Lección 7
Descriptores en los puntos de interés en las imágenes

Hugo Estrada, Reynaldo Martell, Julio Cruz, Gerardo Carrera, Jesús Savage, Marco Negrete

18 de junio de 2021

Análisis de Escala

1. Los objetos de interés pueden tener diferentes tamaños o estar a diferentes distancias de la cámara.
2. No es posible conocer a priori las escalas útiles de una imagen.
3. Es indispensable analizar todas las escalas posibles para obtener la información completa.

Teoría espacio-escala

1. La cantidad de información presente en una imagen cambia cuando un objeto cambia su distancia relativa a la cámara.
2. Una imagen de un objeto distante da una pequeña cantidad de muestras con pocas frecuencias espaciales.
3. Un acercamiento factible para alcanzar el análisis invariable a la escala es muestrear el espacio-escala con la suficiente densidad de tal manera que nos sea posible rastrear la evolución de los detalles que van emergiendo cuando pasamos de una escala a otra.

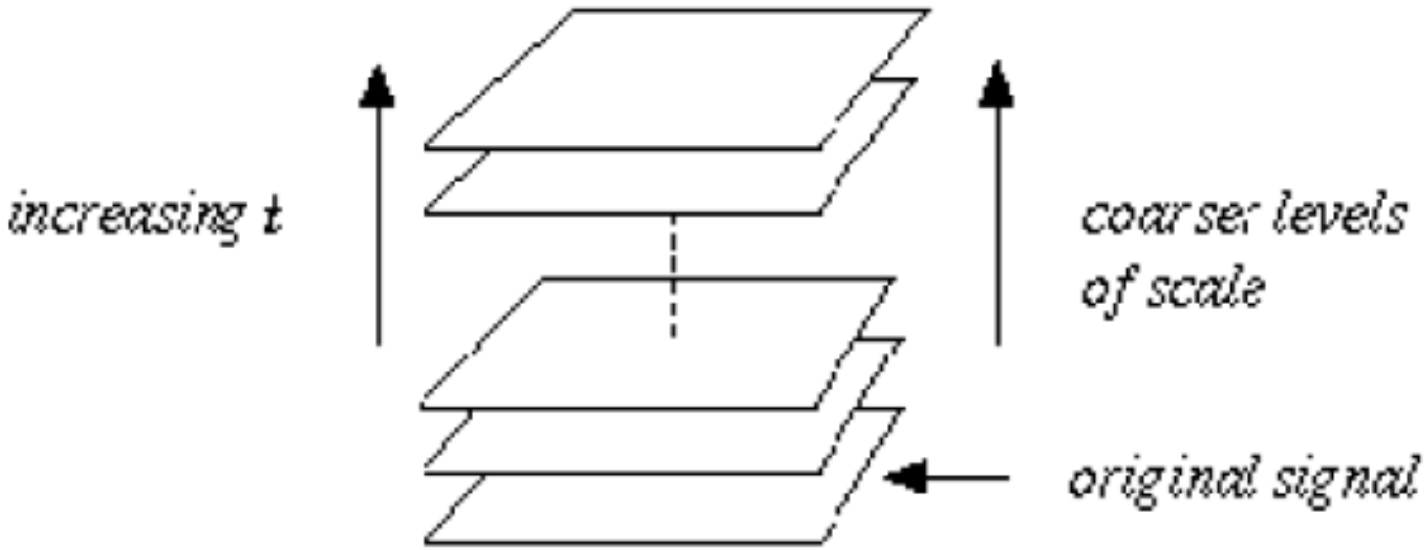
Teoría espacio-escala

El análisis en diferentes escalas es necesario debido a que:

1. Las estructuras y atributos presentes en una imagen existen a lo largo de un rango continuo de tamaños.
2. El tamaño de los atributos específicos en una imagen no es conocido con anterioridad.
3. Es posible seguir el surgimiento de estructuras a lo largo de las escalas, y utilizar estas técnicas para obtener un procesamiento independiente a la escala que sea además computacionalmente eficiente.

Teoría espacio-escala

La representación espacio-escala de una imagen consiste en una familia de imágenes suavizadas a diferentes niveles de detalle (definido por un parámetro de escala t).



Teoría espacio-escala

1. Una imagen $I(x, y)$ genera la familia: $L(x, y; t) = g(x, y; t) * I(x, y)$

   Donde $g(x, y; t) = \frac{1}{2\pi t} e^{-\frac{x^2 + y^2}{2t}}$

   es el núcleo de suavizado en la escala t.
2. Una reestricción razonable es que la estructura a bajo nivel de detalle debe resultar de la simplificación de la estructura a alto nivel de detalle, es decir, no deben aparecer nuevos elementos cuando subimos la escala.

Teoría espacio-escala

1. La familia $L(x, y; t)$ es la representación espacio-escala de la imagen $I(x, y)$.
2. Para $t = 0$; $g(x, y; 0)$ es una función impulso $\delta(x, y)$, de manera que $L(x, y; t) = I(x, y)$ es la imagen original.
3. La desviación estándar de $g(x, y; t)$ es $\sqrt{t}$, de tal manera que $L(x, y; t)$ no contiene detalles menores a $\sqrt{t}$.

Teoría espacio-escala

Las propiedades deseables para el procesamiento multi-escala son:

1. **Invarianza a los desplazamientos:** Isotropía espacial, todas las posiciones espaciales son tratadas en forma equitativa.
2. **Invarianza a la escala:** Homogeneidad espacial, todas las escalas espaciales son tratadas por igual.
3. **Causalidad:**
    1. No deben crearse nuevas curvas de nivel en los espacios-escala.
    2. No deben crearse nuevos extremos locales.
    3. No deben acentuarse los extremos locales, es decir, ningún extremo en una cierta escala se vuelve mayor en las escalas superior e inferior.

