ajustar una parábola a los 3 valores del histograma que están más próximos a cada pico elegido de manera que se interpola la posición del pico de una forma más precisa.

8.2.7. Descriptor

Mediante las operaciones explicadas en las secciones anteriores, a cada uno de los puntos característicos obtenidos, se les ha asignado una posición en la imagen, una escala y una orientación. El último paso del método consiste en calcular un descriptor que sea muy característico y al mismo tiempo lo más invariante posible a las variaciones que quedan, como son los cambios en iluminación y en el punto de vista 3D.

El descriptor propuesto por Lowe (Lowe, 1999) está inspirado en un trabajo realizado por Edelman (Edelman y col., 1997) sobre un modelo de visión biológica. Estos autores indicaron que determinadas neuronas del cortex visual primario responden al gradiente a una determinada orientación y frecuencia espacial, pero que dicho gradiente no es localizado de forma precisa en la retina sino que se permite su desplazamiento sobre un pequeño campo receptivo. Según Edelman, el funcionamiento de estas neuronas complejas permite realizar correspondencias y reconocer objetos 3D desde diferentes puntos de vista.

De forma resumida el cálculo del descriptor consta de los siguientes pasos:

1. Se obtiene la *magnitud y la orientación del gradiente* en una ventana de 16x16 muestras centrada en el punto característico.

2. Se calcula un *histograma de 8 bins con las orientaciones del gradiente para cada una de las 16 regiones* que se obtienen al dividir la ventana inicial en regiones menores de 4x4 muestras. Cada bin se corresponde con 45 grados, el primero de 0 a 44, el segundo de 45 a 89 y así sucesivamente. El histograma contendrá las orientaciones de los gradientes que aparecen en esa ventana.

3. Se *normaliza el vector de 128 elementos* que se obtiene al concatenar los histogramas de 8 valores para cada uno de las 16 divisiones que se han realizado de la región.

Se explican a continuación los detalles de cada uno de los pasos necesarios para calcular el descriptor.

1. Obtención de los gradientes en una ventana de 16x16 muestras alrededor del punto característico

Para obtener los valores de estos gradientes se realizan las siguientes operaciones.

1. Se *calculan a la escala apropiada.*

   Por lo que se selecciona la imagen con el nivel de suavizado Gaussiano que corresponde con la escala del punto característico.

2. Se obtienen con *invarianza a la orientación*.

   Para ello se rotan las coordenadas del descriptor y las orientaciones del gradiente en relación con la orientación del punto característico.

3. Se hace de forma *eficiente.*

   Todos los gradientes se calculan previamente para todos los niveles de la pirámide de Gaussianas. El cálculo es el mismo al realizado para asignar la orientación al descriptor, ya explicado previamente.

4. Se obtienen los valores *ponderados* por una gaussiana.

   Con el objetivo de evitar cambios bruscos en el descriptor ante pequeños cambios en la posición de la ventana y para reducir la influencia de los gradientes alejados del centro del descriptor, al verse éstos más afectados por errores de registro.

Para hacerlo se asigna un peso a la magnitud de cada punto mediante una función gaussiana con una σ igual a la mitad del ancho de la ventana del descriptor.

2. Histograma de orientaciones del gradiente para subregiones de 4x4 muestras

1. Se obtiene un *histograma de 8 direcciones para cada subregión*

   Para cada una de las 16 subregiones de 4x4 muestras se calcula un histograma de 8 intervalos, que son las orientaciones que resultan al subdividir el espacio en regiones de 45 grados.

   El valor que se asigna a cada uno de los 8 intervalos del histograma proviene de la suma de las magnitudes del gradiente de todas las muestras cuya orientación cae en la correspondiente región del histograma.

2. Se *distribuye el valor de cada gradiente en los intervalos adyacentes* del histograma, mediante una interpolación tri-lineal.

   Esto permite evitar los efectos de frontera de manera que el descriptor pueda cambiar de forma abrupta cuando una muestra pudiera pasar de un histograma a otro o de una orientación a otra.

   Cada una de las entradas a un intervalo del histograma se multiplica por un peso de $1 - d$ para cada dimensión, donde d es la distancia de la muestra al valor central del intervalo medida en unidades del espaciado de los intervalos del histograma.

Una de las ventajas de este método es que una muestra del gradiente puede desplazarse diversas posiciones, dentro de estas subregiones de 4x4 muestras, y aún contribuir de la misma forma al histograma obtenido. De esta forma el descriptor es robusto ante posibles desplazamientos locales de la posición de los puntos característicos.

El descriptor se forma concatenando los 16 histogramas de 8 direcciones de manera que el vector de características tendrá una longitud de 4x4x8 = 128 elementos, para cada punto característico.

3. Normalización del vector de características

Como los valores de los gradientes se calculan a partir de diferencias entre niveles de gris de los píxeles, el descriptor obtenido es invariante a cambios afines en la iluminación. Sin embargo, pueden darse cambios no lineales en la iluminación debidos a la saturación de la cámara o producidos como consecuencia de los cambios en la reflexión sobre superficies 3D que presenten diferentes orientaciones.

Para evitar los problemas anteriores y también que medidas locales de alto contraste afecten en exceso al descriptor, se realiza la siguiente normalización en dos pasos:

1. Se *normaliza* el vector *a longitud unidad*.

   Dado el vector de 128 elementos obtenido para un punto característico, se obtiene la longitud de dicho vector, calculado como su norma, y luego se divide cada uno de los elementos de dicho vector por dicha magnitud. De esta forma el vector pasará a tener longitud unidad.

$$\hat{u} = \frac{u}{||u||}, \tag{8.10}$$

2. Umbralización de cada valor a 0.2 y nueva normalización

   Para reducir la influencia de grandes gradientes, cada uno de los elementos del vector de características se umbraliza de manera que ninguno tenga un valor superior a 0.2. Después, el vector es normalizado de nuevo a longitud unidad.

Ejemplo 8.1. *Dada la siguiente región de una imagen, calcular el descriptor SIFT para los 16x16 valores del interior de la línea de trazo grueso de la figura 8.3.*

En este ejercicio se parte de que el punto característico se localizó en el centro de la región indicada. Se asume también que cada una de las muestras corresponde con un píxel y se realizan algunas simplificaciones sobre el método propuesto por Lowe que se irán comentando a lo largo del ejemplo.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	0	5	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0
2	0	0	0	0	5	0	0	0	0	0	5	0	0	0	0	5	0	0	0	0
3	5	5	5	5	0	5	5	10	10	10	0	5	5	10	10	0	5	5	5	0
4	0	0	5	10	5	10	15	0	0	0	5	10	15	0	0	5	0	0	10	5
5	5	10	5	5	5	10	20	15	15	15	5	10	20	15	15	5	10	10	5	0
6	5	10	10	15	20	25	200	25	25	25	20	25	200	25	25	20	10	10	15	5
7	0	0	5	5	10	130	150	150	150	20	10	130	150	150	150	10	0	0	5	0
8	0	0	5	5	5	120	140	140	140	10	5	120	140	140	140	5	0	0	5	0
9	5	10	5	5	5	100	100	100	100	10	5	100	100	100	100	5	10	10	5	0
10	0	5	5	10	100	120	250	250	130	130	100	120	250	190	130	100	5	5	10	0
11	5	5	10	10	120	130	250	250	120	120	120	130	250	250	120	120	5	5	10	0
12	5	10	5	100	160	200	150	250	110	110	160	200	150	250	110	160	100	10	5	0
13	0	0	5	130	180	200	210	120	120	255	255	200	210	120	120	180	130	0	5	0
14	5	10	0	0	190	210	200	120	120	255	255	210	200	120	120	190	10	10	0	0
15	0	0	0	0	200	180	190	130	130	130	200	180	190	130	130	200	0	0	0	10
16	0	5	10	10	200	190	170	140	140	140	200	190	170	140	140	200	5	5	10	0
17	0	5	5	5	10	0	5	0	0	0	10	0	5	0	0	10	5	5	5	0
18	0	0	10	10	0	5	0	0	0	0	0	5	0	0	0	0	0	0	10	0
19	5	0	0	0	5	0	0	10	10	10	5	0	0	10	10	5	0	0	0	0
20	0	0	0	0	5	0	0	10	10	10	5	0	0	10	10	5	0	0	0	0
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Figura 8.3. Región de la imagen sobre la que se calculará el descriptor SIFT

Solución: Como se explicó previamente el cálculo del descriptor se basa en (a) la obtención de la magnitud y la orientación del gradiente de la matriz 16x16 valores que rodean al punto característico, (b) el posterior cálculo del histograma de 8 bins con las orientaciones del gradiente para cada una de las 16 subregiones en las que se subdividen dichos 16x16 valores, y (c) la normalización del vector obtenido.

a) Obtención de la magnitud y la orientación del gradiente

En este primer paso, se asume que se ha seleccionado la escala de la imagen más similar a la escala del punto característico. Posteriormente se obtienen las derivadas horizontal y vertical. Se asume también que este cálculo se realizó previamente, en la etapa de *asignación de la orientación* del punto característico. En ese momento, se calculó la magnitud del gradiente, $m(x,y)$, y también su orientación, $\theta(x,y)$, utilizando diferencias entre los píxeles vecinos. Para facilitar el seguimiento de los cálculos, se ha omitido aplicar sobre la imagen de la figura 8.3 el filtrado gaussiano previo.

