

Análisis de Imagen Biomédica

Resumen

En las ciencias biomédicas no solo se recogen datos; en ocasiones se generan imágenes que posteriormente se deben analizar y de las que se debe extraer información comprensiva. El análisis de imagen biomédica es el proceso de examinar y extraer información útil de imágenes biomédicas, como resonancias magnéticas, tomografías computarizadas (TC) o ecografías, para facilitar diagnósticos, monitorear el progreso de enfermedades y desarrollar nuevas técnicas de diagnóstico y tratamiento. El prefijo "bio" se refiere a la fase preclínica, siendo así las imágenes de modelos animales fundamentalmente.

Índice general

I	Introducción al procesado de imágenes	3
I.1	Formación de imágenes	3
I.1.1	De analógico a digital	4
I.2	Procesado de imágenes digitales	4
I.2.1	Image enhancement	5
I.2.2	Image analysis	6
II	Fundamentos imagen biomédica	8
II.1	Percepción visual	8
II.1.1	Fenómeno de la luz	8
II.1.2	Sistema visual humano	8
II.1.3	Luz	10
II.1.4	Percepción de luminancia y color	11
II.1.5	Brillo	12
II.2	Captura de imágenes	12
II.2.1	Sistema de adquisición	12
II.2.2	Modelo de cámara puntual <i>pin hole</i>	13
II.2.3	Color	13
II.3	Representación de imágenes	14
II.3.1	La imagen digital	14
II.3.2	True Color	15
II.3.3	Resolución vs definición	16
II.3.4	Profundidad	16
II.3.5	Formatos	17
III	Procesado imágenes digitales	18
III.1	Introducción	18
III.2	Operadores puntuales	18
III.2.1	Introducción	18
III.2.2	Modelado de histograma	19
III.2.3	Modificación de niveles	21
III.3	Operadores locales	22
III.3.1	Filtrado espacial	23
III.3.2	Suavizado: lineal y no lineal	24
III.3.3	Realce: lineal	25
IV	Registro de imágenes	27
IV.1	Tipos de registro	27
IV.2	Avances clave en IA en registro de imágenes	28

IV.3	Artículo	29
IV.4	Imágenes, mapas y templates	30
IV.4.1	Imágenes de Difusión en Neuroimagen	31
IV.4.2	Construcción de Mapas de Difusión	31
IV.4.3	Registro de imágenes y templates	31
V	Segmentación de imágenes	32
V.1	Introducción	32
V.2	Técnicas más representativas	33
V.2.1	Umbralización (operador puntual)	33
V.2.2	Clustering (agrupamiento por regiones)	35
V.2.3	Detección y unión de bordes (basados en contornos)	35
V.2.4	Contornos activos (basados en contornos)	36
V.2.5	Redes convolucionales (deep learning)	37
VI	Aplicaciones del Procesamiento Digital de Imágenes: CT, PET & SPECT, Ultrasonido y Microscopía	38
VI.1	Introducción y relación con la MRI	38
VI.2	Tomografía Computarizada (CT/TAC)	38
VI.3	PET y SPECT	40
VI.4	Ultrasonido	41
VI.5	Microscopía	41
VI.6	Integración y flujo de trabajo clínico	42
VI.7	Conclusión	42
VII	Reconocimiento de patrones en imagen biomédica	43
VII.1	Clasificación de imágenes	43
VII.1.1	¿Qué es la clasificación?	43
VII.1.2	Evaluación del rendimiento	44
VII.1.3	Clasificación basada en <i>bag of words</i>	45
VII.2	Detección de objetos	46
VII.2.1	¿Qué es la detección de objetos?	46
VII.2.2	Evaluación del rendimiento	47
VII.2.3	Detector Dalal-Triggs	47

Capítulo I

Introducción al procesado de imágenes

I.1. Formación de imágenes

Hay muchas formas de obtener una imagen, dependiendo de la fuente de energía (p. ej., fotones, ondas de radio, ultrasonido) y del sistema detector que la capture. Los tipos de imágenes se clasifican según si la formación es directa o indirecta, analógica (continua) o digital, y según el tipo de energía utilizada (radiación electromagnética, acústica, etc.).

Una forma de clasificarlas es en función del tipo de energía o radiación utilizada. La luz visible es una pequeña parte del espectro electromagnético. Las imágenes biomédicas pueden generarse utilizando otras partes de este espectro (rayos X, radiofrecuencia), así como otras formas de energía como ondas mecánicas (sonar, ultrasonido) o campos (magnéticos, eléctricos).

Aunque el grueso del curso sea el análisis y procesado de las imágenes, es importante saber de dónde vienen y su contexto.

- La **formación de imagen de rayos X** proviene de una fuente emisora. Estos rayos interaccionan con el tejido biológico (son atenuados principalmente por absorción y dispersión). La radiación que atraviesa el tejido incide sobre un detector (que históricamente era una placa fotográfica o de película, pero hoy son casi exclusivamente detectores digitales como placas de fósforo o detectores planos).
- La **formación de una imagen de Resonancia Magnética (RM)** se basa en aplicar un campo magnético estático y fuerte (B_0) al tejido, alineando los momentos magnéticos de los núcleos de hidrógeno. Se aplican pulsos de radiofrecuencia (RF) para excitar estos núcleos, que al relajarse emiten señales de RF. Esta señal se detecta con bobinas. Utilizando gradientes de campo magnético (que varían el campo de forma lineal en el espacio), se puede codificar espacialmente la señal y reconstruirla digitalmente para asignar diferentes intensidades de señal a cada voxel (elemento de volumen 3D) o píxel (2D).

I.1.1. De analógico a digital

Para convertir una imagen analógica (como una placa de rayos X o una fotografía) a digital se utiliza un dispositivo de digitalización, como un escáner o un detector digital directamente. Este proceso implica **muestrear (sampling)** la imagen, dividiendo el espacio en una matriz discreta de elementos (píxeles), y **cuantizar** la intensidad de luz en cada punto, asignándole un valor digital discreto.

La resolución espacial tiene que ver con el nivel de detalle discernible, determinado por el tamaño del píxel y las propiedades del sistema de imagen. Una imagen con muchos píxeles pequeños (alta resolución) permite visualizar detalles finos y bordes definidos. En cambio, una imagen de baja resolución tiene pocos píxeles grandes, lo que resulta en una pérdida de detalles y una apariencia pixelada o borrosa al ampliarla.

Las imágenes con las que se suele trabajar en diagnóstico por imagen suelen representarse en escala de grises. Esto significa que a cada píxel se le asigna una intensidad lumínica representada por un valor numérico. En el formato estándar de 8 bits, este valor va de 0 (negro) a 255 (blanco), permitiendo $2^8 = 256$ tonos de gris intermedios. La cantidad de bits utilizada para representar el valor de un píxel se denomina profundidad de bits (bit depth).

Cualquier sistema moderno de formación de imagen biomédica integra **sensores** (o detectores) y un **conversor de analógico a digital (ADC)**. La imagen digital se debe guardar, procesar y, en ocasiones, se reconvierte a analógico (por ejemplo, mediante una pantalla) para su visualización.

I.2. Procesado de imágenes digitales

Dentro del procesamiento de imagen hay varias categorías o pasos:

1. **Mejora o Realce de la Imagen (Image Enhancement):** Técnicas para ajustar propiedades como el brillo, el contraste o para aplicar filtros (como la convolución) con el fin de hacer ciertas características más visibles para el observador humano o para un algoritmo posterior. Incluye la corrección de algunos artefactos.
2. **Restauración de Imagen (Image Restoration):** Su objetivo es eliminar o reducir la degradación (como el desenfoque o el ruido) utilizando un modelo de cómo se degradó la imagen. Se busca aproximarse a la imagen original no degradada.
3. **Análisis de la Imagen (Image Analysis):** Procesos como la segmentación (delimitación de regiones de interés), la extracción de características (cálculo de métricas cuantitativas) y la clasificación de objetos.
4. **Compresión de Imagen (Image Compression):** Técnicas para reducir el tamaño de los archivos de imagen para su almacenamiento o transmisión eficiente (p. ej., JPEG, JPEG 2000, DICOM con compresión).

5. **Síntesis de Imagen (Image Synthesis):** Generación de imágenes a partir de modelos o datos, como en la reconstrucción tomográfica o la generación de imágenes mediante Inteligencia Artificial.

1.2.1. Image enhancement

Las imágenes biomédicas se utilizan para el diagnóstico, tanto **cualitativo** (observación visual de morfología) como **cuantitativo** (medición de parámetros fisiológicos o bioquímicos). Para el diagnóstico cualitativo, se tratan las imágenes para reducir ruido o realzar bordes. En las imágenes cuantitativas, el valor numérico del píxel tiene significado físico (p. ej., unidades Hounsfield en TC, concentración de un contraste, valores T1 o T2 en RM). Por ello, al manipular estas imágenes para su análisis, es crucial utilizar técnicas que no alteren o que corrijan de manera controlada los valores subyacentes, preservando la información biológico-física. A menudo, la consistencia en el procesamiento (aplicar el mismo algoritmo a todos los sujetos de un estudio) es más importante que la absolutez del valor.

Los **artefactos** son patrones o estructuras presentes en la imagen que no corresponden a la anatomía o fisiología real del sujeto, sino que son causados por el equipo, el protocolo de adquisición o el propio paciente. Estos incluyen desde artefactos por metal (implantes) hasta los causados por el movimiento voluntario o involuntario del paciente durante una adquisición, especialmente si es larga. La **reducción de ruido** (ruido aleatorio inherente a cualquier medición física) es también esencial para mejorar la relación señal-ruido (SNR).

Hoy día, las técnicas de **aprendizaje automático (machine learning)**, y en particular las **redes neuronales convolucionales (CNN)**, se utilizan extensamente. Un ejemplo es el entrenamiento de redes con pares de imágenes: una entrada degradada (p. ej., con artefactos de movimiento o ruido) y una salida objetivo de referencia ("ground truth") de alta calidad. La red aprende la transformación para mapear la imagen defectuosa a una versión corregida. Es importante destacar que a menudo el "ground truth" perfecto no existe *in vivo*, por lo que los modelos se entrena con datos simulados o con imágenes de alta calidad adquiridas en condiciones ideales.

Para corregir el ruido, una técnica común es el **filtrado espacial**. Un ejemplo básico es el **filtro Gaussiano**, que suaviza la imagen promediando los valores de los píxeles vecinos, ponderados por una función Gaussiana. Esto homogeniza las regiones y reduce el ruido, pero a costa de una potencial pérdida de detalle (suavizado de bordes). La operación se aplica deslizando un kernel (ventana) sobre la imagen, realizando una **convolución** (producto punto entre el kernel y los píxeles subyacentes) en cada posición.

Mejora de la Resolución: El término general es superresolución (super-resolution). Este proceso busca estimar o reconstruir una imagen de alta resolución a partir de una o varias imágenes de baja resolución. El objetivo es recuperar detalles finos perdidos durante la adquisición, lo que conduce a una visualización más clara y puede facilitar un diagnóstico más preciso. Es importante distinguirlo de un simple zoom o interpolación, que agranda los píxeles sin añadir información nueva real. Las técnicas de superresolución pueden ser:

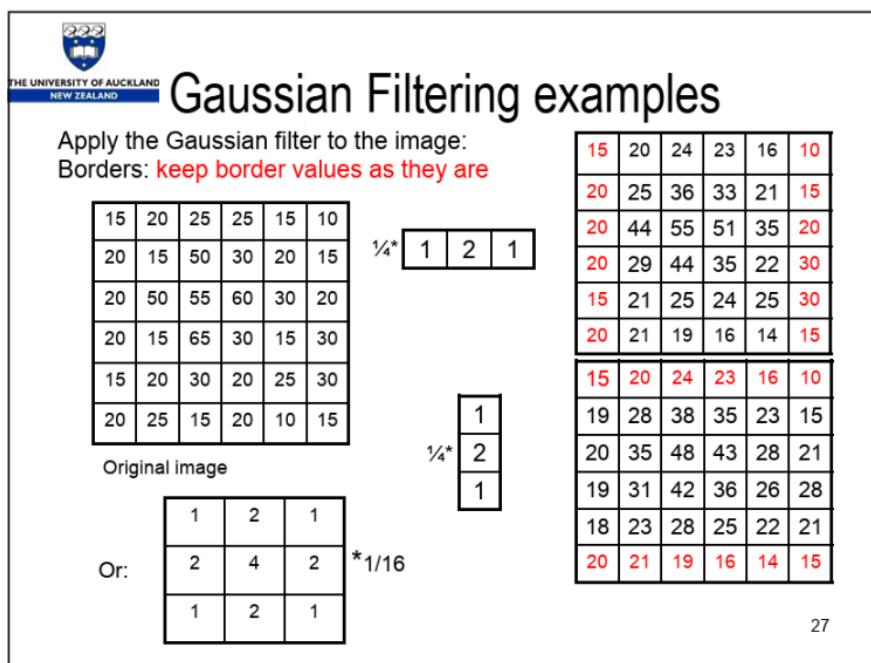


Figura I.1: Ejemplo de filtro Gaussiano. Se aplica la matriz izquierda inferior sobre la imagen original, multiplicando los valores por la matriz de filtro y dividiendo por 16. Esto se realiza sobre cada píxel a excepción de los bordes. Por ejemplo, para el píxel con valor 15 de la segunda fila, segunda columna: $(15 * 1 + 20 * 2 + 25 * 1 + 20 * 2 + 15 * 4 + 50 * 2 + 20 * 1 + 50 * 2 + 55 * 1) / 16 = 455 / 16 = 28,4375 \approx 28$, que es el valor del píxel de la matriz derecha inferior.