Teoría espacio-escala

$t = 0$

$t = 1$

$t = 4$

$t = 16$

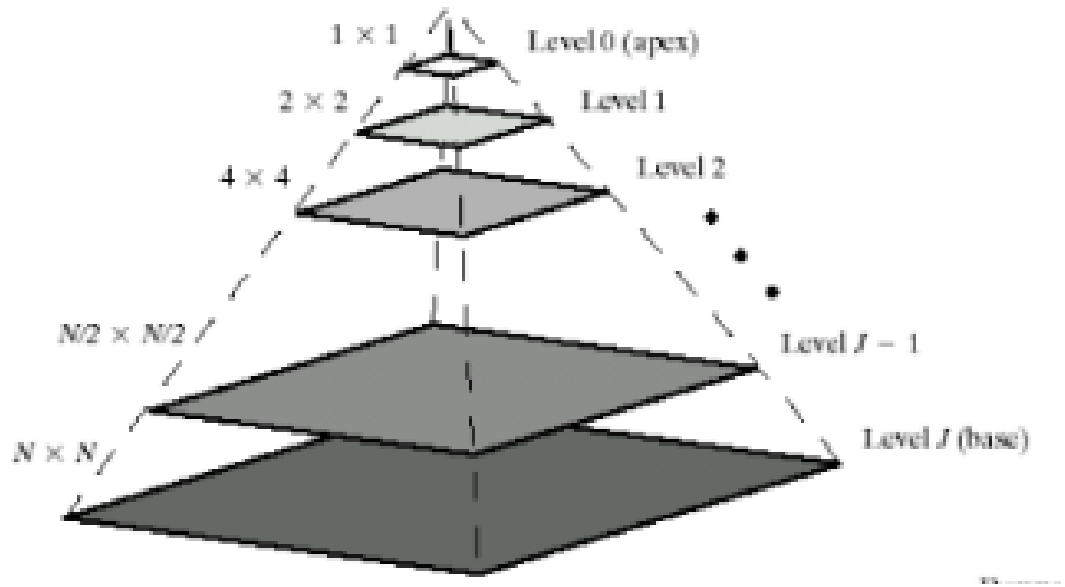
$t = 64$

$t = 256$

Pirámides

1. Ya que los detalles pequeños desaparecen cuando la escala aumenta, se puede pensar en reducir la resolución a escalas más altas.
2. Así se obtiene una pirámide de imágenes, es decir una colección de imágenes con resolución decreciente arregladas en forma de pirámides.
3. La imagen de la base de la pirámide contiene la más alta resolución, mientras que la punta de la pirámide contiene la más baja resolución. Conforme nos movemos hacia arriba de la pirámide, tanto el tamaño como la resolución de las imágenes disminuyen.