En primer lugar se calculan las derivadas horizontal y vertical, para cada píxel.

Derivada horizontal

L(x+1,y)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	0	5	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0
2	0	0	0	0	5	0	0	0	0	0	5	0	0	0	0	5	0	0	0	0
3	5	5	5	0	5	5	10	10	10	0	5	5	10	10	0	5	5	5	5	0
4	0	0	10	5	10	15	0	0	0	5	10	15	0	0	5	0	0	10	10	5
5	5	10	5	5	10	20	15	15	15	5	10	20	15	15	5	10	10	5	5	0
6	5	10	15	20	25	200	25	25	25	20	25	200	25	25	20	10	10	15	15	5
7	0	0	5	10	130	150	150	150	20	10	130	150	150	150	10	0	0	5	5	0
8	0	0	5	5	120	140	140	140	10	5	120	140	140	140	5	0	0	5	5	0
9	5	10	5	5	100	100	100	100	10	5	100	100	100	100	5	10	10	5	5	0
10	0	5	10	100	120	250	250	130	130	100	120	250	190	130	100	5	5	10	10	0
11	5	5	10	120	130	250	250	120	120	120	130	250	250	120	120	5	5	10	10	0
12	5	10	100	160	200	150	250	110	110	160	200	150	250	110	160	100	10	5	5	0
13	0	0	130	180	200	210	120	120	255	255	200	210	120	120	180	130	0	5	5	0
14	5	10	0	190	210	200	120	120	255	255	210	200	120	120	190	10	10	0	0	0
15	0	0	0	200	180	190	130	130	130	200	180	190	130	130	200	0	0	0	0	10
16	0	5	10	200	190	170	140	140	140	200	190	170	140	140	200	5	5	10	10	0
17	0	5	5	10	0	5	0	0	0	10	0	5	0	0	10	5	5	5	5	0
18	0	0	10	0	5	0	0	0	0	0	5	0	0	0	0	0	0	10	10	0
19	5	0	0	0	5	0	0	10	10	10	5	0	0	10	10	5	0	0	0	0
20	0	0	0	0	5	0	0	10	10	10	5	0	0	10	10	5	0	0	0	0
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

L(x-1,y)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	0	5	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0
2	0	0	0	0	5	0	0	0	0	0	5	0	0	0	0	5	0	0	0	0
3	5	5	5	5	5	0	5	5	10	10	10	0	5	5	10	10	0	5	5	0
4	0	0	0	5	10	5	10	15	0	0	0	5	10	15	0	0	5	0	10	5
5	5	10	10	5	5	5	10	20	15	15	15	5	10	20	15	15	5	10	5	0
6	5	10	10	10	15	20	25	200	25	25	25	20	25	200	25	25	20	10	15	5
7	0	0	0	5	5	10	130	150	150	150	20	10	130	150	150	150	10	0	5	0
8	0	0	0	5	5	5	120	140	140	140	10	5	120	140	140	140	5	0	5	0
9	5	10	10	5	5	5	100	100	100	100	10	5	100	100	100	100	5	10	5	0
10	0	5	5	5	10	100	120	250	250	130	130	100	120	250	190	130	100	5	10	0
11	5	5	5	10	10	120	130	250	250	120	120	120	130	250	250	120	120	5	10	0
12	5	10	10	5	100	160	200	150	250	110	110	160	200	150	250	110	160	100	5	0
13	0	0	0	5	130	180	200	210	120	120	255	255	200	210	120	120	180	130	5	0
14	5	10	10	0	0	190	210	200	120	120	255	255	210	200	120	120	190	10	0	0
15	0	0	0	0	0	200	180	190	130	130	130	200	180	190	130	130	200	0	0	10
16	0	5	5	10	10	200	190	170	140	140	140	200	190	170	140	140	200	5	10	0
17	0	5	5	5	5	10	0	5	0	0	0	10	0	5	0	0	10	5	5	0
18	0	0	0	10	10	0	5	0	0	0	0	0	5	0	0	0	0	0	10	0
19	5	0	0	0	5	0	0	10	10	10	5	0	0	10	10	5	0	0	0	0
20	0	0	0	0	5	0	0	10	10	10	5	0	0	10	10	5	0	0	0	0
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

L(x+1,y) -
L(x-1,y)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	0	5	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0
2	0	0	0	0	5	0	0	0	0	0	5	0	0	0	0	5	0	0	0	0
3	5	5	0	-5	0	5	5	5	0	-10	-5	5	5	5	-10	-5	5	0	5	0
4	0	0	10	0	0	10	-10	-15	0	5	10	10	-10	-15	5	0	-5	10	10	5
5	5	10	-5	0	5	15	5	-5	0	-10	-5	15	5	-5	-10	-5	5	-5	5	0
6	5	10	5	10	10	180	0	-175	0	-5	0	180	0	-175	-5	-15	-10	5	15	5
7	0	0	5	5	125	140	20	0	-130	-140	110	140	20	0	-140	-150	-10	5	5	0
8	0	0	5	0	115	135	20	0	-130	-135	110	135	20	0	-135	-140	-5	5	5	0
9	5	10	-5	0	95	95	0	0	-90	-95	90	95	0	0	-95	-90	5	-5	5	0
10	0	5	5	95	110	150	130	-120	-120	-30	-10	150	70	-120	-90	-125	-95	5	10	0
11	5	5	5	110	120	130	120	-130	-130	0	10	130	120	-130	-130	-115	-115	5	10	0
12	5	10	90	155	100	-10	50	-40	-140	50	90	-10	50	-40	-90	-10	-150	-95	5	0
13	0	0	130	175	70	30	-80	-90	135	135	-55	-45	-80	-90	60	10	-180	-125	5	0
14	5	10	-10	190	210	10	-90	-80	135	135	-45	-55	-90	-80	70	-110	-180	-10	0	0
15	0	0	0	200	180	-10	-50	-60	0	70	50	-10	-50	-60	70	-130	-200	0	0	10
16	0	5	5	190	180	-30	-50	-30	0	60	50	-30	-50	-30	60	-135	-195	5	10	0
17	0	5	0	5	-5	-5	0	-5	0	10	0	-5	0	-5	10	5	-5	0	5	0
18	0	0	10	-10	-5	0	-5	0	0	0	5	0	-5	0	0	0	0	10	10	0
19	5	0	0	0	5	0	0	10	10	10	5	0	0	10	10	5	0	0	0	0
20	0	0	0	0	5	0	0	10	10	10	5	0	0	10	10	5	0	0	0	0
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Figura 8.4. Obtención de la derivada horizontal como diferencias entre los píxeles vecinos derecho e izquierdo.

Derivada vertical

L(x,y+1)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	0	5	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0
2	0	0	0	0	5	0	0	0	0	0	5	0	0	0	0	5	0	0	0	0
3	5	5	0	0	5	0	0	0	0	0	5	0	0	0	0	5	0	0	5	0
4	0	0	5	5	0	5	5	10	10	10	0	5	5	10	10	0	5	5	10	5
5	5	10	5	10	5	10	15	0	0	0	5	10	15	0	0	5	0	0	5	0
6	5	10	5	5	5	10	20	15	15	15	5	10	20	15	15	5	10	10	15	5
7	0	0	10	15	20	25	200	25	25	25	20	25	200	25	25	20	10	10	5	0
8	0	0	5	5	10	130	150	150	150	20	10	130	150	150	150	10	0	0	5	0
9	5	10	5	5	5	120	140	140	140	10	5	120	140	140	140	5	0	0	5	0
10	0	5	5	5	5	100	100	100	100	10	5	100	100	100	100	5	10	10	10	0
11	5	5	5	10	100	120	250	250	130	130	100	120	250	250	130	100	5	5	10	0
12	5	10	10	10	120	130	250	250	120	120	120	130	250	250	120	120	5	5	5	0
13	0	0	5	100	160	200	150	250	110	110	160	200	150	250	110	160	100	10	5	0
14	5	10	5	130	180	200	210	120	120	255	255	200	210	120	120	180	130	0	0	0
15	0	0	0	0	190	210	200	120	120	255	255	210	200	120	120	190	10	10	0	10
16	0	5	0	0	200	180	190	130	130	130	200	180	190	130	130	200	0	0	10	0
17	0	5	10	10	200	190	170	140	140	140	200	190	170	140	140	200	5	5	5	0
18	0	0	5	5	10	0	5	0	0	0	10	0	5	0	0	10	5	5	10	0
19	5	0	0	0	5	0	0	10	10	10	5	0	0	10	10	5	0	0	0	0
20	0	0	0	0	5	0	0	10	10	10	5	0	0	10	10	5	0	0	0	0
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