- **Basadas en reconstrucción:** Utilizan múltiples imágenes sub-píxel de la misma escena y algoritmos para fusionarlas.
- **Basadas en aprendizaje:** Utilizan redes neuronales (como CNNs) entrenadas con pares de imágenes (baja/alta resolución) para aprender el mapeo que añade los detalles faltantes.

I.2.2. Image analysis

La **segmentación** es un paso fundamental en el análisis de imágenes. Su objetivo es dividir una imagen en regiones o estructuras significativas, permitiendo extraer y cuantificar la información de interés. Por ejemplo, aislar solo los huesos en una Tomografía Computarizada (TC), delimitar un tumor, o separar el ventrículo izquierdo en una imagen cardiaca.

Se lleva a cabo mediante algoritmos que identifican boundaries (bordes) o regiones homogéneas basándose en propiedades como la intensidad, el color, la textura o el contexto. No se limita solo a la extracción de bordes; existen múltiples enfoques:

- **Basados en umbral (thresholding):** Separan según el valor de intensidad.
- **Basados en regiones (region growing, watershed):** Agrupan píxeles conectados con propiedades similares.

- Basados en bordes (edge detection): Identifican discontinuidades en la intensidad (usando operadores como Sobel, Canny).
- Basados en modelos (active contours, level sets): Utilizan curvas o superficies que evolucionan para ajustarse a los contornos.
- Basados en aprendizaje (Machine/Deep Learning): Las redes neuronales (especialmente las U-Net) aprenden a segmentar a partir de ejemplos etiquetados.

El **registro de imágenes (Image Registration)** es el proceso de alinear geométricamente dos o más conjuntos de imágenes (datasets) adquiridas en diferentes momentos, desde diferentes modalidades o desde distintos puntos de vista. El objetivo es establecer una correspondencia punto a punto entre ellas para poder comparar, fusionar o analizar la información de forma coherente.

Esto se logra encontrando la transformación espacial óptima que mapea los puntos de una imagen (imagen móvil o "moving") sobre los de otra (imagen de referencia o "fixed"). Los tipos de transformación, ordenados de menor a mayor complejidad y flexibilidad, son:

- **Registro rígido (Rigid):** Alinea imágenes solo con rotaciones y traslaciones (desplazamientos) globales. Preserva las distancias y ángulos entre todos los puntos. Es útil para imágenes de la misma anatomía sin cambios internos (p. ej., cabeza en diferentes estudios de RM).
- **Registro por similitud (Similarity):** Añade escalado isotrópico (mismo factor de escala en todos los ejes) a la transformación rígida. Preserva las formas y los ángulos, pero no las distancias absolutas.
- **Registro afín (Affine):** Incluye cizallamiento (shearing) y escalado anisotrópico (diferente factor en cada eje), además de las transformaciones rígidas y de escalado. Las líneas paralelas siguen siéndolo después de la transformación, pero los ángulos pueden no conservarse. Es útil para corregir diferencias de adquisición o geometría entre modalidades.
- **Registro no rígido o deformable (Non-rigid/Deformable):** Maneja deformaciones elásticas o fluidas complejas, localizadas y no lineales. Es esencial para compensar movimientos de órganos (como el corazón o los pulmones), cambios anatómicos (crecimiento de un tumor) o para alinear imágenes de pacientes diferentes en un atlas poblacional.

Capítulo II

Fundamentos imagen biomédica

II.1. Percepción visual

II.1.1. Fenómeno de la luz

Una fuente de luz **monocromática** emite radiación predominantemente en una única **longitud de onda** (o frecuencia), percibida como un color puro. Un ejemplo característico es el láser. Por el contrario, la luz **policromática** está compuesta por una mezcla de múltiples longitudes de onda, como la luz solar o la de una bombilla LED blanca.

Existen dos tipos de fuentes luminosas:

- **Fuentes Primarias o Emisivas:** Generan su propia luz mediante procesos de excitación de átomos o moléculas (ej.: el Sol, una bombilla, un LED).
- **Fuentes Secundarias o Reflectantes:** No generan luz propia, sino que reflejan total o parcialmente la luz que reciben de una fuente primaria (ej.: la Luna, un libro, la mayoría de los objetos que nos rodean). Casi todo lo que vemos son fuentes secundarias.

II.1.2. Sistema visual humano

El proceso de la visión comienza cuando la luz entra en el ojo y es proyectada sobre la **retina**, donde los **fotorreceptores** la captan y la convierten en señales electroquímicas (proceso de **transducción**). Estas señales se transmiten a través del **nervio óptico** al cerebro, donde se interpretan para generar la percepción visual.

Desde un punto de vista óptico, el ojo humano es análogo a una cámara fotográfica:

- **Lente:** El cristalino (junto con la córnea y los humores acuoso y vítreo) se encarga de enfocar la luz, proyectando una imagen nítida sobre la retina. Su forma se ajusta en un proceso llamado acomodación.
- **Diafragma:** El iris (la parte coloreada) actúa como un diafragma, controlando el tamaño de la pupila para regular la cantidad de luz que entra en el ojo.

- **Sensor:** La retina equivale al sensor de una cámara (CCD/CMOS). Es una capa de tejido sensible a la luz ubicada en la parte posterior del ojo.

La retina contiene dos tipos principales de fotorreceptores:

- **Bastones:** Altamente sensibles a la intensidad lumínica (luminancia), pero no al color. Son responsables de la visión escotópica (visión en condiciones de baja iluminación). Se concentran en la periferia de la retina, siendo muy sensibles al movimiento.
- **Conos:** Menos sensibles que los bastones, pero especializados en la percepción del color (visión fotópica, en condiciones de alta iluminación). Se concentran en la fóvea, la zona central de la retina de máxima agudeza visual. Existen tres tipos, cada uno con un pico de sensibilidad a diferentes longitudes de onda: rojo (larga), verde (media) y azul (corta).

La capacidad del ojo tiene limitaciones físicas. La **resolución espacial** (capacidad para discernir detalles finos) y la **resolución temporal** (capacidad para discernir eventos rápidos) son finitas debido al número limitado de fotorreceptores y a su tiempo de respuesta.

La **agudeza visual** es la capacidad de distinguir entre dos puntos separados. Si la separación angular entre ellos es superior a **1 minuto de arco** ($1/60$ de grado), estimulan fotorreceptores y fibras nerviosas diferentes, por lo que el cerebro los percibe como entidades distintas. Si la separación es menor, ambos puntos estimulan el mismo receptor, y el cerebro percibe una **mezcla aditiva espacial (MAE)**, fusionándolos en un solo estímulo de color.

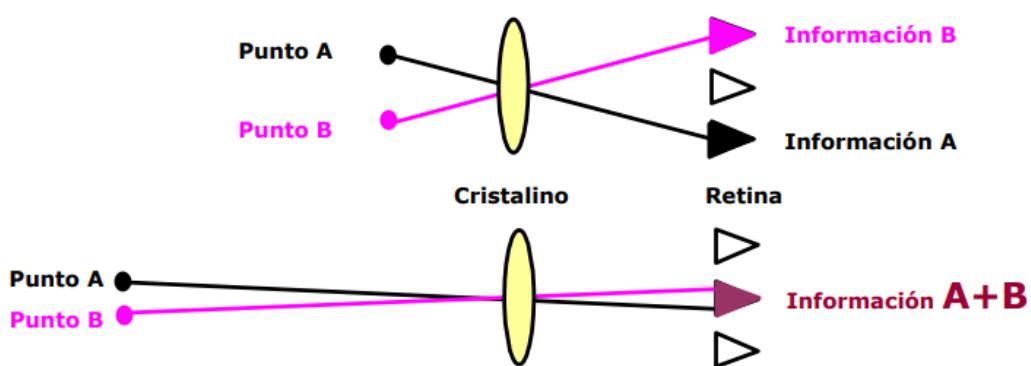


Figura II.1: Ilustración del principio de la mezcla aditiva espacial. Puntos de color suficientemente pequeños y cercanos se perciben como un color uniforme.

Este principio es fundamental en tecnologías de visualización como las pantallas de televisión y monitores, donde la imagen se forma mediante **píxeles** compuestos por subpíxeles rojos, verdes y azules (triadas de colores) cuya separación es inferior al umbral de agudeza visual.

La **memoria visual** o persistencia retiniana es una propiedad por la cual la excitación de los fotorreceptores no cesa instantáneamente tras desaparecer el estímulo, sino que continúa enviando señales al cerebro durante un breve periodo. La integración

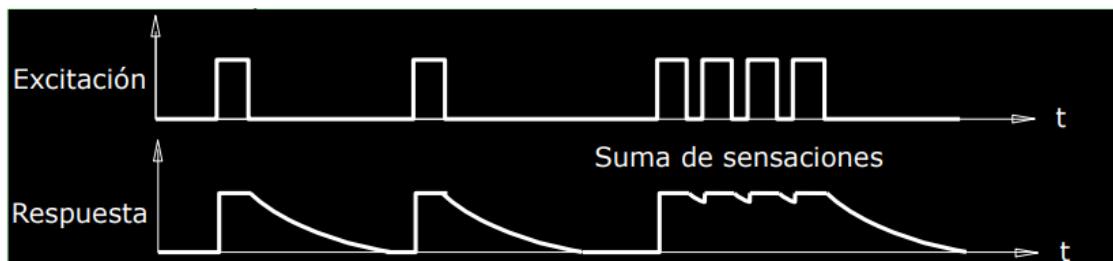


Figura II.2: Ilustración del principio de la mezcla aditiva temporal. Estímulos discretos sucesivos se perciben como un continuo si la frecuencia es suficientemente alta.

de impulsos luminosos consecutivos da lugar a la sensación de continuidad, un fenómeno conocido como **mezcla aditiva temporal (MAT)** o persistencia retiniana.

Si el intervalo entre impulsos es mayor de aproximadamente 40-50 ms (equivalente a 20-25 Hz), se percibe un parpadeo o *flicker* molesto. Por debajo de este umbral, la respuesta al nuevo estímulo se suma a la anterior, creando una sensación de flujo continuo. Esto explica por qué una secuencia de imágenes estáticas (fotogramas) a una velocidad superior a 25 fps (fotogramas por segundo) se percibe como movimiento continuo. El umbral exacto varía con la luminancia y el campo visual estimulado.

II.1.3. Luz

La visión de los objetos no emisivos se debe a la reflexión (y en algunos casos, a la transmisión) de la luz que incide sobre ellos. La intensidad de la luz reflejada (L , luminancia) que llega al ojo depende de la intensidad de la luz incidente (E , iluminancia, medida en lux) y de las propiedades reflectivas del material (R).

En términos generales, se puede modelar como:

$$C_R(X, V, geom, t, \lambda) = E(X, t, \lambda) \cdot r(V, geom, \lambda)$$

Donde: $R(V, geom, \lambda)$ es la reflectancia del material, que depende del ángulo de visión (V), la geometría de la superficie (si es rugosa/difusa o lisa/especular) y la longitud de onda (λ). Un objeto parece de un color porque absorbe selectivamente ciertas longitudes de onda y refleja otras.

Existen dos modelos de síntesis de color:

- **Síntesis Aditiva:** Propia de las fuentes de luz primarias. Los colores se crean sumando diferentes longitudes de onda de luz. La suma de los colores primarios aditivos (Rojo, Verde, Azul - RGB) en su máxima intensidad produce la percepción de blanco. Ej.: pantallas, monitores.
- **Síntesis Sustractiva:** Propia de los pigmentos y materiales (fuentes secundarias). Los colores se crean porque el material absorbe (sustrae) ciertas longitudes de onda de la luz blanca incidente y refleja otras. Los colores primarios sustractivos son Cian, Magenta y Amarillo (CMY). La "suma" teórica de los tres absorbería toda la luz, produciendo negro. Ej.: pintura, impresión.

II.1.4. Percepción de luminancia y color

El sistema visual humano opera en tres regímenes de visión según el nivel de iluminación:

- **Visión Escotópica:** Activada en condiciones de muy baja iluminación (noche). Dominada por los bastones. No permite la percepción del color (visión en escala de grises) y tiene una baja agudeza visual.
- **Visión Fotópica:** Activada en condiciones de alta iluminación (día). Dominada por los conos. Permite la percepción del color y una alta agudeza visual.
- **Visión Mesópica:** Régimen intermedio (amanecer, atardecer, iluminación tenue). Participan tanto bastones como conos. La percepción del color y la agudeza visual son intermedias. Los colores y nuestros conos no son puros. Los conos verdes absorben más que los conos rojos y azules. Nuestro cerebro recibe la información de los tres sensores y los integra para percibir un color u otro.