Pirámides

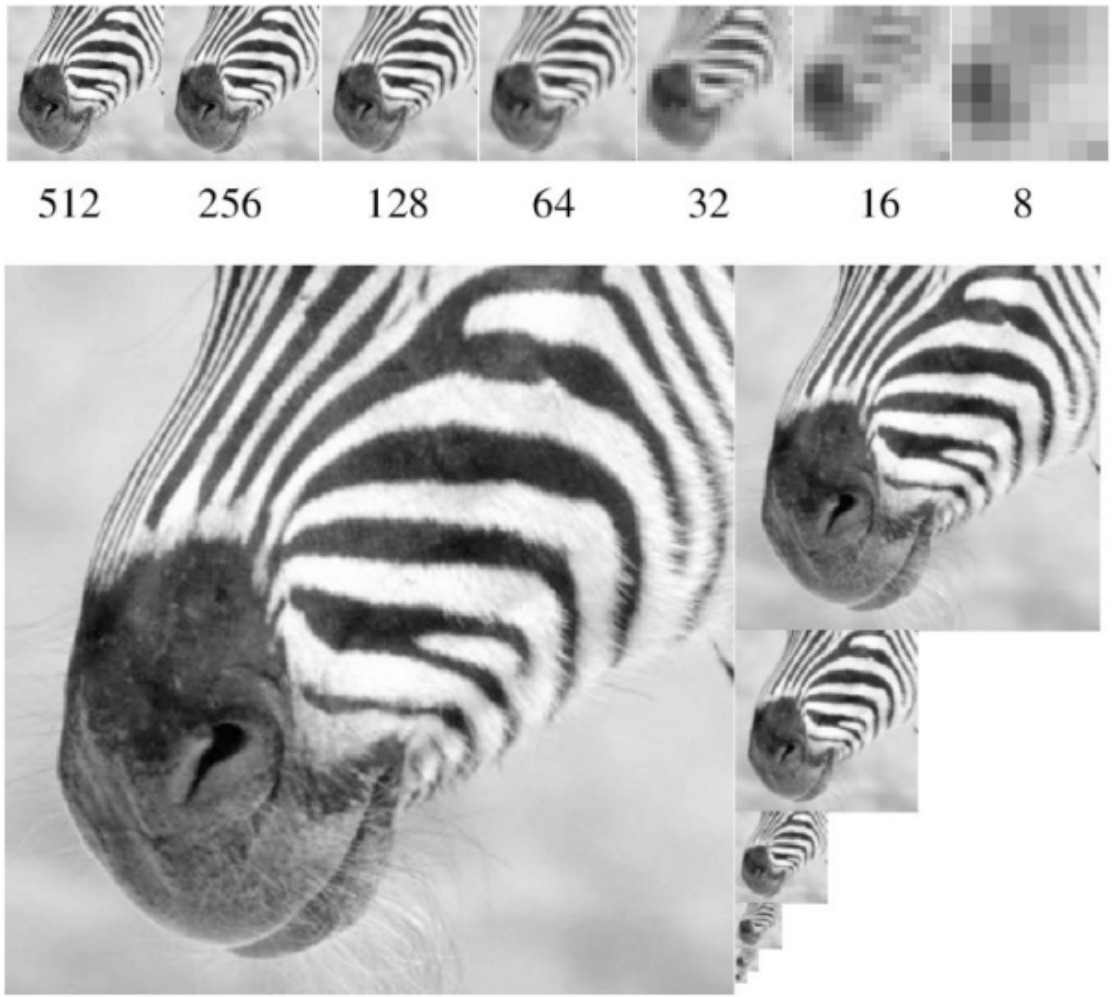


Tamaño: $2^j x 2^j$ donde $0 \leq j \leq J$

Pirámides Gaussianas

En una pirámide Gaussiana cada nivel de la pirámide es suavizado por un kernel Gaussiano simétrico y submuestreado para obtener la siguiente capa.
Notación: $S^{\downarrow}$ submuestrea la imagen I
$P_{Gaussian}(I)_{j-1} = S^{\downarrow}(G_{\sigma} \otimes P_{Gaussian}(I)_j)$
dónde, $P_{Gaussian}(I)_J = I$; es la original.

Pirámides Gaussianas



Aplicaciones

1. **Búsqueda de escala:** Muchos objetos pueden ser representados con pequeños patrones. Por ejemplo las caras tienen patrones muy bien definidos: dos ojos hundidos en fondos oscuros, debajo de dos líneas oscuras (cejas), separados de una luz especular (nariz) y sobre una barra oscura (boca). En una pirámide podemos buscar caras grandes, medianas y pequeñas a o largo de sus diferentes capas.
2. **Búsqueda espacial:** En visión estereoscópica, o análisis de movimiento, se buscan pares de puntos que casen en una escala gruesa y luego se va buscando en escalas más finas, con más detalle pero en zonas de búsqueda significativamente menores.

Aplicaciones

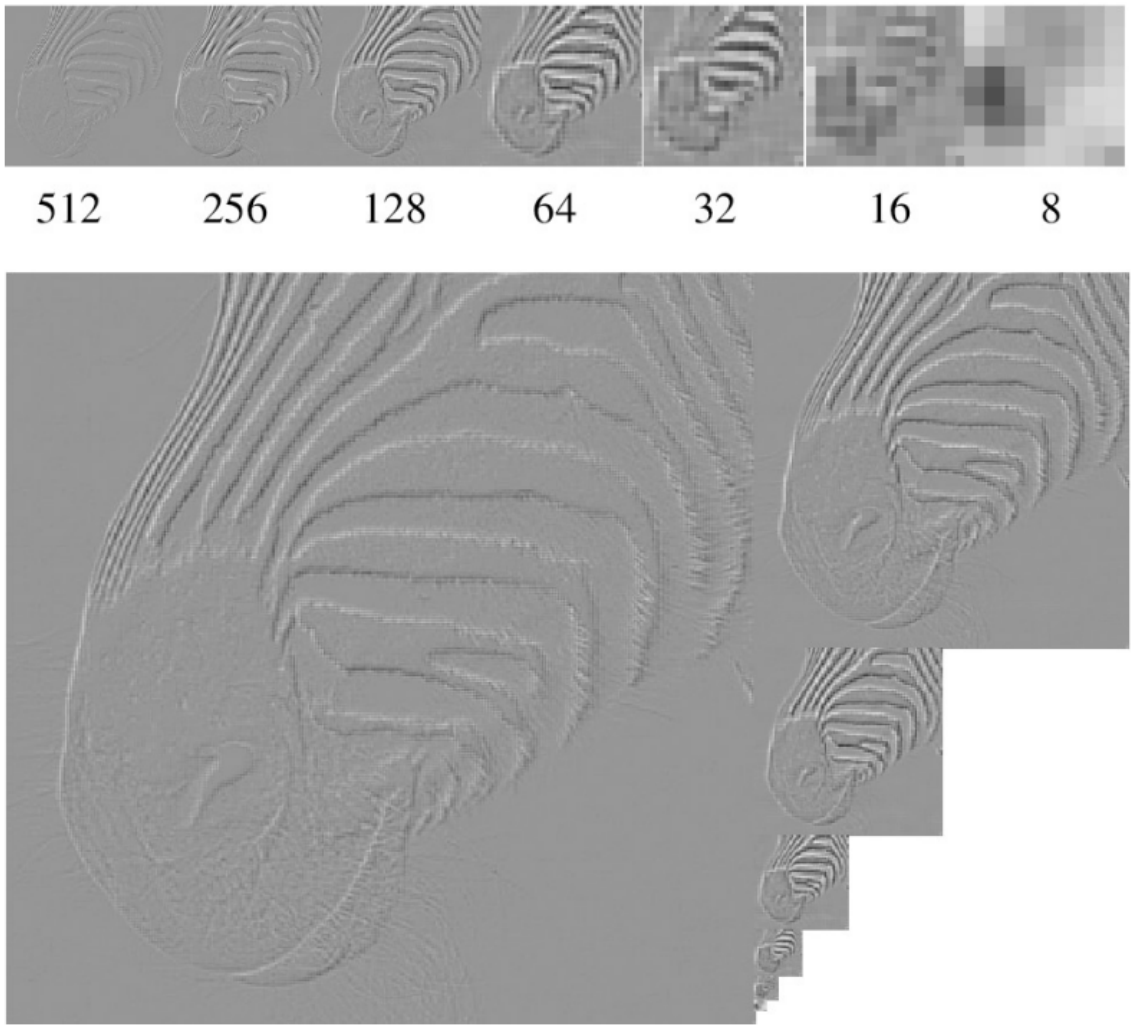
Seguimiento de características (tracking): la mayoría de las cracterísticas encontradas en escalas gruesas están asociadas con cambios bruscos de contraste. Típicamente, encontrar objetos en escalas gruesas subestima tanto el tamaño como la localización de los objetos. Por ejemplo un error de un solo pixel a escala gruesa representa un error de múltiples pixeles en escala fina.
En escalas finas existen muchos eventos asociados con cambios pequeños de contrastes bajos. Una estrategía para mejorar la caracterización de objetos en escalas finas es hacer el seguimiento de la característica de escalas finas a gruesas y quedarse con aquellas que tienen sus respectivos padres a lo largo de la pirámide (por ejemplo, diferencia entre ruido y estructuras de ciertas texturas).