L(x,y-1)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	0	5	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0
2	0	0	0	0	5	0	0	0	0	0	5	0	0	0	0	5	0	0	0	0
3	5	5	5	10	5	10	15	0	0	0	5	10	15	0	0	5	0	0	5	0
4	0	0	5	5	5	10	20	15	15	15	5	10	20	15	15	5	10	10	10	5
5	5	10	10	15	20	25	200	25	25	25	20	25	200	25	25	20	10	10	5	0
6	5	10	5	5	10	130	150	150	150	20	10	130	150	150	150	10	0	0	15	5
7	0	0	5	5	5	120	140	140	140	10	5	120	140	140	140	5	0	0	5	0
8	0	0	5	5	5	100	100	100	100	10	5	100	100	100	100	5	10	10	5	0
9	5	10	5	10	100	120	250	250	130	130	100	120	250	250	130	100	5	5	5	0
10	0	5	10	10	120	130	250	250	120	120	120	130	250	250	120	120	5	5	10	0
11	5	5	5	100	160	200	150	250	110	110	160	200	150	250	110	160	100	10	10	0
12	5	10	5	130	180	200	210	120	120	255	255	200	210	120	120	180	130	0	5	0
13	0	0	0	0	190	210	200	120	120	255	255	210	200	120	120	190	10	10	5	0
14	5	10	0	0	200	180	190	130	130	130	200	180	190	130	130	200	0	0	0	0
15	0	0	10	10	200	190	170	140	140	140	200	190	170	140	140	200	5	5	0	10
16	0	5	5	5	10	0	5	0	0	0	10	0	5	0	0	10	5	5	10	0
17	0	5	10	10	0	5	0	0	0	0	0	5	0	0	0	0	0	0	5	0
18	0	0	0	0	5	0	0	10	10	10	5	0	0	10	10	5	0	0	10	0
19	5	0	0	0	5	0	0	10	10	10	5	0	0	10	10	5	0	0	0	0
20	0	0	0	0	5	0	0	10	10	10	5	0	0	10	10	5	0	0	0	0
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

L(x,y+1) -
L(x,y-1)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	0	5	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0
2	0	0	0	0	5	0	0	0	0	0	5	0	0	0	0	5	0	0	0	0
3	5	5	-5	-10	0	-10	-15	0	0	0	0	-10	-15	0	0	0	0	0	5	0
4	0	0	0	0	-5	-5	-15	-5	-5	-5	-5	-5	-15	-5	-5	-5	-5	-5	10	5
5	5	10	-5	-5	-15	-15	-185	-25	-25	-25	-15	-15	-185	-25	-25	-15	-10	-10	5	0
6	5	10	0	0	-5	-120	-130	-135	-135	-5	-5	-120	-130	-135	-135	-5	10	10	15	5
7	0	0	5	10	15	-95	60	-115	-115	15	15	-95	60	-115	-115	15	10	10	5	0
8	0	0	0	0	5	30	50	50	50	10	5	30	50	50	50	5	-10	-10	5	0
9	5	10	0	-5	-95	0	-110	-110	10	-120	-95	0	-110	-110	10	-95	-5	-5	5	0
10	0	5	-5	-5	-115	-30	-150	-150	-20	-110	-115	-30	-150	-150	-20	-115	5	5	10	0
11	5	5	0	-90	-60	-80	100	0	20	20	-60	-80	100	0	20	-60	-95	-5	10	0
12	5	10	5	-120	-60	-70	40	130	0	-135	-135	-70	40	130	0	-60	-125	5	5	0
13	0	0	5	100	-30	-10	-50	130	-10	-145	-95	-10	-50	130	-10	-30	90	0	5	0
14	5	10	5	130	-20	20	20	-10	-10	125	55	20	20	-10	-10	-20	130	0	0	0
15	0	0	-10	-10	-10	20	30	-20	-20	115	55	20	30	-20	-20	-10	5	5	0	10
16	0	5	-5	-5	190	180	185	130	130	130	190	180	185	130	130	190	-5	-5	10	0
17	0	5	0	0	200	185	170	140	140	140	200	185	170	140	140	200	5	5	5	0
18	0	0	5	5	5	0	5	-10	-10	-10	5	0	5	-10	-10	5	5	5	10	0
19	5	0	0	0	5	0	0	10	10	10	5	0	0	10	10	5	0	0	0	0
20	0	0	0	0	5	0	0	10	10	10	5	0	0	10	10	5	0	0	0	0
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Figura 8.5. Obtención de la derivada vertical como diferencias entre los píxeles vecinos superior e inferior.

Una vez obtenidas las derivadas se obtiene la magnitud y la orientación del gradiente mediante las fórmulas indicadas en la ecuación 8.9.

Magnitud del gradiente $(L(x+1,y)-L(x-1,y))^2$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	0	5	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0
2	0	0	0	0	5	0	0	0	0	0	5	0	0	0	0	5	0	0	0	0
3	5	5	0	25	0	25	25	25	0	100	25	25	25	25	100	25	25	0	5	0
4	0	0	100	0	0	100	100	225	0	25	100	100	100	225	25	0	25	100	10	5
5	5	10	25	0	25	225	25	25	0	100	25	225	25	25	100	25	25	25	5	0
6	5	10	25	100	100	32400	0	30625	0	25	0	32400	0	30625	25	225	100	25	15	5
7	0	0	25	25	15625	19600	400	0	16900	19600	12100	19600	400	0	19600	22500	100	25	5	0
8	0	0	25	0	13225	18225	400	0	16900	18225	12100	18225	400	0	18225	19600	25	25	5	0
9	5	10	25	0	9025	9025	0	0	8100	9025	8100	9025	0	0	9025	8100	25	25	5	0
10	0	5	25	9025	12100	22500	16900	14400	14400	900	100	22500	4900	14400	8100	15625	9025	25	10	0
11	5	5	25	12100	14400	16900	14400	16900	16900	0	100	16900	14400	16900	16900	13225	13225	25	10	0
12	5	10	8100	24025	10000	100	2500	1600	19600	2500	8100	100	2500	1600	8100	100	22500	9025	5	0
13	0	0	16900	30625	4900	900	6400	8100	18225	18225	3025	2025	6400	8100	3600	100	32400	15625	5	0
14	5	10	100	36100	44100	100	8100	6400	18225	18225	2025	3025	8100	6400	4900	12100	32400	100	0	0
15	0	0	0	40000	32400	100	2500	3600	0	4900	2500	100	2500	3600	4900	16900	40000	0	0	10
16	0	5	25	36100	32400	900	2500	900	0	3600	2500	900	2500	900	3600	18225	38025	25	10	0
17	0	5	0	25	25	25	0	25	0	100	0	25	0	25	100	25	25	0	5	0
18	0	0	100	100	25	0	25	0	0	0	25	0	25	0	0	0	0	100	10	0
19	5	0	0	0	5	0	0	10	10	10	5	0	0	10	10	5	0	0	0	0
20	0	0	0	0	5	0	0	10	10	10	5	0	0	10	10	5	0	0	0	0
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

$$(L(x,y+1)-L(x,y-1))^2$$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	0	5	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0
2	0	0	0	0	5	0	0	0	0	0	5	0	0	0	0	5	0	0	0	0
3	5	5	25	100	0	100	225	0	0	0	0	100	225	0	0	0	0	0	5	0
4	0	0	0	0	25	25	225	25	25	25	25	25	225	25	25	25	25	25	10	5
5	5	10	25	25	225	225	34225	625	625	625	225	225	34225	625	625	225	100	100	5	0
6	5	10	0	0	25	14400	16900	18225	18225	25	25	14400	16900	18225	18225	25	100	100	15	5
7	0	0	25	100	225	9025	3600	13225	13225	225	225	9025	3600	13225	13225	225	100	100	5	0
8	0	0	0	0	25	900	2500	2500	2500	100	25	900	2500	2500	2500	25	100	100	5	0
9	5	10	0	25	9025	0	12100	12100	100	14400	9025	0	12100	12100	100	9025	25	25	5	0
10	0	5	25	25	13225	900	22500	22500	400	12100	13225	900	22500	22500	400	13225	25	25	10	0
11	5	5	0	8100	3600	6400	10000	0	400	400	3600	6400	10000	0	400	3600	9025	25	10	0
12	5	10	25	14400	3600	4900	1600	16900	0	18225	18225	4900	1600	16900	0	3600	15625	25	5	0
13	0	0	25	10000	900	100	2500	16900	100	21025	9025	100	2500	16900	100	900	8100	0	5	0
14	5	10	25	16900	400	400	400	100	100	15625	3025	400	400	100	100	400	16900	0	0	0
15	0	0	100	100	100	400	900	400	400	13225	3025	400	900	400	400	100	25	25	0	10
16	0	5	25	25	36100	32400	34225	16900	16900	16900	36100	32400	34225	16900	16900	36100	25	25	10	0
17	0	5	0	0	40000	34225	28900	19600	19600	19600	40000	34225	28900	19600	19600	40000	25	25	5	0
18	0	0	25	25	25	0	25	100	100	100	25	0	25	100	100	25	25	25	10	0
19	5	0	0	0	5	0	0	10	10	10	5	0	0	10	10	5	0	0	0	0
20	0	0	0	0	5	0	0	10	10	10	5	0	0	10	10	5	0	0	0	0
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