La respuesta espectral de los tres tipos de conos se solapa. El cerebro deduce el color percibido (**crominancia**) a partir de la **señal tricromática** comparada que envían estos tres tipos de receptores. La crominancia se describe mediante dos atributos:

- **Tono:** el color en sí mismo.
- **Saturación:** la pureza o intensidad del color.

Por otro lado, la **luminancia** se refiere a la cantidad de luz medida físicamente (en candelas/ m^2), que se percibe subjetivamente como **brillo** en una escala de grises.

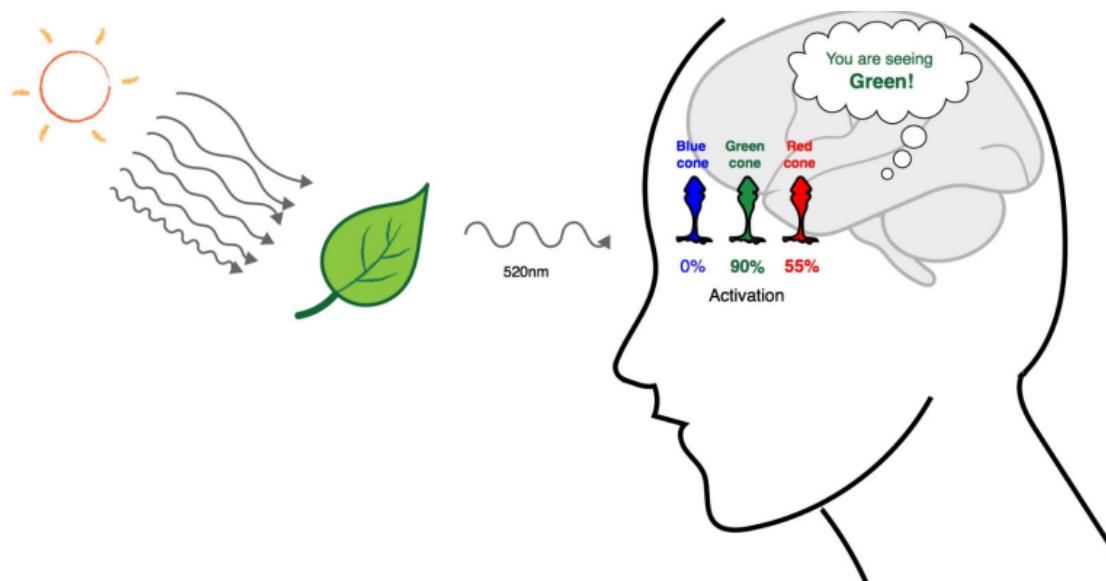


Figura II.3: Representación de la percepción visual, desglosando la información en componentes de luminancia (brillo) y crominancia (tono y saturación).

II.1.5. Brillo

El brillo es la percepción subjetiva de la luminancia. El sistema visual es extraordinariamente adaptable, capaz de funcionar en un rango de aproximadamente 10^{10} niveles de intensidad lumínica mediante el proceso de adaptación (ajuste de la sensibilidad retinal según la luminancia media del entorno).

La percepción del brillo es relativa, no absoluta. Depende críticamente del contraste entre un objeto y su fondo o entorno inmediato (Ley de Weber-Fechner). El ojo es mucho más sensible a las variaciones de luminancia (bordes, movimientos) que a los valores constantes.

No existe una medida física directa del brillo, ya que es una experiencia perceptiva compleja. En condiciones de alta luminancia (visión fotópica), el ojo tiene una mayor agudeza para discernir objetos claros sobre fondos oscuros. En condiciones de baja luminancia (visión escotópica o mesópica), la sensibilidad cambia, y puede resultar más fácil discernir objetos de luminancia media.

Dado que la percepción del brillo es relativa, su evaluación debe considerar necesariamente el entorno. La **evaluación del contraste** es, por tanto, la medición cuantitativa de la diferencia percibida en luminancia (brillo) entre un objeto de interés (por ejemplo, un texto) y su fondo inmediato. Esta diferencia se expresa comúnmente como una razón de contraste (por ejemplo, 4:1, 7:1), que se calcula dividiendo la luminancia relativa de la parte más clara entre la de la parte más oscura. Un contraste alto asegura que la información sea discernible para el sistema visual, lo que es un principio crítico en disciplinas como el diseño de interfaces, la señalética y la accesibilidad web, garantizando que el contenido sea legible para usuarios con diversas capacidades visuales o en condiciones de iluminación variables.

II.2. Captura de imágenes

II.2.1. Sistema de adquisición

Los elementos funcionales fundamentales de un sistema de adquisición de imágenes son: el **sistema óptico** (o grupo óptico), el **sensor de imagen** (con sus distintas tecnologías y características) y la etapa de **procesamiento de la señal** (que convierte la información cruda del sensor en una imagen o señal de video utilizable).

En una cámara, la exposición —la cantidad de luz que alcanza el sensor— se controla mediante dos mecanismos principales:

- El **obturador** controla la duración de la exposición (el intervalo de tiempo durante el cual la luz incide en el sensor).
- El **diafragma** controla la intensidad de la luz que entra a través de la lente, ajustando el tamaño de la abertura.
- El tercer elemento crucial es el propio **sensor**, compuesto por materiales fotosensibles (fotodiodos) que convierten la energía de los fotones (luz) en una señal eléctrica (carga).

II.2.2. Modelo de cámara puntual *pin hole*

El modelo de cámara estenopeica o *pinhole* es el modelo más simple de formación de imágenes. Sus componentes esenciales son:

- Un **centro de proyección** (el orificio estenopeico o *pinhole*).
- Una **distancia focal** (f), que es la distancia entre el centro de proyección y el plano de la imagen.
- Un **plano de imagen** donde se forma la imagen proyectada.

Debido a la propagación rectilínea de la luz, la imagen formada en el plano es **invertida** (tanto vertical como horizontalmente). La principal ventaja de este modelo es que toda la escena está enfocada sin necesidad de un sistema de enfoque; la profundidad de campo es infinita.

Sin embargo, existe un compromiso (*trade-off*) crítico en el tamaño del orificio:

- Si el orificio es demasiado grande, cada punto de la escena proyecta un pequeño círculo de confusión (*circle of confusion*) en el plano de imagen, resultando en una imagen desenfocada debido a la superposición de estos círculos.
- Si el orificio es demasiado pequeño, el fenómeno de la difracción de la luz se vuelve significativo. La luz se dispersa al pasar por la pequeña abertura, provocando que los puntos de la imagen se difuminen entre sí y se pierda definición y nitidez. Existe, por tanto, un tamaño de orificio óptimo que minimiza la combinación de estos dos efectos adversos.

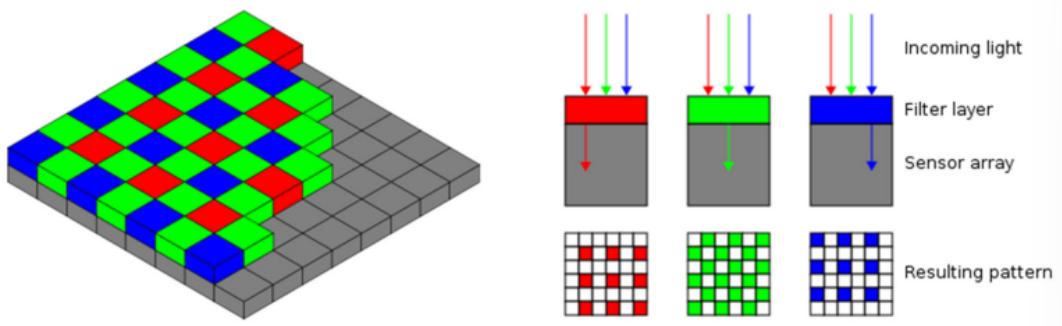
II.2.3. Color

Los sensores de imagen (CCD o CMOS) son inherentemente **monocromáticos**; solo pueden medir la intensidad de la luz, no su longitud de onda (color). Para capturar imágenes en color, se emplean diversas estrategias que implican la separación de la luz en sus componentes espectrales:

- **Sistema de 3 Sensores (3-CCD/3-CMOS)**: Utiliza un prisma dicroico para dividir la luz incidente en sus tres componentes espectrales primarias (rojo, verde y azul). Cada haz de color se dirige hacia un sensor dedicado. Este sistema ofrece la máxima calidad y fidelidad de color, ya que cada píxel de la imagen final se genera con información de intensidad completa para los tres canales. Su principal desventaja es el alto coste y la complejidad mecánica.
- **Filtro de Color Rotativo**: Se coloca un filtro de color (rojo, verde o azul) giratorio delante de un único sensor. La cámara captura secuencialmente un fotograma para cada color primario. Este método es más económico que el de 3 sensores, pero introduce graves inconvenientes: baja calidad de color (especialmente con objetos en movimiento, que producen artefactos de *ghosting*), una tasa de captura efectiva menor y la necesidad de una iluminación constante durante la rotación.

- **Matriz de Filtros de Color (CFA - *Color Filter Array*):** Es el método más común en cámaras consumer y profesionales. Consiste en un mosaico de microfiltros coloreados, depositado directamente sobre la superficie del sensor, donde cada filtro corresponde a un único fotodiodo (píxel). Cada píxel del sensor captura únicamente la intensidad de una componente de color (R, G o B). El patrón más utilizado es el filtro de Bayer (desarrollado por Bryce Bayer en Kodak, 1976), que utiliza un 50 % de filtros verdes, un 25 % de rojos y un 25 % de azules, imitando la mayor sensibilidad del ojo humano al verde. La principal limitación de

Filtro de Bayer:



los CFA es que la resolución espacial de color es inferior a la resolución nominal del sensor. Para generar una imagen en color completa (donde cada píxel tenga valores R, G y B), es necesario aplicar un algoritmo de interpolación cromática o *demosaicing*. Este proceso estima los componentes de color faltantes en cada píxel basándose en la información de los píxeles vecinos, lo que puede introducir artefactos como el *moiré* o falseado de color (*color aliasing*).

II.3. Representación de imágenes

II.3.1. La imagen digital

Una imagen digital es una representación bidimensional de una escena que ha sido **muestreada espacialmente y cuantificada en amplitud**. Esto significa que está definida en:

- Un dominio espacial discreto: un número finito de posiciones (píxeles) organizadas en una rejilla regular (matriz $M \times N$).
- Un rango de valores discreto: las intensidades solo pueden tomar un conjunto finito de valores (generalmente potencias de 2).

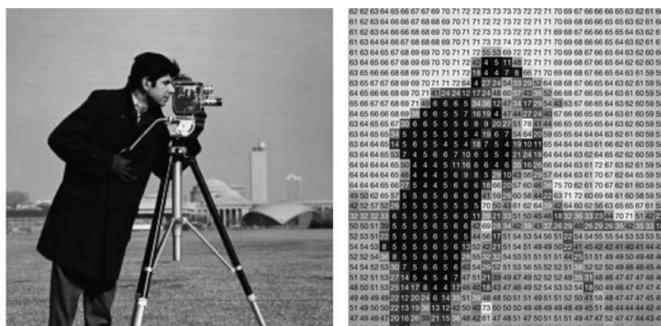
El **píxel** (elemento de imagen) es la unidad mínima de información espacial en una imagen digital. Cada píxel almacena un valor (o un conjunto de valores) que representa la intensidad luminosa y/o el color capturado por el sensor en esa posición específica. Así, una imagen digital puede representarse matemáticamente como una matriz de valores numéricos.

II.3.2. True Color

El término True Color se refiere a un método de representación de color que utiliza una codificación directa de los componentes cromáticos, típicamente capaz de representar más de 16 millones de colores distintos.

Es crucial hacer una distinción conceptual:

- **Imagen en Escala de Grises (Luminancia):** Se almacena un único valor por píxel, que representa la luminancia (la medida física de la intensidad de la luz), no la percepción subjetiva del brillo. Con una profundidad de 8 bits por píxel (bpp), se pueden representar 256 (2^8) niveles de gris, donde 0 suele representar el negro absoluto y 255 el blanco máximo.



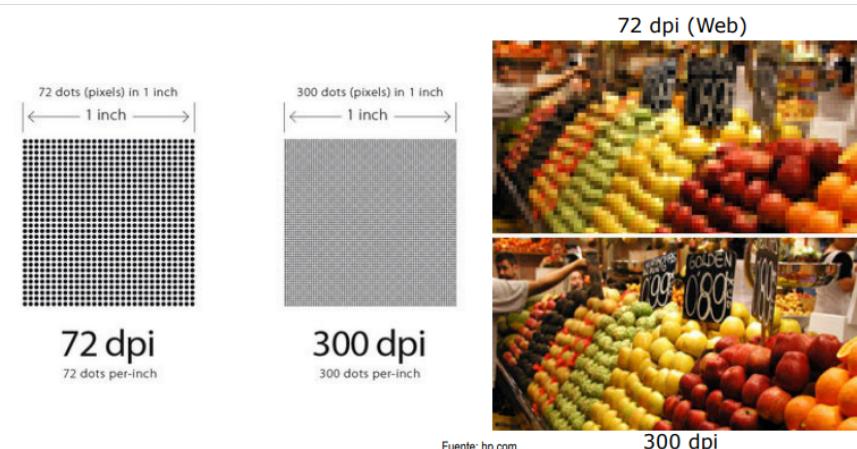
- **Imagen True Color a Color:** Utiliza tres canales independientes (generalmente Rojo, Verde y Azul - RGB). Cada canal tiene una profundidad de 8 bits, resultando en un total de 24 bits por píxel ($8 + 8 + 8$). Esto permite representar $2^{24} = 16.777.216$ colores distintos. Este tipo de imagen no es una matriz bidimensional simple, sino una estructura tridimensional de tamaño $M \times N \times 3$, a menudo conceptualizada como "un cubo de información" donde cada "capa" corresponde a un canal de color.