Pirámides Laplacianas

Las pirámides Laplacianas hacen uso del hecho de que una capa gruesa de la pirámide Gaussiana predice la apariencia de la siguiente capa fina. Si utilizamos un operador de submuestreo que pueda producir una versión de capa gruesa del mismo tamaño que la siguiente capa fina, entonces sólo necesitamos almacenar la diferencia entre estás dos predicciones y la siguiente capa fina.
Es decir, es una secuencia de imágenes de error, diferencias de dos capas de la pirámide Gaussiana, cada una de las capas finas de la pirámide Laplaciana es la diferencia entre una capa de pirámide Gaussiana y una predicción obtenida submuestreando la siguiente capa Gaussiana de la pirámide.

$$P_{Laplaciana}(I)_k = P_{Gaussian}(I)_k - S^{\uparrow}(P_{Gaussian}(I)_{k+1})$$

Pirámides Laplacianas



Operadores diferenciales

Generalmente con el uso de operadores diferenciales se puede extraer la información necesaria para el análisis local de señales, se puede observar en mediante la expansión de la serie de Taylor.

$$f(x) = f(x_0)+f'(x_0)(x-x_0)+\frac{f''(x_0)}{2!}(x-x_0)^2+...+\frac{f^n(x_0)}{n!}(x-x_0)^n \quad (1)$$

Para nuestro caso la señal en análisis es una imagen bidimensional y se asume que es una señal diferenciable. La expansión de la serie de Taylor provee truncada hasta el segundo orden, ha sido demostrado que localmente aproxima la estructura de una imagen alrededor de un punto.

Operadores diferenciales

$$f(x) \approx f(x_0) + x\,\nabla f(x_0)^T + x^T \mathcal{H}(x_0)\,x$$

Donde ∇ es el operador gradiente, el cuál es un vector que apunta en la dirección de máximo crecimiento, y esta definido como:

$$\nabla f(x_1, x_2, ..., x_n) = (\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, ..., \frac{\partial f}{\partial x_n})$$

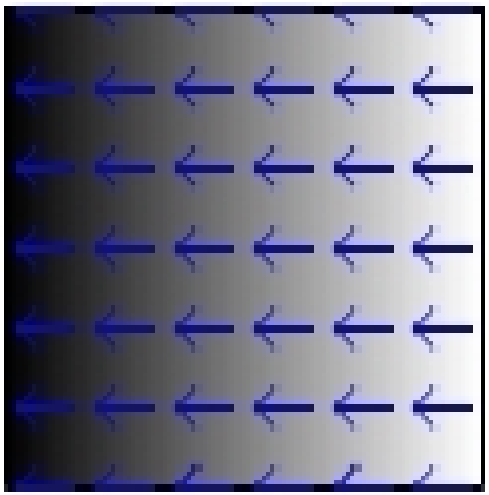


Figura: Gradiente: Dirección del máximo crecimiento

Operadores diferenciales

$$f(x) \approx f(x_0) + x\,\nabla f(x_0)^T + x^T \mathcal{H}(x_0)\,x$$

Donde $(\mathcal{H})$ es la matriz Hessiana, definida por:

$$\mathcal{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial {x_1}^2} & \frac{\partial^2 f}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 x_n} \\ \frac{\partial^2 f}{\partial x_2 x_1} & \frac{\partial^2 f}{\partial {x_2}^2} & \cdots & \frac{\partial^2 f}{\partial x_2 x_n} \\ \frac{\partial^2 f}{\partial x_n x_1} & \frac{\partial^2 f}{\partial x_n x_2} & \cdots & \frac{\partial^2 f}{\partial {x_n}^2} \end{bmatrix}$$

Los operadores diferenciales anteriores generalmente son usados de diferentes maneras, el gradiente normalmente es usado para la detección de características locales o para la descripción de estructuras locales de la imagen. Una manera de hacer un descriptor invariante a rotaciones es considerar las derivadas direccionales alrededor del punto de interés, con esto se obtiene la dirección dominante de la estructura local.