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

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0,0	0,0	5,0	5,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	5,0	0,0
2	0,0	0,0	0,0	0,0	5,0	0,0	0,0	0,0	0,0	0,0	5,0	0,0	0,0	0,0	0,0	5,0	0,0	0,0	0,0	0,0
3	5,0	5,0	5,0	11,2	0,0	11,2	15,8	5,0	0,0	10,0	5,0	11,2	15,8	5,0	10,0	5,0	5,0	0,0	5,0	0,0
4	0,0	0,0	10,0	0,0	5,0	11,2	18,0	15,8	5,0	7,1	11,2	11,2	18,0	15,8	7,1	5,0	7,1	11,2	10,0	5,0
5	5,0	10,0	7,1	5,0	15,8	21,2	185,1	25,5	25,0	26,9	15,8	21,2	185,1	25,5	26,9	15,8	11,2	11,2	5,0	0,0
6	5,0	10,0	5,0	10,0	11,2	216,3	130,0	221,0	135,0	7,1	5,0	216,3	130,0	221,0	135,1	15,8	14,1	11,2	15,0	5,0
7	0,0	0,0	7,1	11,2	125,9	169,2	63,2	115,0	173,6	140,8	111,0	169,2	63,2	115,0	181,2	150,7	14,1	11,2	5,0	0,0
8	0,0	0,0	5,0	0,0	115,1	138,3	53,9	50,0	139,3	135,4	110,1	138,3	53,9	50,0	144,0	140,1	11,2	11,2	5,0	0,0
9	5,0	10,0	5,0	5,0	134,4	95,0	110,0	110,0	90,6	153,1	130,9	95,0	110,0	110,0	95,5	130,9	7,1	7,1	5,0	0,0
10	0,0	5,0	7,1	95,1	159,1	153,0	198,5	192,1	121,7	114,0	115,4	153,0	165,5	192,1	92,2	169,9	95,1	7,1	10,0	0,0
11	5,0	5,0	5,0	142,1	134,2	152,6	156,2	130,0	131,5	20,0	60,8	152,6	156,2	130,0	131,5	129,7	149,2	7,1	10,0	0,0
12	5,0	10,0	90,1	196,0	116,6	70,7	64,0	136,0	140,0	144,0	162,2	70,7	64,0	136,0	90,0	60,8	195,3	95,1	5,0	0,0
13	0,0	0,0	130,1	201,6	76,2	31,6	94,3	158,1	135,4	198,1	109,8	46,1	94,3	158,1	60,8	31,6	201,2	125,0	5,0	0,0
14	5,0	10,0	11,2	230,2	211,0	22,4	92,2	80,6	135,4	184,0	71,1	58,5	92,2	80,6	70,7	111,8	222,0	10,0	0,0	0,0
15	0,0	0,0	10,0	200,2	180,3	22,4	58,3	63,2	20,0	134,6	74,3	22,4	58,3	63,2	72,8	130,4	200,1	5,0	0,0	10,0
16	0,0	5,0	7,1	190,1	261,7	182,5	191,6	133,4	130,0	143,2	196,5	182,5	191,6	133,4	143,2	233,1	195,1	7,1	10,0	0,0
17	0,0	5,0	0,0	5,0	200,1	185,1	170,0	140,1	140,0	140,4	200,0	185,1	170,0	140,1	140,4	200,1	7,1	5,0	5,0	0,0
18	0,0	0,0	11,2	11,2	7,1	0,0	7,1	10,0	10,0	10,0	7,1	0,0	7,1	10,0	10,0	5,0	5,0	11,2	10,0	0,0
19	5,0	0,0	0,0	0,0	5,0	0,0	0,0	10,0	10,0	10,0	5,0	0,0	0,0	10,0	10,0	5,0	0,0	0,0	0,0	0,0
20	0,0	0,0	0,0	0,0	5,0	0,0	0,0	10,0	10,0	10,0	5,0	0,0	0,0	10,0	10,0	5,0	0,0	0,0	0,0	0,0
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Figura 8.6. Obtención de la magnitud del gradiente para cada píxel de la región de interés.

Los valores obtenidos para la magnitud del gradiente deben ponderarse por una gaussiana que tenga una σ igual a la mitad del ancho de la ventana del descriptor. Para simplificar los cálculos y facilitar la

interpretación de los valores obtenidos en el cálculo del gradiente, en este ejercicio no realizaremos dicha ponderación.

Orientación del gradiente

A partir de las derivadas verticales y horizontales se obtiene, y se muestra en la figura 8.7, la orientación en cada píxel. Esta orientación se calculó como el arco cuya tangente es la relación entre la derivada vertical y la derivada horizontal, utilizando los valores calculados previamente mediante la fórmula expresada en la ecuación 9, que recordamos era la siguiente:

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

Orientación en grados, entre 180 y -180, calculado a partir del arco tangente.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	0	5	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0
2	0	0	0	0	5	0	0	0	0	0	5	0	0	0	0	5	0	0	0	0
3	5	5	-90	-117	0	-63	-72	0	0	0	0	-63	-72	0	0	0	0	0	5	0
4	0	0	0	0	-90	-27	-124	-162	-90	-45	-27	-27	-124	-162	-45	-90	-135	-27	10	5
5	5	10	-135	-90	-72	-45	-88	-101	-90	-112	-108	-45	-88	-101	-112	-108	-63	-117	5	0
6	5	10	0	0	-27	-34	-90	-142	-90	-135	-90	-34	-90	-142	-92	-162	135	63	15	5
7	0	0	45	63	7	-34	72	-90	-139	174	8	-34	72	-90	-141	174	135	63	5	0
8	0	0	0	0	2	13	68	90	159	176	3	13	68	90	160	178	-117	-63	5	0
9	5	10	0	-90	-45	0	-90	-90	174	-128	-47	0	-90	-90	174	-133	-45	-135	5	0
10	0	5	-45	-3	-46	-11	-49	-129	-171	-105	-95	-11	-65	-129	-167	-137	177	45	10	0
11	5	5	0	-39	-27	-32	40	0	171	90	-81	-32	40	0	171	-152	-140	-45	10	0
12	5	10	3	-38	-31	-98	39	107	0	-70	-56	-98	39	107	0	-99	-140	177	5	0
13	0	0	2	30	-23	-18	-148	125	-4	-47	-120	-167	-148	125	-9	-72	153	0	5	0
14	5	10	153	34	-5	63	167	-173	-4	43	129	160	167	-173	-8	-170	144	0	0	0
15	0	0	-90	-3	-3	117	149	-162	-90	59	48	117	149	-162	-16	-176	179	90	0	10
16	0	5	-45	-2	47	99	105	103	90	65	75	99	105	103	65	125	-179	-45	10	0
17	0	5	0	0	91	92	90	92	90	86	90	92	90	92	86	89	135	90	5	0
18	0	0	27	153	135	0	135	-90	-90	-90	45	0	135	-90	-90	90	90	27	10	0
19	5	0	0	0	5	0	0	10	10	10	5	0	0	10	10	5	0	0	0	0
20	0	0	0	0	5	0	0	10	10	10	5	0	0	10	10	5	0	0	0	0
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Orientación expresada en grados, entre 0 y 360

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	0	5	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0
2	0	0	0	0	5	0	0	0	0	0	5	0	0	0	0	5	0	0	0	0
3	5	5	270	243	0	297	288	0	0	0	0	297	288	0	0	0	0	0	5	0
4	0	0	0	0	270	333	236	198	270	315	333	333	236	198	315	270	225	333	10	5
5	5	10	225	270	288	315	272	259	270	248	252	315	272	259	248	252	297	243	5	0
6	5	10	0	0	333	326	270	218	270	225	270	326	270	218	268	198	135	63	15	5
7	0	0	45	63	7	326	72	270	221	174	8	326	72	270	219	174	135	63	5	0
8	0	0	0	0	2	13	68	90	159	176	3	13	68	90	160	178	243	297	5	0
9	5	10	0	270	315	0	270	270	174	232	313	0	270	270	174	227	315	225	5	0
10	0	5	315	357	314	349	311	231	189	255	265	349	295	231	193	223	177	45	10	0
11	5	5	0	321	333	328	40	0	171	90	279	328	40	0	171	208	220	315	10	0
12	5	10	3	322	329	262	39	107	0	290	304	262	39	107	0	261	220	177	5	0
13	0	0	2	30	337	342	212	125	356	313	240	193	212	125	351	288	153	0	5	0
14	5	10	153	34	355	63	167	187	356	43	129	160	167	187	352	190	144	0	0	0
15	0	0	270	357	357	117	149	198	270	59	48	117	149	198	344	184	179	90	0	10
16	0	5	315	358	47	99	105	103	90	65	75	99	105	103	65	125	181	315	10	0
17	0	5	0	0	91	92	90	92	90	86	90	92	90	92	86	89	135	90	5	0
18	0	0	27	153	135	0	135	270	270	270	45	0	135	270	270	90	90	27	10	0
19	5	0	0	0	5	0	0	10	10	10	5	0	0	10	10	5	0	0	0	0
20	0	0	0	0	5	0	0	10	10	10	5	0	0	10	10	5	0	0	0	0
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Figura 8.7. Obtención de la orientación del gradiente para cada píxel de la región de interés.

b) Histograma de orientaciones para las 4x4 subregiones

Codificación de las orientaciones del gradiente

Para el cálculo del descriptor, las orientaciones del gradiente se codifican en ocho direcciones de manera que si la orientación se encuentra entre 0 y 44 grados se codifica como 1, entre 45 y 89 como 2 y así sucesivamente. La figura 8.8 presenta el resultado de aplicar dicha codificación a las orientaciones obtenidas previamente que se muestran en la figura 8.7.