II.3.3. Resolución vs definición

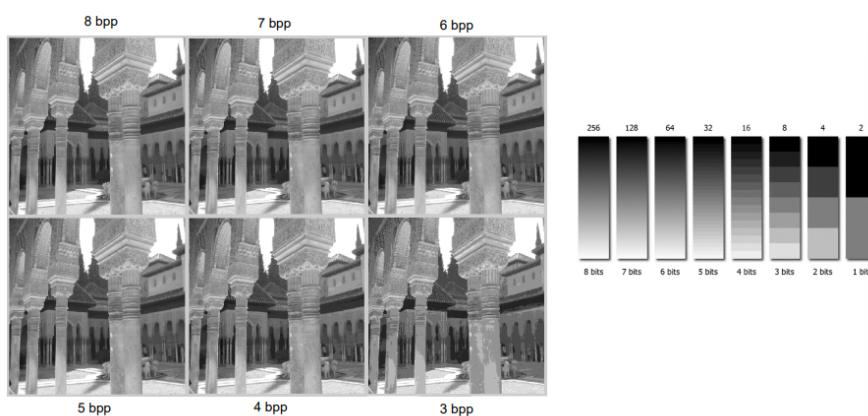
La **resolución de la imagen (Pixel Dimensions)** se refiere al número absoluto de píxeles que componen la imagen en sus dimensiones de ancho y alto (e.g., 1920×1080 px). Es un atributo intrínseco del archivo digital y determina la cantidad de detalle espacial que la imagen contiene. El usuario puede seleccionarla durante la captura o el post-procesado (remuestreo).

La **Definición o Densidad de Píxeles (PPI - Pixels Per Inch / DPI - Dots Per Inch)** es una medida de densidad que relaciona la resolución en píxeles con un tamaño físico real. Indica cuántos píxeles (PPI para pantallas) o puntos de tinta (DPI para impresión) hay en una pulgada lineal. Este valor define cómo de grandes o pequeños se verán los píxeles al reproducir la imagen en un dispositivo de salida (monitor, impresora). Está intrínsecamente ligado a la calidad del proceso de captura (óptica, sensor) y a las técnicas de procesamiento (interpolación, submuestreo) utilizadas.



II.3.4. Profundidad

La profundidad de color se mide en bits por píxel (bpp) y determina cuánta información puede almacenar cada píxel, es decir, cuántos colores o tonos de gris diferentes puede representar una imagen.



Aprovechando la menor sensibilidad del sistema visual humano a la información de color (crominancia) en comparación con la información de luminancia, se han desarrollado formatos que permiten comprimir imágenes reduciendo selectivamente los datos de color. Una alternativa al almacenamiento True Color (24 bpp) son las **imágenes indexadas**. En lugar de almacenar los tres valores RGB para cada píxel, se utiliza una paleta de colores o tabla de búsqueda (Color Look-Up Table - CLUT o Color Map). Esta paleta es un array de hasta 256 entradas (para 8 bpp) donde cada entrada contiene un color RGB de 24 bits. La imagen en sí misma no almacena colores, sino índices (valores de 0 a 255) que apuntan a una posición en la paleta. Esto reduce drásticamente el tamaño del archivo. Se almacena una matriz de $M \times N$ de 8 bits (los índices) y una pequeña tabla auxiliar de 256×3 bytes (la paleta), en lugar de tres matrices de $M \times N$ de 8 bits. No obstante, limita la imagen a un máximo de 256 colores simultáneos ($2^{n\text{bits}}$), lo que puede producir posterización (*banding*) en imágenes con degradados suaves o muchas variaciones de color. Es ideal para gráficos con áreas planas de color.

II.3.5. Formatos

Hay muchos formatos de imagen:

- Sin compresión: BMP, RAW, PPM
- Compresión sin pérdidas: PNG, GIF, TIFF
- Compresión con pérdidas: JPEG, TIFF

Capítulo III

Procesado imágenes digitales

III.1. Introducción

El preprocesado y la mejora de imágenes tienen como objetivo:

- Realzar o mejorar el contraste.
- Recortar (*cropping*) o seleccionar regiones de interés.
- Registrar: ajustar la geometría de una imagen para alinearla con otra.
- Reducir o eliminar ruido y artefactos.
- Corregir el desenfoque (*deblurring*) o reconstruir áreas faltantes (*inpainting*).

Un operador puntual se define de la siguiente manera: Dada una imagen de entrada, el operador calcula cada píxel de la imagen de salida únicamente en función del valor del píxel en la misma posición (x, y) de la imagen de entrada, mediante una función matemática.

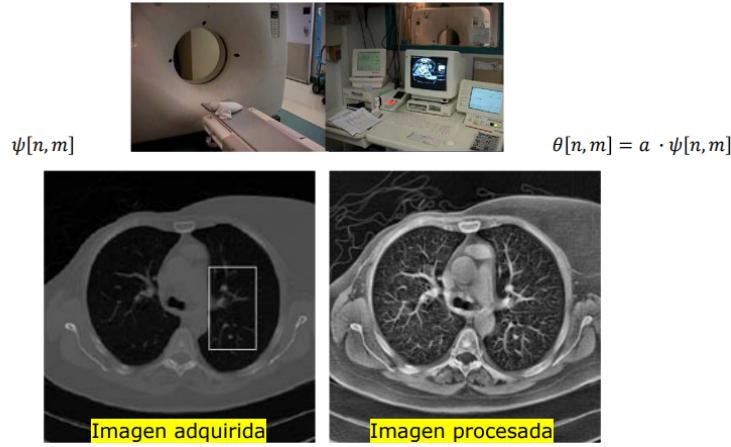
Un operador local, en cambio, calcula cada píxel de salida en función del píxel en la misma posición y de los píxeles de su vecindario en la imagen de entrada. La dependencia del vecindario está implícita en la función del operador.

III.2. Operadores puntuales

III.2.1. Introducción

Una transformación píxel a píxel donde el vecindario considerado es de 1×1 , es decir, el propio píxel. Esto implica que las coordenadas espaciales del píxel son irrelevantes para la transformación. Se define como una transformación de los valores de intensidad de entrada, r_k , a valores de salida, s_k . Así, es un operador que modifica el nivel de gris de cada píxel individualmente. Estos operadores alteran la amplitud de los píxeles de acuerdo con una operación específica, lo que puede expandir la escala de grises para mejorar la visualización de detalles.

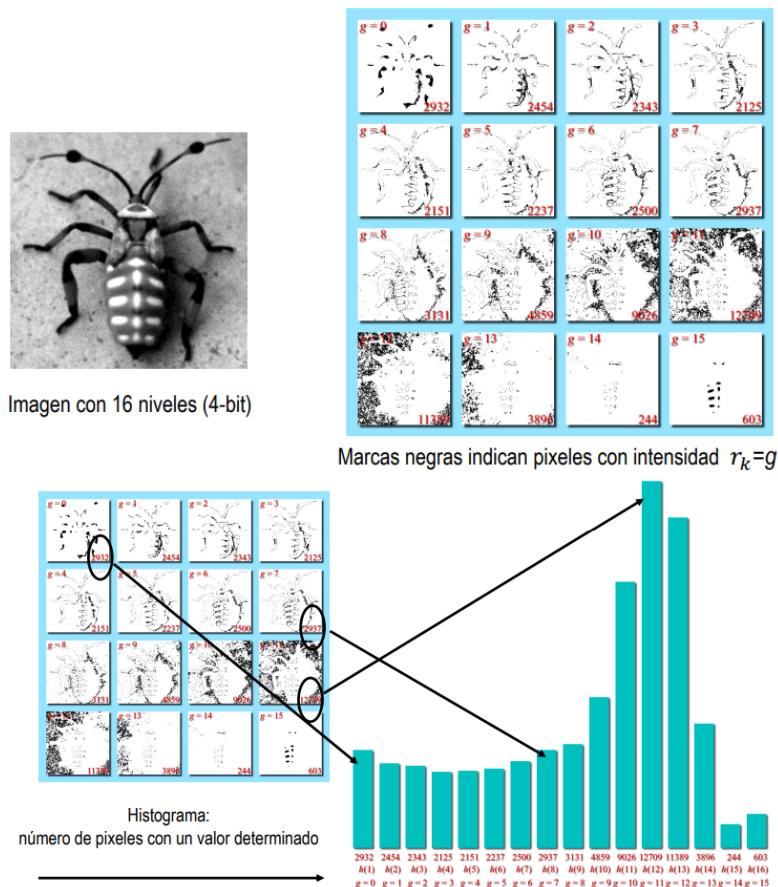
Ejemplo (mejora de la imagen)



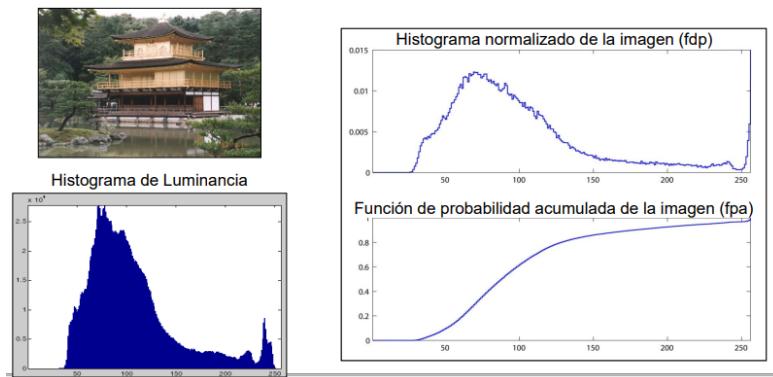
III.2.2. Modelado de histograma

El histograma de una imagen $\psi[n, m]$: $h(r_k)$ denotado como $h(r_k)$, es una representación gráfica de la frecuencia de los niveles de gris. $h(r_k)$ indica el número de píxeles que tienen el valor r_k .

Mientras la imagen original es bidimensional (coordenadas n, m), el histograma es unidimensional, ya que solo representa la distribución de frecuencias de los niveles de gris, sin información espacial.



El histograma normalizado, obtenido al dividir cada valor del histograma por el número total de píxeles (ancho x alto de la imagen), estima la **función de densidad de probabilidad (FDP)** de los niveles de gris. La suma de todos sus valores es 1. Al acumular estos valores, se obtiene la **función de distribución acumulativa (FDA)** de la imagen, que siempre culmina en 1.

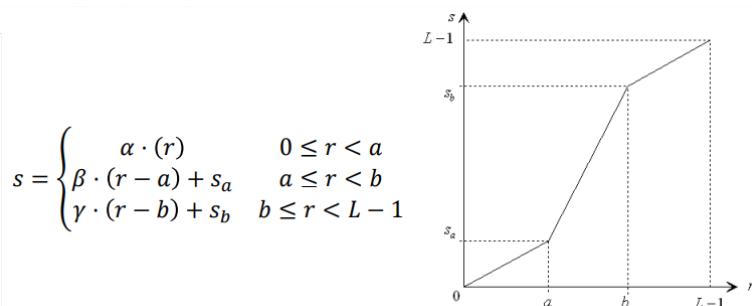


Una limitación importante del histograma es que **no contiene información espacial** sobre la disposición de los píxeles en la imagen.

III.2.2.1. Ajuste de contraste

El objetivo es realizar imágenes con bajo contraste, causado por condiciones de baja iluminación, un rango dinámico limitado del sensor o errores en la configuración de la cámara (como la apertura del diafragma). La solución consiste en expandir o comprimir el rango dinámico de las intensidades según sea necesario.

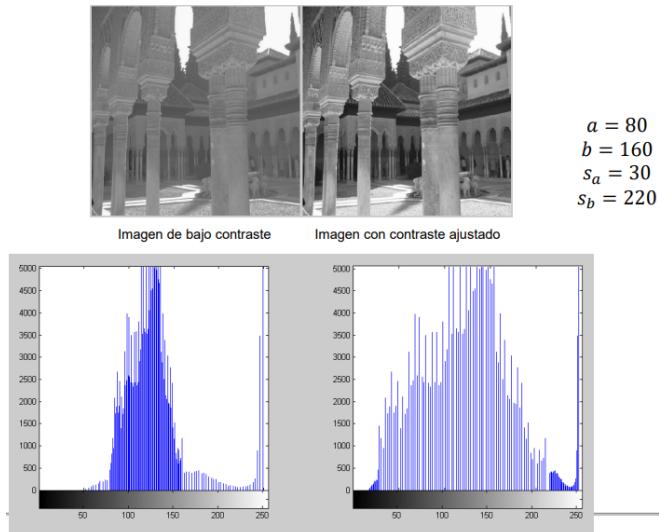
Para expandir el contraste, se asigna un rango más amplio de valores de salida a un rango estrecho de valores de entrada en el histograma. Para reducir el contraste, se hace lo contrario. El objetivo es aprovechar todo el rango disponible de niveles de gris.