Detector de puntos de interés de Harris

La matriz Hessiana ($\mathcal{H}$) puede ser usada también, tanto para detectar como para describir las propiedades de estructuras locales de una imagen. Particularmente el uso de la traza y el determinante de esta matriz son de gran interés.
En álgebra lineal, la traza de una matriz cuadrada A, denotada tr (A), se define como la suma de elementos en la diagonal principal (de la parte superior izquierda a la inferior derecha) de A.
La traza de la matriz denota el filtro Laplaciano, generalmente usado para la detección isotrópica (independiente de la dirección) de bordes.
El uso de segundas derivadas como es el caso de la matriz Hessiana nos da una respuesta pequeña exactamente en el punto donde el cambio en la señal es significativo, a diferencia de la primera derivada, en la cual el máximo no es localizado exactamente en el punto donde la señal cambia significativamente, sino en su vecindad.

Detector de puntos de ínteres de Harris

Los antecedentes de este algoritmo provienen del detector hecho por Moravec, uno de los primeros detectores de puntos de interés basados en la intensidad de la señal. El detector de Moravec está basado en la función de auto-correlación de la señal. Esta función mide las diferencias entre los valores de gris de una ventana comparada con ventanas movidas en varias direcciones. Se consideran 3 casos:

a) Si la región en donde está la ventana es plana (constante en intensidad) entonces todos los cambios al moverse la ventana son pequeños.

b) Si la ventana está sobre un borde, los cambios van a resultar pequeños cuando el movimiento es a lo largo del eje pero grandes cuando el movimiento es perpendicular al borde.

c) Si la ventana está sobre una esquina o un punto aislado entonces todos los movimientos producen cambios grandes.

Detector de puntos de ínteres de Harris

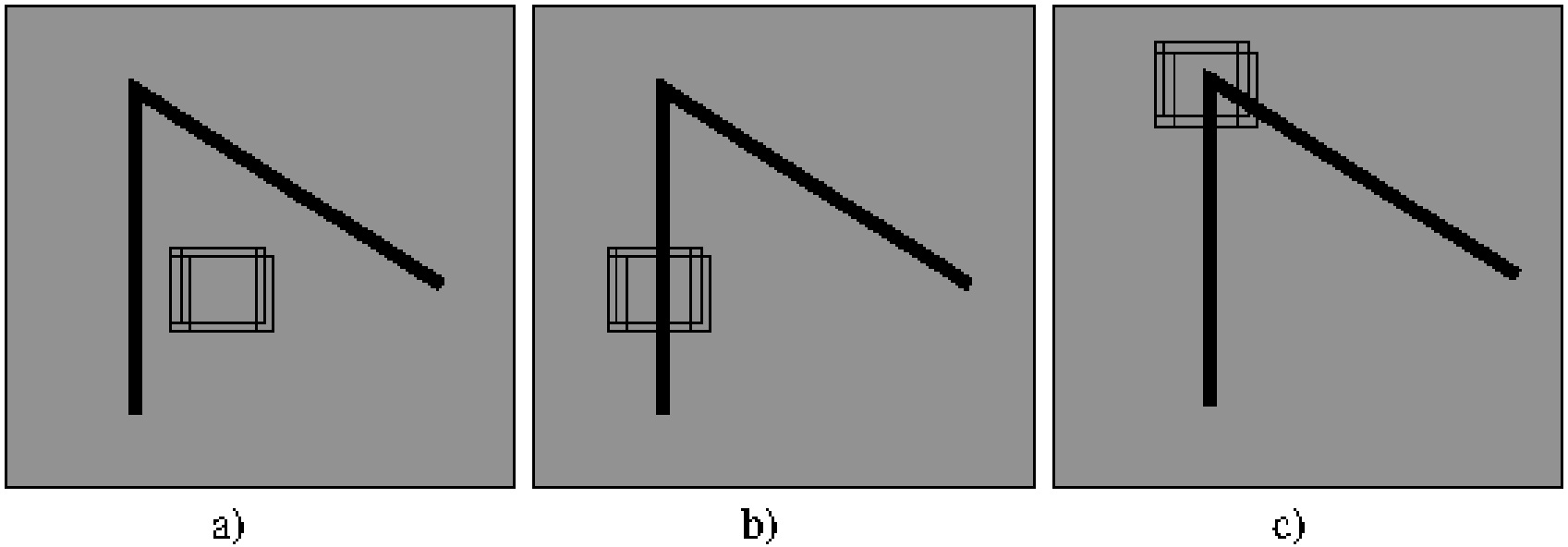


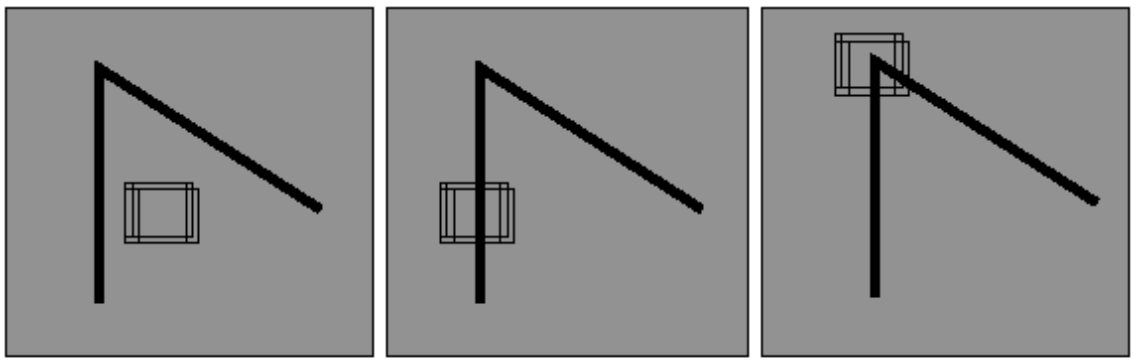
Figura: Movimientos de la ventana en diferentes direcciones discretas: a) Región plana, b) Borde, c) Esquina

Detector de puntos de ínteres de Harris

Se puede observar en la función de auto-correlación, dado un cambio $(\Delta x, \Delta y)$ y un punto (x, y) la función está dada por:

$$f(x, y) = \sum_{W} (I(x_k, y_k) - I(x_k + \Delta x, y_k + \Delta y))^2 \tag{2}$$

donde I es la imagen evaluada en (x_k, y_k) dentro de la ventana W centrada en (x, y), los movimientos de la ventana son discretos: $(1, 0), (1, 1), (0, 1), (-1, 1)$, los cuales representan un movimiento de la ventana de 45 grados. El detector busca si el mínimo de la función, $min(f)$, es mayor que un umbral, si es así entonces existe un punto de interés.