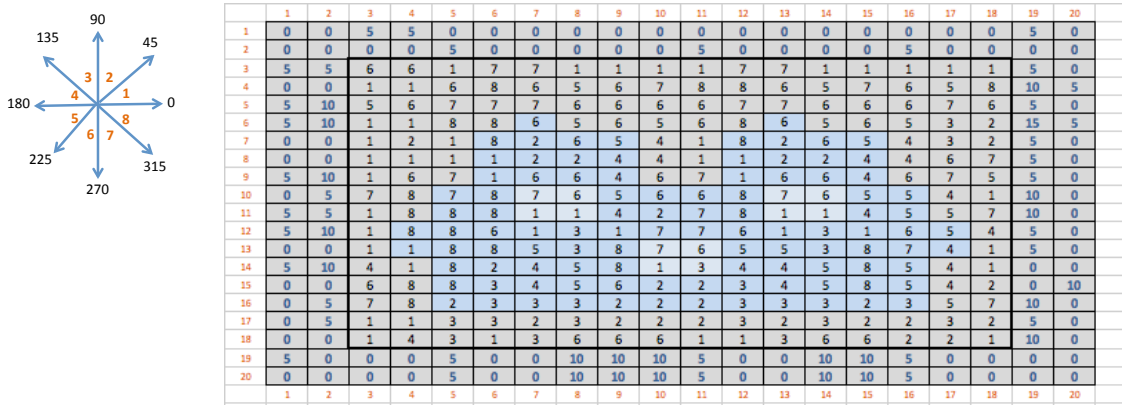


Figura 8.8. Codificación de las orientaciones del gradiente en 8 direcciones.

División de la región de interés en 4x4 subregiones

El descriptor SIFT se forma concatenando los histogramas de orientaciones obtenidos al dividir la región de interés en 4x4 subregiones. En la figura 8.9 se muestra esta subdivisión tanto para los valores de la magnitud del gradiente (a), como para su orientación (b).

(a)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0,0	0,0	5,0	5,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	5,0	0,0
2	0,0	0,0	0,0	0,0	5,0	0,0	0,0	0,0	0,0	0,0	5,0	0,0	0,0	0,0	0,0	5,0	0,0	0,0	0,0	0,0
3	5,0	5,0	5,0	11,2	0,0	11,2	15,8	5,0	0,0	10,0	5,0	11,2	15,8	5,0	10,0	5,0	5,0	0,0	5,0	0,0
4	0,0	0,0	10,0	0,0	5,0	11,2	18,0	15,8	5,0	7,1	11,2	11,2	18,0	15,8	7,1	5,0	7,1	11,2	10,0	5,0
5	5,0	10,0	7,1	5,0	15,8	21,2	185,1	25,5	25,0	26,9	15,8	21,2	185,1	25,5	26,9	15,8	11,2	11,2	5,0	0,0
6	5,0	10,0	5,0	10,0	11,2	216,3	130,0	221,0	135,0	7,1	5,0	216,3	130,0	221,0	135,1	15,8	14,1	11,2	15,0	5,0
7	0,0	0,0	7,1	11,2	125,9	169,2	63,2	115,0	173,6	140,8	111,0	169,2	63,2	115,0	181,2	150,7	14,1	11,2	5,0	0,0
8	0,0	0,0	5,0	0,0	115,1	138,3	53,9	50,0	139,3	135,4	110,1	138,3	53,9	50,0	144,0	140,1	11,2	11,2	5,0	0,0
9	5,0	10,0	5,0	5,0	134,4	95,0	110,0	110,0	90,6	153,1	130,9	95,0	110,0	110,0	95,5	130,9	7,1	7,1	5,0	0,0
10	0,0	5,0	7,1	95,1	159,1	153,0	198,5	192,1	121,7	114,0	115,4	153,0	165,5	192,1	92,2	169,9	95,1	7,1	10,0	0,0
11	5,0	5,0	5,0	142,1	134,2	152,6	156,2	130,0	131,5	20,0	60,8	152,6	156,2	130,0	131,5	129,7	149,2	7,1	10,0	0,0
12	5,0	10,0	90,1	196,0	116,6	70,7	64,0	136,0	140,0	144,0	162,2	70,7	64,0	136,0	90,0	60,8	195,3	95,1	5,0	0,0
13	0,0	0,0	130,1	201,6	76,2	31,6	94,3	158,1	135,4	198,1	109,8	46,1	94,3	158,1	60,8	31,6	201,2	125,0	5,0	0,0
14	5,0	10,0	11,2	230,2	211,0	22,4	92,2	80,6	135,4	184,0	71,1	58,5	92,2	80,6	70,7	111,8	222,0	10,0	0,0	0,0
15	0,0	0,0	10,0	200,2	180,3	22,4	58,3	63,2	20,0	134,6	74,3	22,4	58,3	63,2	72,8	130,4	200,1	5,0	0,0	10,0
16	0,0	5,0	7,1	190,1	261,7	182,5	191,6	133,4	130,0	143,2	196,5	182,5	191,6	133,4	143,2	233,1	195,1	7,1	10,0	0,0
17	0,0	5,0	0,0	5,0	200,1	185,1	170,0	140,1	140,0	140,4	200,0	185,1	170,0	140,1	140,4	200,1	7,1	5,0	5,0	0,0
18	0,0	0,0	11,2	11,2	7,1	0,0	7,1	10,0	10,0	10,0	7,1	0,0	7,1	10,0	10,0	5,0	5,0	11,2	10,0	0,0
19	5,0	0,0	0,0	0,0	5,0	0,0	0,0	10,0	10,0	10,0	5,0	0,0	0,0	10,0	10,0	5,0	0,0	0,0	0,0	0,0
20	0,0	0,0	0,0	0,0	5,0	0,0	0,0	10,0	10,0	10,0	5,0	0,0	0,0	10,0	10,0	5,0	0,0	0,0	0,0	0,0
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

(b)

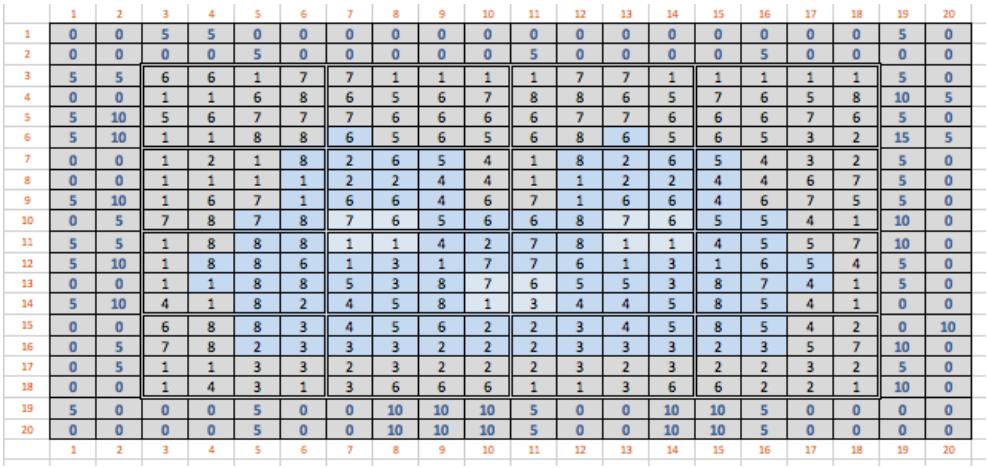


Figura 8.9. División de la región de interés en 4x4 subregiones. (a) Magnitud del gradiente y (b) Orientaciones del gradiente.

Histograma de 8 direcciones para las primeras 4 subregiones

Por claridad y para poder representar la información en el espacio disponible, se obtendrá el vector SIFT únicamente para las 4 primeras regiones de las 16 anteriores.