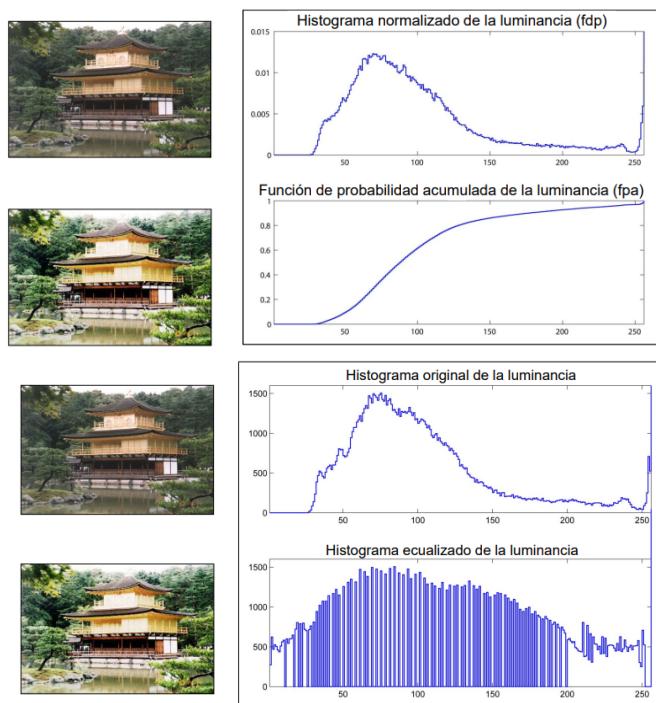
En el ejemplo, un valor de entrada de 80 se transforma en 30, y un valor de 160 en 220. Esto expande el rango central de intensidades y comprime los extremos, logrando una mejor distribución del contraste.

III.2.2.2. Igualación o ecualización (*equalization*)

La ecualización busca que la imagen resultante tenga un histograma lo más uniforme posible, es decir, que todos los niveles de gris tengan una frecuencia similar. Para ello,



se utiliza la función de distribución acumulativa (FDA) del histograma normalizado como función de transformación. El efecto es una mejora continua y automática del contraste, a diferencia del ajuste por tramos.



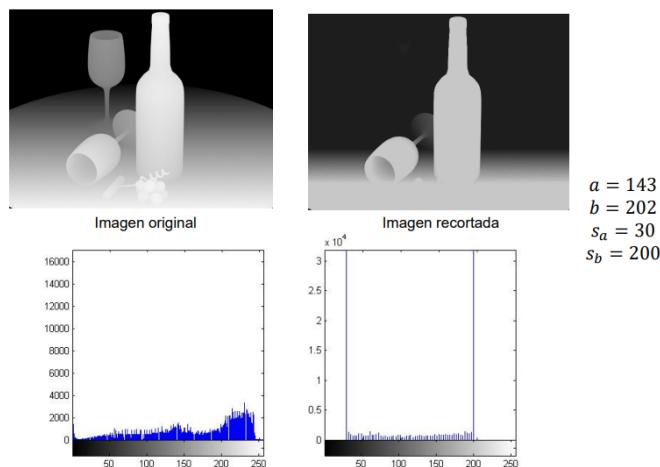
La nueva intensidad de un píxel se calcula aplicando su FDA al rango máximo de grises (ej., 255). Por ejemplo, si la FDA de un nivel de gris es 0.6, su nuevo valor será $0.6 \cdot 255 \approx 153$.

III.2.3. Modificación de niveles

Estas técnicas permiten seleccionar información específica o crear efectos especiales modificando el histograma indirectamente. Los tipos principales son:

- **Recorte (*clipping*):**

Similar al ajuste de contraste, pero "recorta" los valores por encima y por debajo de unos umbrales, asignándoles el valor máximo o mínimo. Las pendientes fuera del rango de interés se vuelven cero. Es útil para aislar regiones de interés o eliminar ruido cuando se conoce el rango válido de intensidades.



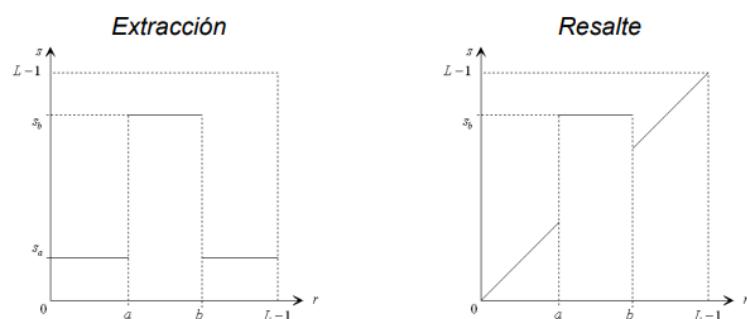
Se utiliza, por ejemplo, en sistemas de ayuda a la conducción para ignorar obstáculos lejanos, o en robótica para limitar la percepción a distancias relevantes.

- **Negativo o Inversión del eje de intensidades:**

Invierte los valores de intensidad (el blanco se vuelve negro y viceversa). Es útil en el análisis de imágenes médicas, como radiografías, para cambiar la perspectiva visual. El histograma resultante es la imagen especular del original.

- **Seccionado de niveles (*slicing*):**

Busca aislar una banda específica de niveles de gris, ya sea para extraerla o para resaltarla sobre el fondo.



III.3. Operadores locales

Los operadores locales efectúan una transformación en el dominio espacial, conocida como *filtrado espacial*.

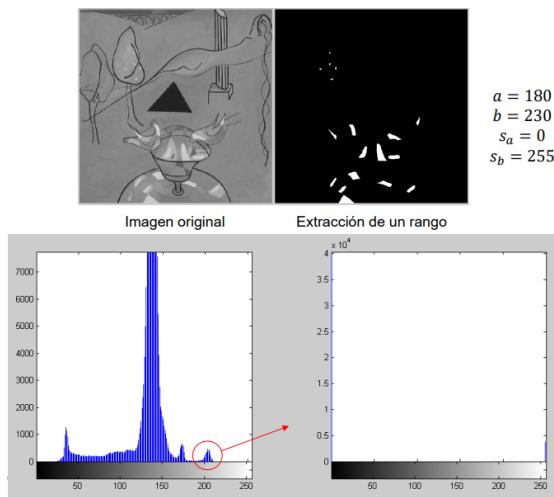


Figura III.1: Ejemplo de extracción. Los píxeles dentro del rango seleccionado se muestran en blanco; el resto en negro.

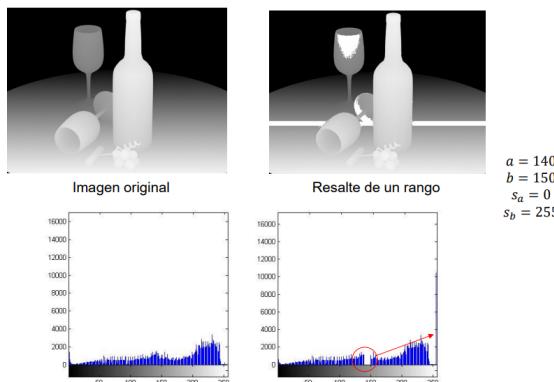


Figura III.2: Ejemplo de resalte. Una banda específica (ej., 140-150) se resalta en blanco. Puede usarse en sistemas de asistencia al conductor para marcar los límites de la distancia de seguridad.

III.3.1. Filtrado espacial

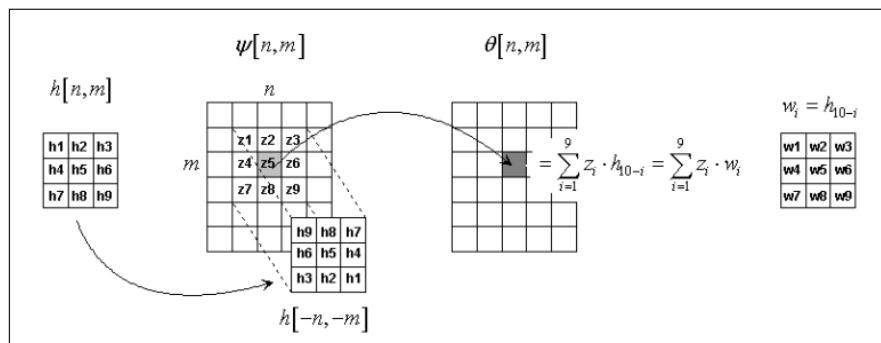
La operación de convolución se expresa como:

$$\theta[n, m] = \psi[n, m] * h[n, m] = \sum_{k=-1}^a \sum_{l=-b}^b \psi[k, l] \cdot h[n - k, m - l]$$

Para calcular $\theta[n, m]$, se centra la máscara invertida $h[-n, -m]$ sobre el píxel $\psi[n, m]$ y se calcula la suma de los productos de los píxeles de la vecindad por los coeficientes correspondientes de la máscara.

La convolución de una imagen con un impulso (delta) desplaza la imagen. Por ejemplo, una máscara con un único '1' en una esquina desplazará la imagen. Si la máscara contiene varios impulsos, el resultado es una superposición de imágenes desplazadas y promediadas, lo que genera un efecto de borronamiento.

Si en lugar de guardar solo 1 píxel de la matriz guardamos tres impulsos, se produce un desplazamiento y superposición. Se deben dividir entre 3 por normalización, y se



guardan el píxel central, el píxel arriba a la izquierda y el píxel abajo a la derecha. Como se promedia, el resultado es una imagen movida emborronada.

Ahora guardamos 5 valores, las 4 esquinas y el valor central. La imagen resultante sigue estando movida. Esto se debe a que usualmente el vecindario no es tan grande. Si en lugar de utilizar un vecindario 16×16 se utiliza un vecindario 3×3 , la convolución se convierte en un filtro de emborronado o suavizado que a simple vista no se aprecia, pero con zoom sí se aprecia que los bordes son más suaves al estar promediados.

En la práctica, se utilizan máscaras pequeñas (ej., 3×3). La convolución con una máscara de promediado produce un suavizado o desenfoque, haciendo los bordes menos nítidos.

Un desafío al aplicar convoluciones es el tratamiento de los bordes, donde la máscara se extiende más allá de la imagen. Las estrategias comunes para manejar esto son:

- **Relleno de ceros (Zero padding):** Los píxeles fuera de la imagen se consideran 0.
- **Réplica de píxeles (Replication):** Se replican los valores de los píxeles del borde. Esta opción suele preferirse en imágenes médicas.
- **Convolución circular:** La imagen se trata como periódica, tomando píxeles del lado opuesto.

III.3.2. Suavizado: lineal y no lineal

Los filtros de suavizado se utilizan para reducir ruido o difuminar detalles. En los filtros lineales, los coeficientes de la máscara son positivos y suman 1 para mantener el nivel de brillo promedio.

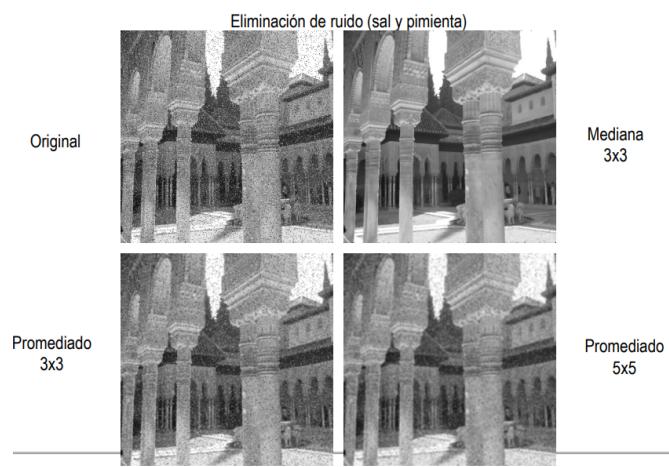
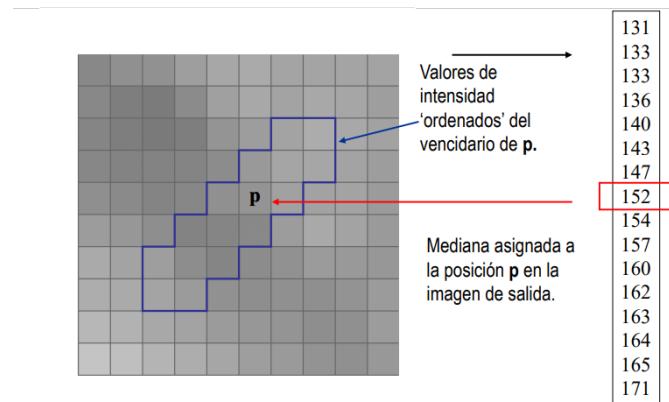
Los diseños comunes incluyen:

- **Filtro de media (averaging):** Todos los coeficientes de la máscara son iguales.
- **Filtro de media ponderada (weighted average):** Los coeficientes son mayores en el centro, dando más peso al píxel central y a sus vecinos más cercanos. Un caso particular es el filtro binomial, que es separable y produce un suavizado más gradual.