Detector de puntos de ínteres de Harris

El detector de puntos de interés de Harris mejora la aproximación hecha por Moravec debido a que ésta presenta diferentes problemas, los cuales son mencionados a continuación:

- La respuesta es anisotrópica, es decir, dependiente de la dirección de búsqueda,
- La respuesta es ruidosa debido a que la ventana es binaria y rectangular,
- El operador responde muy fácilmente a los bordes debido a que sólo el mínimo de f es tomado en consideración.

Usando la matriz de auto-correlación y una ventana Gaussiana, se pueden solucionar los problemas anteriores. La matriz $A(x, y)$ promedia las derivadas de una señal en una ventana W alrededor del punto (x, y):

$$A(x,y) = \begin{bmatrix} \sum_W (I_x(x_k, y_k))^2 & \sum_W I_x(x_k, y_k) I_y(x_k, y_k) \\ \sum_W I_y(x_k, y_k) I_x(x_k, y_k) & \sum_W (I_y(x_k, y_k))^2 \end{bmatrix}$$

Detector de puntos de ínteres de Harris

Para la matriz A se tienen dos valores característicos α y β los cuales son proporcionales a las curvaturas de la función local de auto-correlación por lo que se tienen igualmente 3 casos:

a) Si los dos valores característicos son pequeños entonces la región es homogénea.

b) Si uno de los valores característicos es grande entonces indica la presencia de un borde.

c) Si los dos valores característicos son grandes entonces esto indica una esquina.

En lugar de usar una ventana binaria y rectangular, se utiliza una función Gaussiana para pesar las derivadas dentro de la ventana, es decir:

$$W(x,y) = \frac{1}{2\pi\sigma^2} exp^{\frac{-(x^2+y^2)}{2\sigma^2}}$$

Detector de puntos de ínteres de Harris

No solamente es necesario detectar los bordes o las esquinas sino también medir la calidad de las respuestas, para lograrlo se hace uso de la traza de A y del determinante de A, evitando así el cálculo explícito de los valores característicos.

$$cornerness = Det(A) - k\ Tr(A)^2$$

donde k es una constante tal que $0 < k < 0.25$, usualmente $k = 0.04$. Una vez que se aplica la ecuación a cada pixel, se realiza una detección del máximo en una vecindad de 3×3 pixeles. Para poder hacer una detección más confiable, se hace uso de un umbral, para el cual aquellos pixeles que tengan un valor de *cornerness* arriba del umbral, se considerarán como puntos de interés, aquellos que estén por debajo del umbral se desechan.

Scale Invariant Feature Transform (SIFT)

Con el algoritmo de Harris se pueden encontrar esquinas aunque están se encuentren giradas, ya que son invariantes a la rotación, lo que significa que, incluso si se gira la imagen, podemos encontrar las mismas esquinas. Es obvio porque las esquinas siguen siendo esquinas en la imagen rotada también. Pero es posible que una esquina no sea una esquina si la imagen está escalada, una esquina en una imagen pequeña dentro de una ventana pequeña es plana cuando se amplía en la misma ventana. Entonces la esquina de Harris no es invariante en la escala y se tiene que encontrar otro algoritmo que si lo sea, ese algoritmo es SIFT (Scale Invariant Feature Transform).
Este algoritmo fue propuesto por David Lowe y ha ganado mucha popularidad en los últimos años en diferentes aplicaciones donde es necesaria la extracción de características locales dentro de una imagen, como el reconocimiento de objetos y escenas, seguimiento de objetos, reconocimiento de rostros, etc.

Detector de puntos de ínteres usando SIFT

Entre las características más importantes están: invariancia a rotación y escala, es parcialmente invariante a cambios en la iluminación y a la posición $3D$ de la cámara, y además es muy distintivo.
El algoritmo SIFT logra la invariancia a escala buscando características estables dentro de todas las posibles escalas construyendo una representación piramidal de espacio-escala. Para detectar eficientemente las posiciones de los puntos de interés se propone el uso de la función de DoG (Difference-of-Gaussians), la cual se aproxima a la función Laplaciana de Gaussianas (LoG).
La función DoG (D) se logra restando imágenes de escalas sucesivas L separadas por un factor k, es decir:
$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma)$

Scale Invariant Feature Transform (SIFT)

El uso de esta función es computacionalmente menos costoso que construir la LoG ya que previamente se construye la representación escala-espacio y solamente es necesaria una resta entre imágenes.