(a) (b) (c)

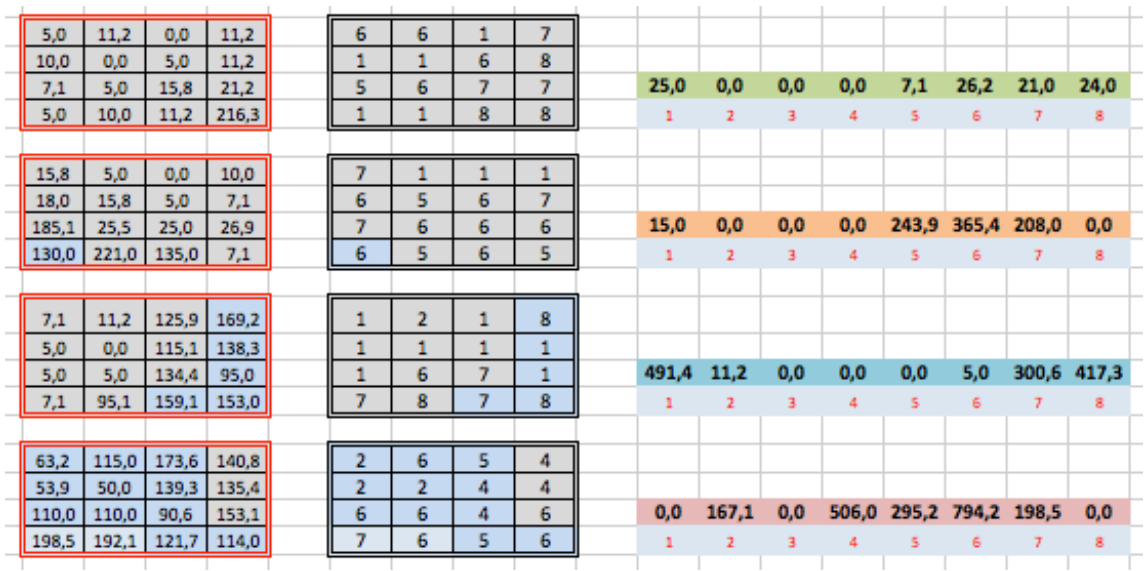


Figura 8.9. Obtención del histograma de orientaciones del gradiente para 4 subregiones. En columnas: (a) valores de la magnitud del gradiente, (b) orientaciones del gradiente, (c) histograma de orientaciones para 8 direcciones.

Como puede verse el histograma acumula para cada orientación la suma de los valores de la magnitud del gradiente de los píxeles con dicha orientación. Si nos fijamos, por ejemplo, en la primera región de la figura 8.9, puede verse que para la orientación 1, que correspondía de 0 a 44 grados, hay 1 píxel en la primera fila, 2 en la segunda y 2 en la cuarta. También puede observarse que el primer elemento del histograma (columna c de la figura 8.9) contiene la suma de los valores de dichas magnitudes que son 0,0 + 10,0 + 0,0 + 5,0 + 10, 0 = 25,0. La distribución del valor de cada gradiente

en los intervalos adyacentes del histograma mediante interpolación tri-lineal también se ha omitido en este ejercicio.

Una vez realizada dicha operación para las 16 subregiones, el vector de características se obtiene concatenando los histogramas para todas ellas. En nuestro caso, al calcularlo solo para 4 subregiones, se obtendría el vector de la figura 8.10.

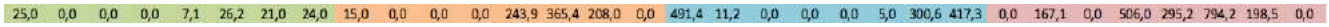


Figura 8.10. Vector de histogramas orientados de gradientes para las primeras 4 subregiones.c) Normalización del vector de características obtenido

Primera normalización a longitud unidad

Para realizar la primera normalización a longitud unidad se eleva cada elemento al cuadrado, se suman todos los elementos y se obtiene la raíz cuadrada, resultando una longitud, para el vector anterior de 1336,48.

Figura 8.11. Primera normalización del vector de histogramas orientados de gradientes para las primeras 4 subregiones.

Segunda normalización a longitud unidad

Para la realizar la segunda normalización, primero se truncan todos los valores a 0,2, como puede verse en la figura 8.11.

Figura 8.12. Segunda normalización del vector de características: valores truncados a 0,2.

Y después se normaliza nuevamente a longitud unidad. En este caso la longitud del vector es 0,6138 por lo que el vector SIFT resultante es el que se muestra en la figura 8.12.

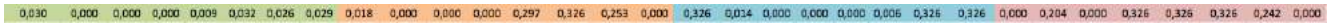


Figura 8.13. Segunda normalización de vector de características: nueva normalización a longitud unidad.

Haciendo este proceso para las 16 subregiones definidas alrededor del punto característico se obtiene el descriptor SIFT de cada keypoint.

8.3. Aplicación de SIFT al reconocimiento de objetos

Una de las principales aplicaciones de los métodos basados en características locales invariantes, como es el caso de SIFT, es el reconocimiento de objetos, y especialmente cuando hay oclusión o el fondo está lleno de objetos desordenados (*clutter*).

Cuando se utiliza este tipo de técnicas, el reconocimiento de un objeto se basa en encontrar correspondencias correctas entre el suficiente número de puntos característicos del objeto "consulta" con uno o varios de los objetos típicamente almacenados en una base de datos.

Inicialmente se realiza la correspondencia entre cada uno de los descriptores de los puntos característicos de la imagen consulta contra la base de datos que contiene todos los descriptores de los puntos característicos extraídos del conjunto de imágenes de entrenamiento. Muchas de las correspondencias iniciales serán erróneas debido a características que son ambiguas o que proceden de un fondo lleno de objetos desordenados. Para mejorar el reconocimiento, se identifican grupos de al menos tres características que sean coherentes con un objeto y su pose, asumiendo que es mucho más probable que estos grupos sean más correctos que las correspondencias obtenidas individualmente. Después, se comprueba cada grupo mediante un ajuste geométrico detallado al modelo, de manera que, en función del resultado de este ajuste, se rechaza o acepta la correspondencia entre los objetos.

8.3.1. Correspondencias de vecinos más cercanos

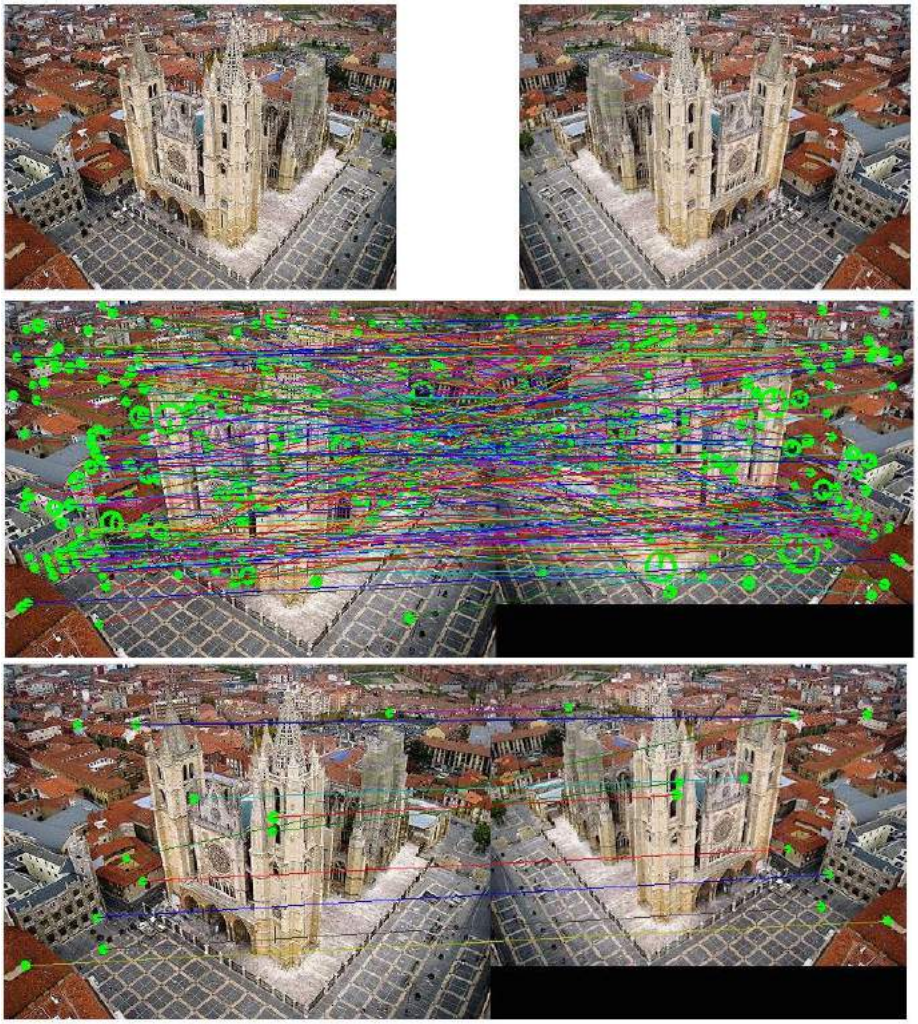
Como comentábamos antes, la mejor correspondencia candidata para cada punto característico viene dada por el vecino más cercano a su descriptor de entre todos los puntos almacenados en la base de datos que contiene los descriptores de las imágenes de entrenamiento. Para SIFT, el vecino más cercano se define como aquel punto característico con una distancia Euclídea mínima entre su vector de características invariante, que es su descriptor, y el de la consulta. Un ejemplo de estas correspondencias puede verse en la figura 8.14.