La máscara de filtrado promediado ponderada es un caso particular de filtro binomial, familia de filtros separables resultante de la aplicación sucesiva y en ambas dimensiones de la máscara. El promediado da un resultado más borroso que el binomial.

El filtrado no lineal no se basa en una suma ponderada. Preserva mejor los bordes y es efectivo contra ruido impulsivo (como ruido sal y pimienta). Un caso paradigmático es el filtro de mediana, que reemplaza el valor del píxel central por la mediana de los valores en su vecindario. Otros filtros de estadísticas de orden son el de mínimo y el de máximo.

Filtro de mediana: en lugar de mezclar por igual, se aplica la función de la mediana. Se define un píxel central y un vecindario, y al vecindario se aplica la mediana. Eso mismo se puede hacer con el valor mínimo (filtro de mínimo), máximo (filtro de máximo) y filtro de posición. El filtro de mediana funciona muy bien para el ruido de sal-pimienta (puntos blancos y negros aleatorios).



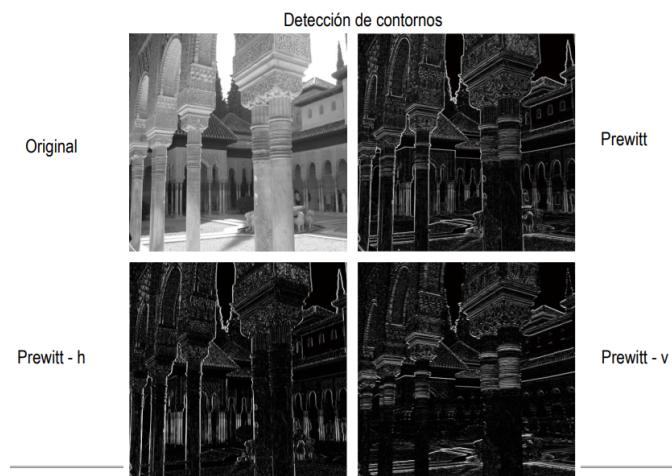
III.3.3. Realce: lineal

Los filtros de realce resaltan detalles finos y bordes. Mientras el suavizado se asocia con la integración (promediado), el realce se relaciona con la derivación, que responde fuertemente a las discontinuidades (bordes) y al ruido.

En imágenes discretas, las derivadas se aproximan mediante diferencias. La primera derivada se implementa típicamente calculando la magnitud del gradiente. Los

operadores más comunes para esto en máscaras 3x3 son los operadores de Prewitt, Sobel y Roberts.

El **operador de Prewitt** es un filtro de derivada parcial utilizado para calcular una aproximación del gradiente de la imagen en cada punto, lo que permite resaltar los bordes. Se compone de dos máscaras o kernels 3x3: una para calcular la derivada en la dirección horizontal (G_x) y otra para la dirección vertical (G_y).



Capítulo IV

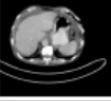
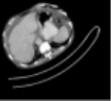
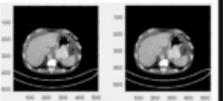
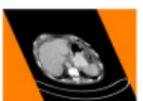
Registro de imágenes

El registro de imágenes hace referencia a alinear en un mismo sistema de coordenadas unas imágenes. Esto tiene varias aplicaciones. Si en un mismo sujeto adquieres varios tipos de imagen en distintos puntos temporales, se puede alinear para ver la evolución temporal de esa zona. También se puede hacer entre sujetos para automatizar un estudio. Además, cada tipo de imagen médica (TAC, PET) puede proporcionar una información distinta, por lo que al alinearlos se puede tener una visión más completa. Por todo esto, el proceso de alinear imágenes es diferente en función de lo que se quiera conseguir. Así, los tipos de abordajes de registro dependen de lo que se busque. En general, se busca minimizar la diferencia entre la imagen de entrada y la salida, o maximizando la similitud.

IV.1. Tipos de registro

Hay distintos tipos de registro:

- **Registro rígido:** El registro rígido solo permite rotar, manteniendo distancias y ángulos, y mover en el plano.
- **Registro de similaridad:** Se añade un multiplicador de escala. Puede rotar, mover y ampliar. Esta transformación escala todas las direcciones en la misma medida, conserva los ángulos y las proporciones relativas, e incluye rotación y traslación.
- **Transformación afín:** Además de todo lo anterior, permite el shearing, que es la alteración de la imagen al modificar un solo eje. Mientras que el escalado mantiene la forma, el shearing no. No es uniforme, ya que puede tener distintas escalas en cada eje.
- **Registro deformable:** Permite deformaciones locales que puede ser distinto para cada píxel.
- **Registro longitudinal:** Alinea múltiples imágenes del mismo sujeto a lo largo del tiempo, y puede seguir un registro rígido o deformable dependiendo de si se deben detectar cambios sutiles sin introducir deformaciones artificiales.

Identity $T_g = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ 	Scaling $T_g = \begin{bmatrix} \alpha_x & 0 & 0 \\ 0 & \alpha_y & 0 \\ 0 & 0 & 1 \end{bmatrix}$ 	Rotation $T_g = \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$ 
Translation $T_g = \begin{bmatrix} 1 & 0 & \delta_x \\ 0 & 1 & \delta_y \\ 0 & 0 & 1 \end{bmatrix}$ 	y - shearing $T_g = \begin{bmatrix} 1 & 0 & 0 \\ \zeta_y & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ 	x - shearing $T_g = \begin{bmatrix} 1 & \zeta_x & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ 

- **Registro multimodal:** Alinea imágenes obtenidas mediante diferentes técnicas de imagen, como la resonancia magnética (estructural) y la tomografía por emisión de positrones (funcional). Las diferentes modalidades pueden tener contrastes distintos; entre las métricas de similitud comunes se incluyen la información mutua y la correlación cruzada. El registro multimodal preciso es fundamental para fusionar la información anatómica y molecular (por ejemplo, planificación quirúrgica, localización de tumores).

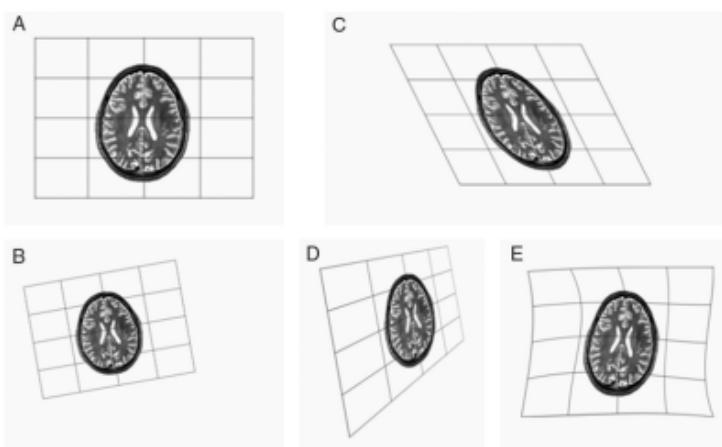


Figura IV.1: Un registro de imágenes médicas utilizando diferentes matrices de transformación. A) Una imagen original. B) Transformación de rotación y traslación. C) Transformación de rotación, traslación y cizallamiento. D) Transformación proyectiva. E) Transformación deformable.

IV.2. Avances clave en IA en registro de imágenes

La IA está mejorando significativamente el registro y el co-registro de imágenes médicas al hacer que el proceso sea más rápido, más preciso y más robusto en entornos clínicos y de investigación.

Los modelos de aprendizaje profundo, como las redes neuronales convolucionales (CNN) y los modelos basados en transformadores, han revolucionado el registro de imágenes al aprender directamente cómo alinear imágenes a partir de ejemplos, saltándose el complejo diseño manual de características y la optimización iterativa.

Las arquitecturas CNN, como U-Net, predicen campos de deformación densos que describen los desplazamientos espaciales píxel a píxel necesarios para alinear las imágenes de forma rápida y precisa.

Los modelos basados en transformadores aportan la capacidad de aprender relaciones globales y dependencias contextuales en toda la imagen, lo que mejora la precisión del registro, especialmente en el caso de variaciones anatómicas grandes o complejas.

Los modelos de IA pueden realizar registros rígidos (traslación, rotación) y deformables (deformación no lineal), lo que ayuda a alinear órganos o tejidos que pueden haber cambiado de forma entre exploraciones.

A diferencia de los métodos de registro clásicos que se basan en procesos iterativos y métricas elaboradas manualmente, los algoritmos de IA realizan predicciones en un solo paso, lo que acelera enormemente el proceso de alineación y lo hace escalable para los flujos de trabajo clínicos.

Los enfoques de registro de IA se generalizan mejor en imágenes multimodales (por ejemplo, de PET a MRI) utilizando funciones de similitud aprendidas, superando las limitaciones de las métricas tradicionales basadas en la intensidad que tienen dificultades con los diferentes contrastes de las imágenes.

Estos avances permiten aplicaciones como la monitorización de la progresión de enfermedades, la planificación de tratamientos y estudios de investigación que requieren una alineación espacial precisa de grandes conjuntos de datos de imágenes.

IV.3. Artículo

Evaluación de un procedimiento de registro conjunto de imágenes multimodales (TC, RM y PET) en fantomas y pacientes con cáncer de cabeza y cuello: precisión, reproducibilidad y consistencia

En el artículo busca mejorar la radioterapia en tumores. Se realiza una radiografía para ver el tumor y poder crear una terapia dirigida. Hay que tener en cuenta la intensidad de la radiación y la geometría del haz y dónde va a impactar en la anatomía del tumor. Para ello, se validan unas técnicas de corregistro entre distintas imágenes para delimitar lo mejor posible el volumen tumoral. Es un trabajo multi-técnica al utilizar PET (tomografía por emisión de positrones), TAC y resonancia magnética. Esto lo prueban en un phantom (tejido simulado para probar la técnica) y en pacientes.

El objetivo del estudio era validar un método interactivo de corregistro rígido basado en segmentación de superficies para fusionar imágenes de CT, MRI y PET. Como sujetos se utilizaron fantomas diseñados con estructuras que simulan el cuello y 4 pacientes con carcinoma de células escamosas faríngeo-laringeo.

Como referencia se utilizaba la tomografía computerizada (CT). Un observador ajustaba manualmente la traslación y rotación de una imagen sobre dicha referencia hasta lograr una superposición visual óptima. Los parámetros de transformación se guardaban.

El accuracy se definió como la diferencia entre la transformación aplicada por un observador y el valor promedio de todas las transformaciones realizadas por todos los observadores para la misma modalidad de imagen. Se expresó como 2 desviaciones estándar de la distribución de las traslaciones y rotaciones. También se usó el vector euclíadiano para combinar traslaciones en x, y, z y obtener un desplazamiento total en mm.

Se utilizó la reproducibilidad para medir cuán consistentes son los resultados cuando el mismo proceso lo realiza diferente gente o la misma persona en diferentes momentos. Se calculó la variación intra-observación (varianza de los resultados de coregistración de una misma persona 4 veces con días de separación) y variación inter-observador (4 observadores diferentes). Se comparó esta varianza con la Varianza Residual (la varianza total que no se puede explicar por el observador), para ver si el factor "observador" era significativo. Las variaciones intra e inter-observador fueron muy pequeñas (varianzas $< 0.25 \text{ mm}^2$) y muy inferiores a la varianza residual. Esto significa que el factor "observador" es insignificante y que cualquier clínico entrenado puede obtener resultados similares.

Finalmente, la consistencia medida si el método produce resultados coherentes cuando se aplica a diferentes tipos de imágenes del mismo sujeto. Para ello, se compararon los parámetros de coregistro obtenidos para las secuencias MRI T1 y MRI T2 de un mismo paciente. Como ambas se adquieren en el mismo equipo sin mover al paciente, deberían coregistrarse con los mismos parámetros. Cualquier diferencia indica una inconsistencia del método. No hubo diferencias significativas al coregistrar MRI T1 vs. T2, excepto en una dirección (y), lo que sugiere que el método es robusto frente a cambios en el contraste de la imagen.

IV.4. Imágenes, mapas y templates

En el procesamiento de imágenes biomédicas conviene establecer una clara diferenciación entre lo que constituye una **imagen** y lo que representa un **mapa**. Las imágenes corresponden a datos directos adquiridos por los equipos de imagenología, como tomografías computarizadas, resonancias magnéticas o estudios de tomografía por emisión de positrones. Estos datos pueden presentarse en forma cruda o reconstruida, conteniendo información anatómica o funcional directa del paciente.

Por otro lado, los mapas representan datos derivados obtenidos mediante procesamiento matemático avanzado. Se construyen a partir de múltiples imágenes y encapsulan parámetros fisiológicos o físicos cuantitativos, proporcionando una representación numérica de propiedades tisulares específicas que no son directamente observables en las imágenes originales.

IV.4.1. Imágenes de Difusión en Neuroimagen

Las imágenes de difusión por resonancia magnética se fundamentan en el movimiento browniano de las moléculas de agua en los tejidos biológicos. Desde la perspectiva física, esta técnica emplea gradientes de campo magnético específicamente diseñados para sensar el desplazamiento molecular. El movimiento de las moléculas de agua modula la señal de resonancia magnética de manera característica, permitiendo cuantificar las restricciones a la difusión mediante la atenuación de la señal observada.