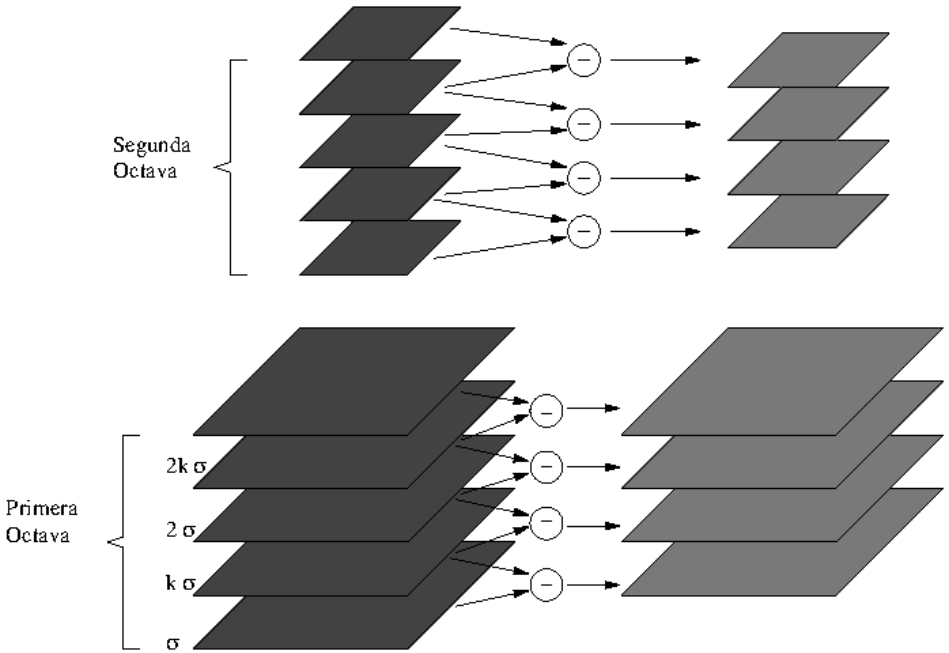


Figura: Construcción de escala-espacio: La figura de la izquierda hace la representación de escala-espacio como pirámide Gaussiana, la figura de la derecha representa la DoG la cual es construida con la resta de imágenes sucesivas en la pirámide Gaussiana

Scale Invariant Feature Transform (SIFT)

Una vez que se encuentra este DoG, se buscan imágenes para los extremos locales sobre la escala y el espacio. Un píxel de una imagen se compara con sus 8 vecinos, así como con 9 píxeles en la siguiente escala y 9 píxeles en las escalas anteriores. Si es un extremo local, es un punto clave potencial. Básicamente significa que el punto clave está mejor representado en esa escala.

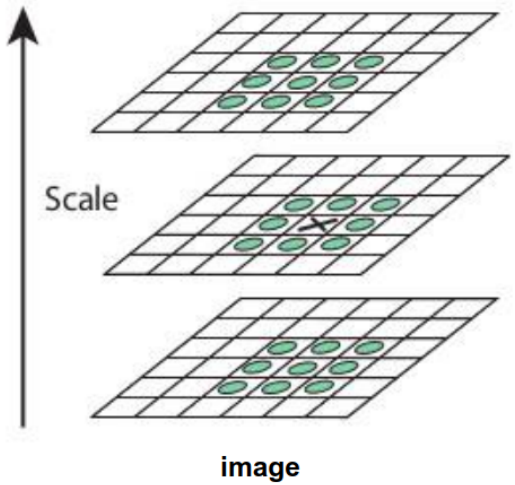


Figura: Referencia:
https : *//docs.opencv.org/master/da/df5/tutorial_py_sift_intro.html*

Scale Invariant Feature Transform (SIFT)

DoG tiene una respuesta muy alta para las orillas, por lo que éstas también deben eliminarse. Para ello, se utiliza un concepto similar al detector de esquinas de Harris. Utilizaron una matriz hessiana de 2x2 (H) para calcular la curvatura principal.

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

Si esta relación es mayor que un umbral r, ese punto clave se descarta. Por lo tanto, elimina los puntos clave de bajo contraste y los puntos clave de borde y lo que queda son puntos de interés fuertes

Descripción de puntos de interés usando SIFT

El primer paso para la construcción del descriptor es asignar una orientación consistente a cada punto de interés, incorporando esta orientación al descriptor se logra la invariancia a la rotación. La manera de obtener la orientación dominante es calculando los vectores gradientes en una ventana alrededor del punto de interés, para cada muestra dentro de la ventana se obtiene la magnitud $m(x, y)$ y la orientación $\theta(x, y)$ del gradiente usando diferencias entre pixeles vecinos,

$$\theta(x, y) = tan^{-1}\left(\frac{L(x, y+1) - L(x, y-1)}{L(x+1, y) - L(x-1, y)}\right), \tag{3}$$

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

Descripción de puntos de interés usando SIFT

Con estas orientaciones se construye un histograma de 36 intervalos cubriendo un rango de 360 grados. Cada muestra es añadida al histograma y es pesada por la magnitud de su gradiente y por una ventana circular Gaussiana con σ igual a 1.5 veces la escala del punto de interés.
En el histograma de orientación se detecta el pico más alto, así como todos aquellos picos que estén por arriba del 80 % del pico más alto por lo que para un mismo punto se pueden tener diferentes orientaciones dominantes por lo que se crean otros puntos de interés con cada orientación dominante.