Para descartar características que no tienen una buena correspondencia Lowe (2004) evaluó la posibilidad de utilizar un umbral global llegando a la conclusión de que no es una buena medida ya que algunos descriptores son más discriminativos que otros. Por ello, propuso utilizar una medida obtenida al comparar la distancia entre el vecino más cercano y el segundo vecino más cercano. En el caso de que en el conjunto de entrenamiento haya más de una imagen del mismo objeto, se define el segundo vecino más cercano como aquel que procede de un objeto diferente al primer vecino, utilizando para ello imágenes de las que se conozca que contienen diferentes objetos. Según Lowe (2004) esta medida funciona bien porque, para que sea fiable, las correspondencias correctas necesitan tener el vecino más cercano significativamente más próximo que la más cercana de las correspondencias indirectas (ver figura 8.14).

Para la implementación realizada, Lowe (2004) rechazó todas las correspondencias en las que el ratio de distancias de la segunda sobre la primera fuera mayor de 0.8. Con eso, se eliminaron el 90% de las correspondencias erróneas, mientras que se descartaron menos del 5% de correspondencias correctas.

8.3.2. Correspondencia entre puntos mediante la aproximación del mejor-recipiente-primero

Ante la falta de algoritmos que pudieran identificar el vecino más cercano en espacios de dimensionalidad alta, que fueran más eficientes que una búsqueda exhaustiva, Lowe (2004) utilizó un algoritmo previamente publicado conjuntamente con Beis (Beis y Lowe, 1997). El algoritmo, llamado Best-Bin-First (BBF), que puede traducirse como el mejor-recipiente-primero, utiliza un ordenamiento modificado de la búsqueda para el algoritmo k-d tree, de manera que en el espacio de características los recipientes, *bins,* se buscan en el orden de la distancia más próxima a la ubicación de la consulta. En el artículo original de Lowe (2004) pueden encontrarse algunos detalles sobre su implementación y una pequeña discusión sobre por qué es apropiado su uso en este problema.



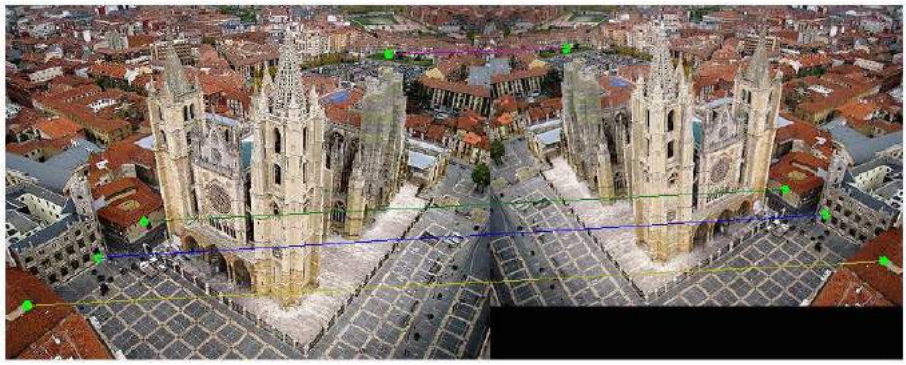


Figura 8.14. Ejemplo de dos imágenes del mismo objeto, la catedral de León, en distintas poses y las correspondencias de descriptores SIFT encontradas. En la primera fila las *imágenes originales*. En la segunda fila aparecen todas las correspondencias encontradas mediante el *primer vecino más cercano*. En la tercera fila se ha aplicado *el test del segundo vecino más cercano* a cada correspondencia. En la cuarta, se ha impuesto una *verificación geométrica* a las correspondencias.

8.3.3. Reconocimiento de objetos mediante la transformada de Hough afin

Para aplicar el descriptor SIFT al reconocimiento de objetos Lowe (2004) propuso utilizar una técnica basada en la transformada de Hough donde se aceptaba como válida una correspondencia cuando al menos tres puntos de un objeto, representados por sus correspondientes descriptores SIFT votaban a favor de ello. Pretendía hacer factible la identificación de objetos con el menor número posible de correspondencias entre características, pensando especialmente en el reconocimiento de objetos pequeños o con mucha oclusión.
La correspondencia de descriptores SIFT se ha establecido como la técnica base del estado del arte para el reconocimiento y recuperación de objetos. En el estudio realizado por Mikolajczyk y Schmid (2005), estos autores llegaron a la conclusión de que el descriptor SIFT es más robusto a deformaciones de la imagen que muchas de las otras técnicas que se estaban usando hasta ese momento, como son los filtros dirigidos, los invariantes diferenciales y complejos o los momentos invariantes, entre otros.

8.4. Extensiones de SIFT

Después de la aparición de SIFT diversos autores lo extendieron de manera que funcionara mejor en problemas para los que no había sido concebido inicialmente, como son las imágenes en color, el muestreo denso, el vídeo o la reducción de su dimensionalidad.

La primera propuesta para aplicar SIFT a imágenes en color la realizó Bosch y otros (2006) que calcularon los descriptores SIFT sobre los tres canales de la imagen tras convertirla al espacio de color HSV, obteniendo así un vector de 3x128 dimensiones. Otra propuesta similar la realizó Van de Weijer y Schmid (2006) quienes concatenaron el descriptor SIFT original con histogramas ponderados del tono o de ángulo oponente, descubriendo que mejoraba las correspondencias de puntos al evaluarlo en diferentes datasets. Otra extensión de SIFT que utiliza el color fue propuesta por Burghouts y Geusebroek (2009). Estos autores definieron un conjuntos de descriptores de imagen que se basan en invariantes de color expresados mediante un modelo de color Gaussiano. Su idea consistió en

reemplazar el gradiente de escala de grises que utiliza SIFT por diversos gradientes de color invariantes a combinaciones de niveles de intensidad local, sombras y zonas iluminadas. Llamaron a su descriptor C-colour-SIFT y comprobaron que funcionaba mejor que los métodos comentados previamente para clasificación de imágenes en categorías y para correspondencia de puntos de interés. En relación con extensiones basadas en color, comentar que uno de los trabajos más completos es el estudio de las propiedades invariantes de diferentes representaciones de color realizado por Koen y sus colegas (Van de Sande y otros, 2010). En este trabajo se estudian representaciones de color basadas en histogramas, momentos e invariantes de color juntamente con descriptores de color basados en SIFT. Su conclusión fue que el descriptor OpponentSIFT es el que mejor rendimiento produce para clasificación de categorías de objetos.

Además de las extensiones que añaden el color a SIFT han aparecido otras como son PCA-SIFT (Ke y Suktahnakar, 2004) que primero utilizan regiones de interés alrededor del keypoint mayores, de 39x39 píxeles, con el objetivo de obtener invarianza a la escala y luego, debido a la alta dimensionalidad, la reducen mediante análisis de componentes principales (PCA) a un vector de 20 dimensiones, siendo, según sus atores, más rápido y distintivo que SIFT.

Y finalmente comentar que una de las extensiones de SIFT más utilizadas ha sido denseSIFT, principalmente para clasificar imágenes en categorías. Este método consiste en calcular el descriptor SIFT sobre una rejilla densa colocada sobre la imagen, en lugar de hacerlo alrededor de los puntos de interés detectados. El motivo para hacerlo es que se considera que de esta forma se obtiene más información que la que se consigue con los mismos descriptores evaluados en los puntos característicos que se encontrarán, sin duda, más dispersos. Este planteamiento lo inició Bosch y otros (2006) y en la actualidad es el procedimiento habitual seguido en problemas relacionados con clasificación de imágenes en categorías.

8.5. Descriptores de imagen relacionados

En los últimos años han aparecido numerosas propuestas de métodos basados total o parcialmente en los conceptos subyacentes en SIFT que son: localización de puntos característicos y descripción de la región de interés, alrededor de dichos puntos, mediante algún método generalmente basado en histogramas de orientaciones del gradiente. En esta sección se comenta brevemente cuatro de los más establecidos, aunque descriptores como FERNS, DAISY, FAST, ORB, BRISK o BRIEF son otras propuestas recientes que se postulan como más rápidas, más eficientes o más apropiadas para diversos problemas.

8.5.1. Histogramas de campo receptivo

Las propuestas de descriptores de la imagen basados en el histograma se inició a principios de los años 90 cuando Swain y Ballard (Swain y Ballard, 1991) mostraron que era posible reconocer objetos si se comparaban histogramas RGB de las imágenes de dichos objetos sin tener en cuenta ningún tipo de relación espacial entre las características de la imagen. Una propuesta posterior (Shiele and Crowley, 2000) confirmó que también era posible hacerlo utilizando tanto histogramas de derivadas

parciales de primer orden, como combinaciones de las respuestas del Laplaciano y magnitudes del gradiente calculadas a múltiples escalas.