En la práctica clínica, las imágenes de difusión han demostrado un valor incalculable para la detección de isquemia cerebral aguda. Durante un evento isquémico, se produce una restricción significativa de la difusión molecular que se manifiesta como una hiperintensidad característica en las imágenes ponderadas por difusión. Esta propiedad convierte a esta modalidad de imagen en una herramienta esencial para el diagnóstico temprano de accidentes cerebrovasculares, permitiendo intervenciones terapéuticas oportunas.

IV.4.2. Construcción de Mapas de Difusión

La transición desde imágenes de difusión hacia mapas paramétricos implica un proceso de reconstrucción matemática sofisticado. El mapa de Coeficiente de Difusión Aparente (ADC) se obtiene mediante un ajuste por mínimos cuadrados realizado voxel por voxel, aplicando el modelo exponencial $S(b) = S(0) \cdot \exp(-b \cdot ADC)$, donde $S(b)$ representa la intensidad de señal con el gradiente de difusión aplicado, $S(0)$ denota la intensidad de señal sin gradiente de difusión, b corresponde al factor de difusión expresado en s/mm^2 y ADC simboliz el coeficiente de difusión aparente en mm^2/s .

El flujo de procesamiento comienza con la adquisición de múltiples imágenes de difusión utilizando diferentes valores del factor b . Posteriormente, se ejecuta la reconstrucción paramétrica mediante ajuste exponencial para generar el mapa de ADC voxel por voxel. La fase final incluye la validación y control de calidad del mapa resultante, asegurando la confiabilidad de las mediciones cuantitativas obtenidas.

IV.4.3. Registro de imágenes y templates

En el contexto del análisis espacial estandarizado, los templates representan imágenes de referencia que definen espacios coordenados normalizados, como los sistemas MNI o Talairach. Estos templates poseen resoluciones espaciales predefinidas, típicamente de $1mm^3$ o $2mm^3$, y sirven como base fundamental para análisis multi-sujeto y comparaciones poblacionales.

El proceso de registro espacial implica la alineación geométrica de imágenes individuales con el template de referencia. Desde una perspectiva práctica, es la imagen individual la que normalmente se somete a re-muestreo para emparejar la resolución del template, estrategia conocida como *downsampling*. Esta aproximación se justifica por el hecho de que los algoritmos de registro funcionan de manera óptima cuando las imágenes involucradas presentan resoluciones espaciales similares. Cabe destacar que en situaciones excepcionales donde el template presenta mayor resolución que la imagen individual, podría aplicarse el proceso inverso de *upsampling*.

Capítulo V

Segmentación de imágenes

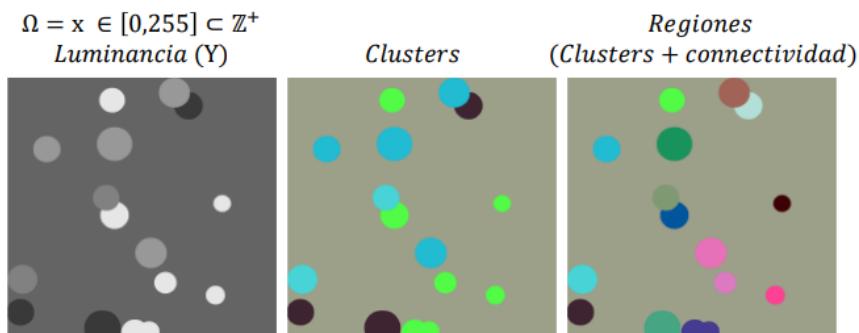
V.1. Introducción

La segmentación de imágenes tiene como objetivo la partición de una imagen en regiones o segmentos homogéneos con respecto a alguna característica (color, forma, distancia, tamaño) con el fin de simplificarla e identificar elementos de interés para una aplicación concreta. Toda la imagen está segmentada y la unión de los fragmentos recupera por completo la imagen.

En medicina, las aplicaciones de la segmentación incluyen la detección del borde coronario en angiogramas, medición del volumen tumoral y su respuesta a la terapia, mapeo funcional, clasificación automatizada de células sanguíneas, detección de microcalcificaciones en mamografías, extracción de imágenes del corazón a partir de cineangiogramas cardíacos, detección de tumores, etc.

La región Ω_j se define por un conjunto de píxeles "similares" de acuerdo a alguna característica definida sobre los píxeles de la imagen.

Al medir la luminancia y conectividad, se miden si los píxeles se tocan (son conexos) o no (están segmentados). En una conectividad de 4, los píxeles que son conexos son los de arriba y abajo con respecto al píxel en cuestión, mientras que una conectividad de 8 implica que todos los de alrededor (incluidos los de la diagonal) son conexos. La luminancia y conectividad nos permiten diferenciar correctamente.



Se utiliza una **etiqueta o label** como identificador de la región en la segmentación. Puede tener un número o color aleatorio y simplemente indica las regiones independientes y por tanto diferentes a las demás.

Un **representante** es un descriptor de las características de la región de igual dimensión (d) que el espacio de decisión. Se realiza la estadística de lo que se quiere representar (media, mediana, moda).

El **contorno** se define como el conjunto de píxeles alrededor de la región Ω_j en cuyo vecindario 9 hay al menos un píxel perteneciente a la región j y un píxel no perteneciente a la región j .

Sin embargo, esto está mal condicionado, ya que no hay un criterio para definir una región. Diferentes usuarios pueden considerar diferentes regiones y diferentes anotaciones. Además, el resultado depende de la escala y el conocimiento previo que se pueda tener ayuda a tomar las decisiones.

Hay distintos tipos de segmentación:

- **Clases:** separar objetos de su entorno, como por ejemplo separar una vaca del fondo de la imagen.
- **Instancias:** separar objetos individuales, para poder diferenciar por ejemplo dos personas juntas del fondo y entre ellas.
- **Partes:** se identifican las distintas partes de los objetos
- **Partes (alto nivel):** se siguen identificando las distintas partes de los objetos, pero sin tanto detalle, siendo así un punto intermedio entre instancias y partes.

V.2. Técnicas más representativas

Las técnicas más representativas se pueden clasificar en manual, semiautomáticas o automáticas, si son segmentación a bajo nivel o basada en modelo, o si son clásicas, basadas en estadísticos, difusas o redes convolucionales.

V.2.1. Umbralización (operador puntual)

La umbralización o *thresholding* es un caso particular de recorte para $a = b = T$. Así, la primera y última pendiente son nulas y la intermedia es absoluta. El threshold es un binario que divide la imagen en dos regiones, siendo así muy útil para imágenes bimodales.

Es útil para eliminar niveles cuando se sabe que el original sólo tiene dos y, en general, en la toma de decisiones binarias, como por ejemplo para finalizar la separación de objetos en procesos de segmentación.

Umbralización de imágenes bimodales Se utiliza un algoritmo que itera y, cuando la diferencia obtenida entre el nuevo umbral y el anterior es menor que la unidad, se detiene.

Método de Otsu Realiza una búsqueda exhaustiva del umbral que minimiza la varianza intra-clase. Busca el punto medio, pero teniendo en cuenta las distribuciones (variabilidad), permitiendo identificar objetos.

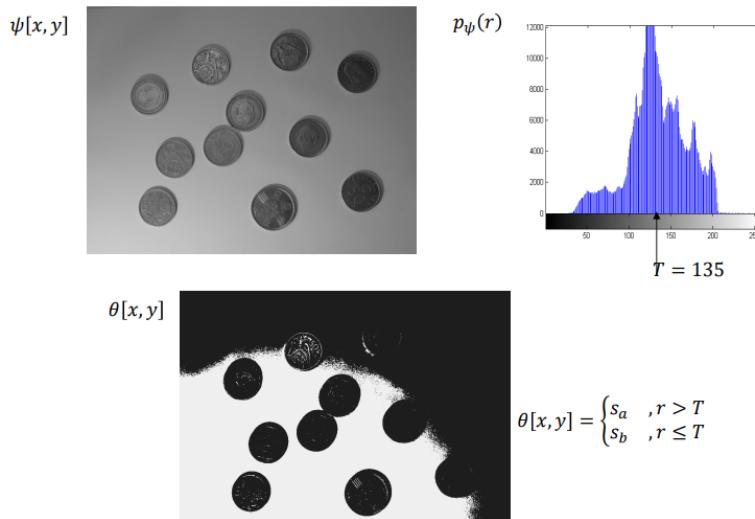
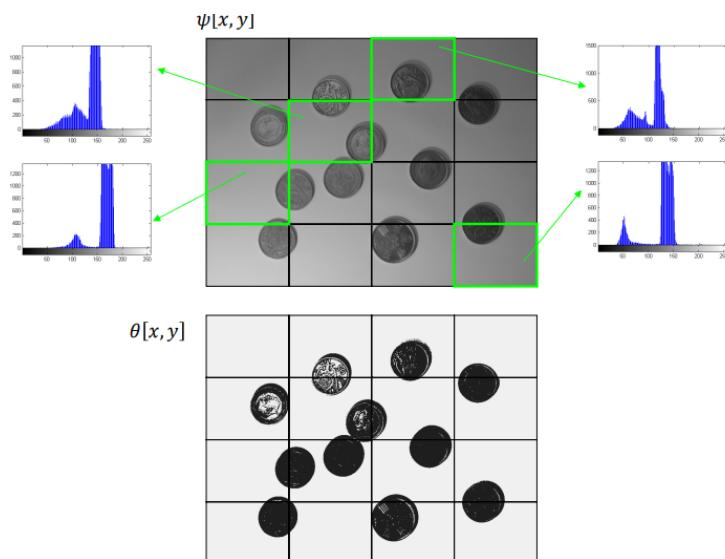


Figura V.1: Si se utiliza la varianza para la umbralización, la imagen resultante (inferior) es muy mala al no realizar la separación entre fondo y monedas bien. Tanto la umbralización de imágenes bimodales como el método de Otsu son binomiales, y para este caso se debe segmentar en ventanas y aplicar la umbralización a cada ventana.



Multimodal Para imágenes multimodales, se pueden aplicar los métodos anteriores bimodales repetidamente. De esta forma, primero se separan las dos regiones más grandes, y posteriormente de cada región se obtienen otras dos regiones contenidas en ella, etc. Esto se utiliza en la clasificación de tejidos mediante tomografía computarizada (CT). Mediante la consulta a expertos, se conoce la distribución y desviación de los tejidos. Estos datos se utilizan en algoritmos y se hacen medias dos a dos para poner un threshold.

V.2.2. Clustering (agrupamiento por regiones)

En el clustering, la imagen se divide entre K regiones más representativas. Las imágenes en color tienen un volumen de datos mayor, y el agrupamiento es difícil si se basa en el sistema visual humano.

Las aproximaciones son espacialmente "ciegas". La imagen se analiza como un conjunto de puntos d-dimensionales en el espacio de decisión. La técnica más representativa es **K-means**. Se busca la mejor frontera de separación entre los píxeles de sus vecinos más cercanos.

Inicialmente hay tantos núcleos como píxeles. Cada núcleo tiene su celda, y se irán fusionando celdas si la distancia es menor que X. Esto se repite hasta obtener el número de clústeres que se quieran. El clustering clásico no tiene en cuenta el contorno, pero en imágenes RGB son 5 las variables que hay que tener en cuenta (rojo, verde, azul, x, y). Los resultados dependen de la característica utilizada (feature space). Aunque nos regiones de la imagen sean del mismo color, se pueden segmentar en dos regiones utilizando similitud y proximidad.

El número de clústeres se escoge en función de lo que se busque ver.



V.2.3. Detección y unión de bordes (basados en contornos)

Se utiliza el detector de Canny, un algoritmo multietapa cuyo objetivo es localizar todos los bordes de una imagen con máxima precisión y sólo una vez cada borde. Las etapas se dividen en:

1. Suavizado de la imagen con un filtrado gaussiano. Esto sirve para que una imagen no se vea abrupta, sino continua.
2. Aproximación de la magnitud del gradiente y ángulo (Sobel/Prewitt).

Esto permite diferenciar los bordes en grados (ángulos de giro). Se hace la tangente y se calcula el ángulo. Saber el ángulo permite que, esté donde esté y tenga el grosor que tenga, se pueda obtener el punto medio analizando su perpendicular.