Descripción de puntos de interés usando SIFT

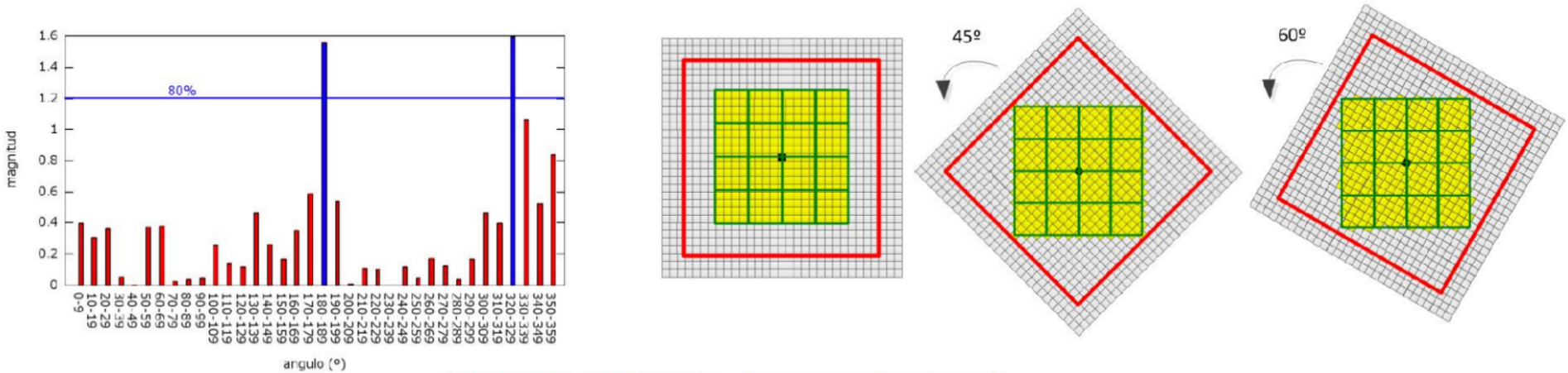


Figura: Histograma de orientaciones

Descripción de puntos de interés usando SIFT

Para que el descriptor pueda ser invariante a la rotación, la ventana es rotada con respecto a la orientación dominante, es decir, las orientaciones de los gradientes dentro de la ventana son rotados relativamente a la orientación dominante del punto de interés, es decir, cada punto de interés tiene una orientación dominante que siempre apunta en la misma dirección, aun y cuando la imagen esté rotada, al girar la ventana con respecto a esta orientación, se logra que los valores del descriptor siempre sean los mismos.

Descripción de puntos de interés usando SIFT

Para construir el descriptor cada muestra es pesada con una función Gaussiana con σ igual a la mitad del tamaño de la ventana del descriptor.
La ventana se divide en subregiones de 4×4 en donde se obtiene un histograma con 8 orientaciones por cada subregión.
El propósito de la ventana Gaussiana es evitar cambios repentinos en el descriptor con pequeñas variaciones en la posición de la ventana, así como dar más énfasis a las muestras mas cercanas al punto.

Descripción de puntos de interés usando SIFT

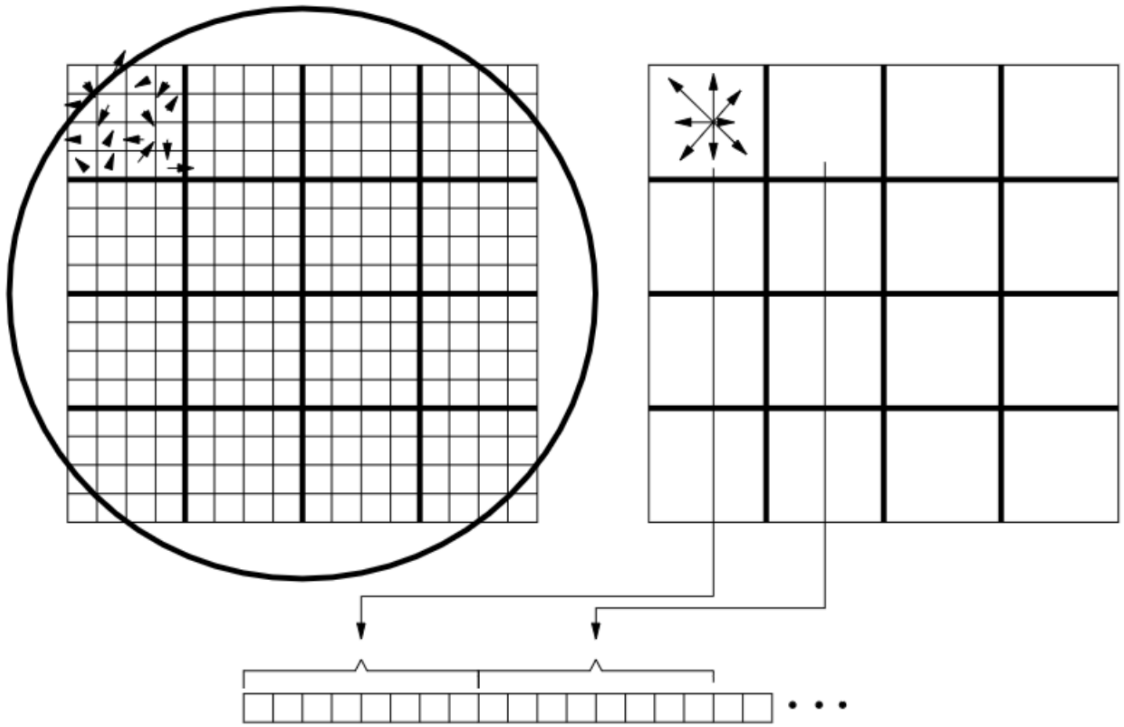


Figura: Construcción del descriptor de SIFT: La figura de la izquierda representa la ventana cuadrada de donde se obtiene la orientación dominante, donde cada gradiente es pesado con una función Gaussiana. La figura de la derecha representa la ventana rotada a lo largo de la orientación dominante y dividida en subregiones de 4×4 a partir de los cuáles se construye el descriptor. La figura de abajo muestra el descriptor representado como un vector de dimensión 128.