Esta propuesta ha sido generalizada aún más por Linde y Lindeberg (Linde y Lindeberg 2004; Linde y Lindeberg 2012) quienes propusieron histogramas de campo receptivo más general a partir de combinaciones de derivadas de Gaussianas o invariantes diferenciales. Posteriormente, realizaron una evaluación exhaustiva de este tipo de descriptores basados en histograma en cuanto a su rendimiento para reconocimiento y clasificación de objetos.

8.5.2. HoG (Histogram of oriented gradients)

HoG, conocido como Histograma de Gradientes Orientados, fue propuesto por Delal y Triggs (Delal y Triggs, 2005) en el CVPR'05 (Computer Vision and Pattern Recognition Conference). El método se basa en la idea de que la apariencia local y la forma de un objeto pueden describirse mediante la distribución de gradientes de intensidad o direcciones de los bordes. La técnica cuenta el número de ocurrencias de una orientación del gradiente en determinadas regiones de la imagen. El método tiene similitudes con SIFT (Lowe, 1999) y con *contextos de forma* (Belongie y otros, 2002) aunque se diferencia de estos métodos en que se calcula en una rejilla densa de celdas espaciadas uniformemente y que utiliza una normalización del contraste local para incrementar la precisión.

Inicialmente esta técnica (Delal y Triggs, 2005) fue propuesta para detectar peatones en imágenes estáticas pero su empleo se ha extendido a la detección de coches y diversos animales comunes, tanto en vídeo como en imágenes estáticas.

8.5.3. SURF (Speed up robust features)

SURF es un descriptor propuesto por Bay y otros (2006) que es similar a SIFT en cuanto que se basa en obtener puntos característicos y posteriormente describir la región alrededor de dichos puntos. Este método se caracteriza porque el espacio escala lo crea mediante wavelets de Haar, en lugar de una pirámide de diferencias de gaussianas. Además, los puntos característicos se detectan con el determinante del Hessiano y la región de interés se describe como sumas, absolutas y no, de derivadas de primer orden, en lugar de un histograma orientado de gradientes.

Si bien se considera que la capacidad de detectar y caracterizar regiones es similar o un poquito inferior a SIFT, es más rápido al basar su detección en wavelets de Haar.

8.5.4. GLOH (Gradient location and orientation histogram)

La técnica del histograma de orientaciones y ubicaciones del gradiente fue propuesta por Mikolajczyk y Schmid (2005) como una alternativa a SIFT. Se parece al método propuesto por Lowe en cuanto a que se basa en un histograma local de orientaciones del gradiente alrededor de un punto de interés. Se diferencia, de dicho método, en que:

- El histograma se calcula sobre una rejilla polar en lugar de una rejilla rectangular.
- Utiliza 16 bins para cuantizar las direcciones del grandiente en lugar de los 8 bins que utiliza SIFT.

- Utiliza análisis de componentes principales (PCA) para reducir la dimensionalidad del descriptor de la imagen.

Según sus autores, este descriptor funciona mejor, ofreciendo mayores tasas de acierto en correspondencias de puntos cuando la imagen contiene escenas estructuradas, mientras que SIFT funciona mejor cuando las escenas tienen textura.

8.6. Bibliografía

Attneave, F. (1954) "Some informational aspects of visual perception," Psychological Review, vol. 61, pp. 183–193.

Bay, H.; Tuytelaars, T. and van Gool, Luc (2006). SURF: Speeded up robust features. Proc. 9th European Conference on Computer Vision (ECCV'06) Springer Lecture Notes in Computer Science 3951: 404-417. Doi: 10.1007/11744023_32.

Beis, J.S.; Lowe, D.G., "Shape indexing using approximate nearest-neighbour search in high-dimensional spaces," Computer Vision and Pattern Recognition, 1997. Proceedings, 1997 IEEE Computer Society Conference on , vol., no., pp.1000,1006, 17-19 Jun 1997. doi: 10.1109/CVPR.1997.609451

Belongie, S. ; Malik, J. and Puzicha J. (2002). "Shape Matching and Object Recognition Using Shape Contexts". IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (24): 509–521. doi:10.1109/34.993558.

Bosch, A.; Zisserman, A. and Munoz, X. (2006). Scene classification via pLSA. Proc. 9th European Conference on Computer Vision (ECCV'06) Springer Lecture Notes in Computer Science 3954: 517~530.

Brown, M. and Lowe, D. G. "Invariant features from interest point groups," British Machine Vision Conference, BMVC 2002, Cardiff, Wales (September 2002), pp. 656-665.

Burt, Peter and Adelson, Ted (1983). The Laplacian pyramid as a compact image code. IEEE Transactions on Communications 9(4): 532–540. doi:10.1109/tcom.1983.1095851.

Burghouts, G. J. and Geusebroek, J-M (2009). Performance evaluation of local colour invariants. Computer Vision and Image Understanding 113: 48-62. doi:10.1016/j.cviu.2008.07.003.

Crowley, J. L. and Stern, R. M. (1984). Fast computation of the difference of low pass transform. IEEE Transactions on Pattern Analysis and Machine Intelligence 6(2): 212-222. doi:10.1109/tpami.1984.4767504.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. Proc. Computer Vision and Pattern Recognition (CVPR'05) (San Diego, CA): I:886-893.

Edelman, S., Intrator, N. and Poggio, T. (1997). Complex cells and object recognition. Unpublished manuscript: http://kybele.psych.cornell.edu/~edelman/Archive/nips97.pdf

Harris, C. and Stephens, M. (1988). "A combined corner and edge detector". Proceedings of the 4th Alvey Vision Conference. pp. 147–151.

Ikeuchi, K. (ed.) (2014). Computer Vision: A Reference Guide, Springer, pages 701–713. DOI: 10.1007/978-0-387-31439-6.

Ke, Y. and Sukthankar, R. (2004). PCA-SIFT: A more distinctive representation for local image descriptors. Proc. Computer Vision and Pattern Recognition (CVPR'04) (Pittsburgh, PA): II:506-513.

Koenderink, J.J. (1984). The structure of images. Biological Cybernetics, 50:363-396.
Lindeberg, T. (1994). Scale-space theory: A basic tool for analysing structures at different scales. Journal of Applied Statistics, 21(2):224-270.
Lindeberg, T. (2012) Scale Invariant Feature Transform. Scholarpedia, 7(5):10491.
Linde, O. and Lindeberg, T. (2004). Object recognition using composed receptive field histograms of higher dimensionality. Proc 17th International Conference on Pattern Recognition (ICPR'04) (Cambridge, U.K.): I:1-6. doi:10.1109/ICPR.2004.1333965.
Linde, O. and Lindeberg, T. (2012). Composed complex-cue histograms: An investigation of the information content in receptive field based image descriptors for object recognition. Computer Vision and Image Understanding 116: 538-560. doi:10.1016/j.cviu.2011.12.003.
Lowe, D. G. (1999). "Object recognition from local scale-invariant features". Proceedings of the International Conference on Computer Vision 2. pp. 1150–1157. doi:10.1109/ICCV.1999.790410.
Lowe, D. G., (2004) “Distinctive Image Features from Scale-Invariant Keypoints”, International Journal of Computer Vision, 60, 2, pp. 91-110.
Mikolajczyk, K., and Schmid, C. (2002). An affine invariant interest point detector. In European Conference on Computer Vision (ECCV), Copenhagen, Denmark, pp. 128-142.
Mikolajczyk, K and Schmid, C. (2005). A performance evaluation of local descriptors. International IEEE Transactions on Pattern Analysis and Machine Intelligence 27(19): 1615--1630. doi:10.1109/tpami.2005.188.
Schiele, B. and Crowley, J. L. (2000). Recognition without correspondence using multidimensional receptive field histograms. International Journal of Computer Vision 26(1): 31-50. doi:10.1007/bfb0015571.
Swain, M. J. and Ballard, D. H. (1991). Color indexing. International Journal of Computer Vision 7(1): 11-32. doi:10.1007/bf00130487.
SIFT Tutorial. AI Shack web page. Disponible online: http://www.aishack.in/tutorials/sift-scale-invariant-feature-transform-introduction/. (accedido 10 abril 2015)
Scale space. (2014, November 28). In Wikipedia, The Free Encyclopedia. Disponible online: http://en.wikipedia.org/w/index.php?title=Scale_space&oldid=635767690. (accedido 20 abril 2015).
Tuytelaars, T and Mikolajczyk, K (2008) Local invariant feature detectors: A survey. Foundations and Trends in Computer Graphics and Vision, 3 (3). 177 - 280. ISSN 1572-2759.
Van de Sande, K.; Gevers, Th. and Jan-Snoek, C. G. M. (2010). Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9): 1582-1596. doi:10.1109/tpami.2009.154.
Van de Weijer, J. and Schmid, C. (2006). Coloring local feature extraction. Proc. 9th European Conference on Computer Vision (ECCV'06) Springer Lecture Notes in Computer Science 3952: 334-348. doi:10.1007/11744047_26.