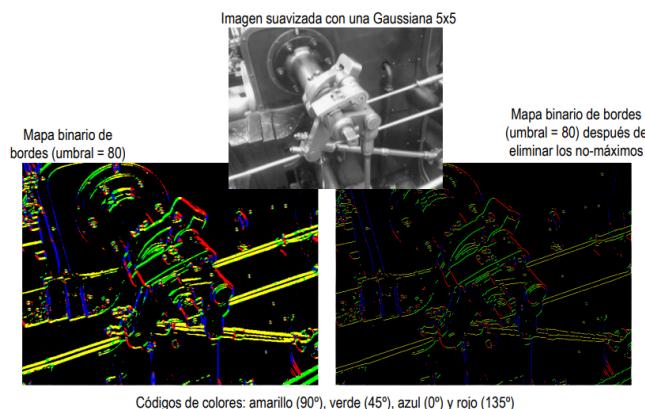
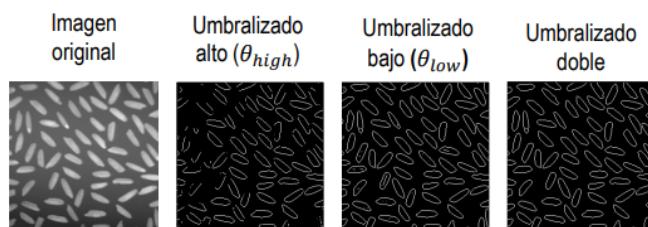


Figura V.2: La imagen izquierda tiene un grosor indefinido. La imagen derecha garantiza que todos tengan un borde de 1 píxel.

3. Umbralizado doble para detectar píxeles pertenecientes a bordes fuertes/débiles. Los bordes fuertes tienen grandes cambios de grises, mientras que los bordes suaves permiten cambios mínimos de grises.
4. Rechazo de píxeles en bordes débiles no conectados a bordes fuertes. En el umbralizado alto, sólo se mantienen los bordes con cambios de gris muy exagerados de manera que hay granos de arroz que no terminan de asomarse.



V.2.4. Contornos activos (basados en contornos)

Los contornos activos snakes son un método global para la búsqueda de los contornos de los objetos en el espacio de decisión usando la imagen como soporte.

Es un proceso iterativo que busca rodear un objeto de interés, como puede ser un tumor. La snake se muestrea en un conjunto de puntos de control y busca minimizar una función de energía que combina:

- Energía interna: controla la deformación de la snake para evitar el sobreajuste.
- Energía externa: controla el ajuste de la snake a los contornos para definir la calidad de la segmentación.
- Energía restrictiva: restricciones de diseño, generalmente en función de la aplicación (ad hoc) o para incrementar la robustez al ruido.

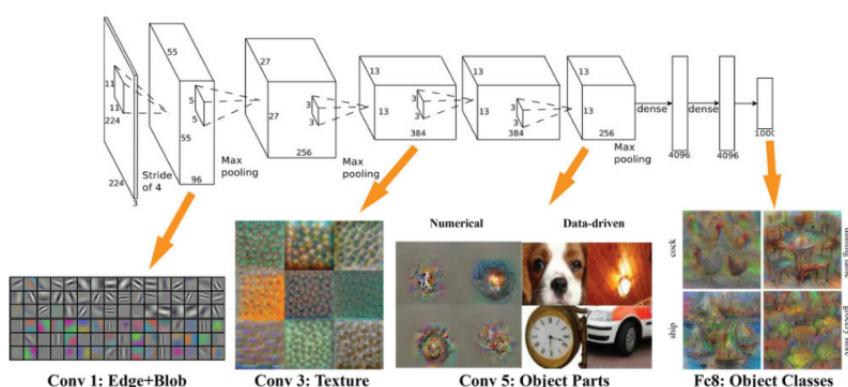
El ajuste de snake se hace sobre una imagen de gradiente donde la curva tiende a meterse o salir del valle. Cuantas más iteraciones se hagan, mayor detalle se consigue.

V.2.5. Redes convolucionales (deep learning)

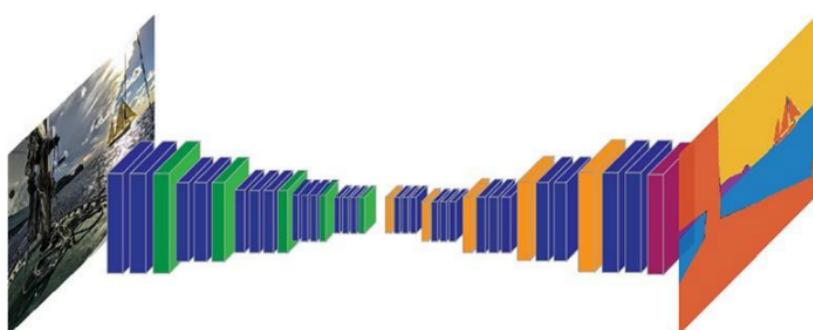
Las redes convolucionales (CNNs) son arquitecturas diseñadas para trabajar con datos estructurados espacialmente, como las imágenes médicas (microscopía, resonancia magnética, TAC, histología digital, etc.).

Un kernel (o filtro) es una pequeña matriz de pesos entrenables que se desplaza por la imagen (operación de convolución). Cada kernel extrae un tipo de información: bordes, texturas, formas, o patrones más abstractos conforme se avanza en las capas. Al principio, los kernels suelen detectar características de bajo nivel (bordes, contrastes), y en capas más profundas, rasgos de alto nivel (estructuras celulares, tejidos, órganos).

La idea es aplicar sucesivamente diferentes kernels entrenables de diferentes tamaños, diferentes niveles semánticos y combinar sucesivamente kernels a algunos más complejos. Al aplicar convoluciones y pooling (submuestreo), la imagen va reduciendo su tamaño (se pierde resolución espacial). A cambio, las representaciones internas ganan en complejidad y profundidad: el número de canales (o mapas de características) aumenta. Esto genera un “código comprimido” de la imagen, útil para reconocer patrones globales.



En clasificación, basta con decir si en la imagen hay un tumor o no. Pero en segmentación biomédica queremos una predicción a nivel de píxel (ej: qué píxeles son tumor, tejido sano, vasos sanguíneos, etc.). El **encoder** reduce la imagen extrayendo características relevantes. El **decoder** hace el proceso inverso, usando convoluciones transpuestas (deconvoluciones) o interpolaciones para recuperar la resolución original. Así, se obtiene un mapa de segmentación con la misma dimensión que la imagen inicial.



Capítulo VI

Aplicaciones del Procesamiento Digital de Imágenes: CT, PET & SPECT, Ultrasonido y Microscopía

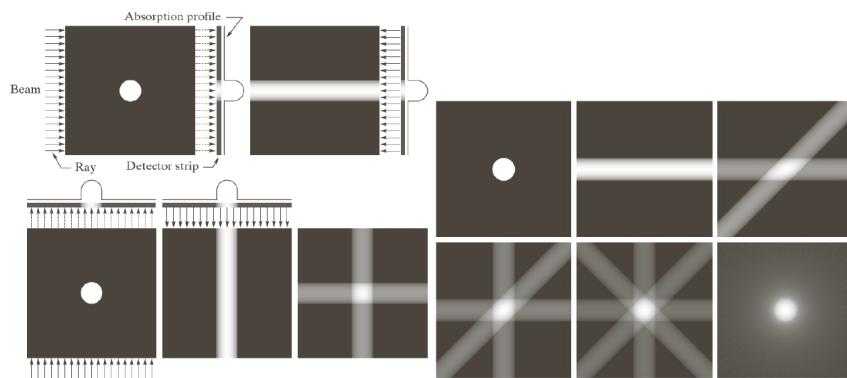
VI.1. Introducción y relación con la MRI

La resonancia magnética (MRI) utiliza campos magnéticos y pulsos de radiofrecuencia, produciendo imágenes de tejidos blandos con alto contraste. Al igual que la MRI, la tomografía computarizada (CT) y la tomografía por emisión de positrones (PET) emplean reconstrucción tomográfica (a partir de información 2D se genera información 3D), aunque con diferentes fundamentos físicos y datos. Comprender los principios físicos de cada modalidad ayuda a anticipar los desafíos del procesamiento digital de imágenes. Una pregunta relevante sería: ¿qué características de las imágenes en el procesamiento de MRI se compartirán o diferirán en CT o PET? Todas son imágenes que se toman *in vivo* de todo el cuerpo. En cuanto al procesamiento, todos tienen el problema del movimiento, teniendo que corregir los artefactos de movimiento. PET tiene una resolución espacial más grande, por lo que el movimiento puede notarse menos, pero sigue afectando. En resonancia, hay un sesgo por el campo no ajustado. Los tipos de sensores que utilizan CT y PET tienen que sentir una radiación, teniendo otro tipo de sensibilidad.

VI.2. Tomografía Computarizada (CT/TAC)

La CT utiliza una fuente de rayos X que rota alrededor del paciente, y los detectores miden la atenuación de los rayos X al atravesar los tejidos. Cada píxel de la imagen corresponde a una unidad de Hounsfield (HU) que representa la densidad del tejido.

En cuanto a las técnicas de reconstrucción, la retroproyección filtrada (FBP) aplica la transformada de Radon inversa para reconstruir la imagen a partir de proyecciones, mientras que los métodos iterativos mejoran el control del ruido y los artefactos, especialmente en exploraciones de baja dosis.



Las imágenes de CT presentan un alto contraste entre hueso y tejido blando, pero suelen ser más ruidosas que las de MRI. Los artefactos más comunes incluyen el **endurecimiento del haz** o *beam hardening*, la dispersión y los artefactos metálicos. El endurecimiento del haz ocurre cuando los rayos X de menor energía son absorbidos preferentemente por estructuras densas, como hueso o implantes metálicos, lo que altera la interpretación del detector y genera efectos como bandas o “oscurecimiento” central.

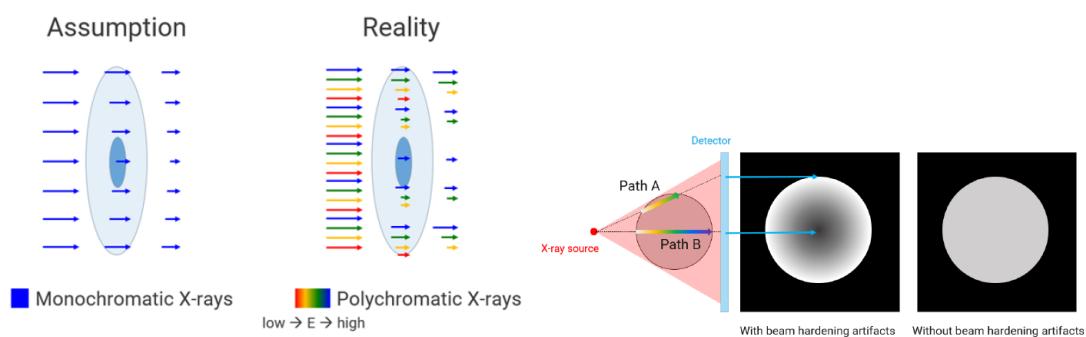


Figura VI.1: La longitud del camino es corto para el camino A y largo para el B. Esto hace que haya menos endurecimiento del haz en el camino A y poca energía resultante de rayos X, además de alta tasa de absorción calculada y alta densidad reconstruida (claro). El camino B tiene más endurecimiento, mucha energía de rayos X, poco ratio de absorción calculado y poca densidad (oscuro).

La **dispersión**, causada principalmente por el efecto Compton, desvía los rayos X de su trayectoria, reduciendo el contraste y aumentando el ruido. Finalmente, los **artefactos metálicos** aparecen como bandas brillantes u oscuras alrededor de implantes metálicos.

El procesamiento de imágenes CT requiere preprocesamiento cuidadoso, reducción de ruido y artefactos. La segmentación enfrenta dificultades para diferenciar tejidos blandos y patológicos, y los datos volumétricos 3D exigen algoritmos eficientes para renderizado y análisis. Clínicamente, la CT se usa ampliamente en traumatismos, enfermedades pulmonares y óseas, y guía la planificación quirúrgica o de radioterapia. La física de la atenuación de rayos X influye directamente en los tipos de ruido y artefactos esperados.

VI.3. PET y SPECT

La PET utiliza trazadores radiactivos que emiten positrones; estos se aniquilan con electrones, produciendo dos fotones de 511 keV detectados de forma coincidente por un anillo de detectores, permitiendo la reconstrucción tomográfica de la distribución del trazador, que refleja actividad metabólica o molecular. La SPECT, en cambio, usa trazadores que emiten fotones gamma simples, capturados mediante una cámara gamma que rota alrededor del paciente. Trazadores comunes incluyen Tecnecio-99m o Yodo-123. Las proyecciones obtenidas se reconstruyen en mapas tridimensionales mediante retroproyección filtrada o métodos iterativos.

La PET alcanza generalmente mayor sensibilidad y mejor resolución espacial que la SPECT, la cual está limitada por las estadísticas de fotones y la eficiencia del colimador. Ambas modalidades presentan un ruido considerable y baja resolución comparadas con CT o MRI, requiriendo algoritmos avanzados de corrección y eliminación de ruido. Las técnicas de aprendizaje profundo se aplican cada vez más para la mejora y reconstrucción de imágenes.

Debido al bajo conteo de fotones, las imágenes PET y SPECT suelen ser ruidosas, siendo SPECT la más afectada. La corrección por atenuación y dispersión es más compleja en SPECT, ya que depende de la emisión de fotones individuales y de la atenuación variable según los tejidos. Las imágenes PET y SPECT suelen fusionarse con exploraciones anatómicas (CT o MRI), lo que exige algoritmos de registro precisos para evitar desalineaciones que afecten la interpretación clínica. La segmentación y cuantificación presentan desafíos adicionales, como la definición precisa de regiones y la compensación de factores físicos y fisiológicos.

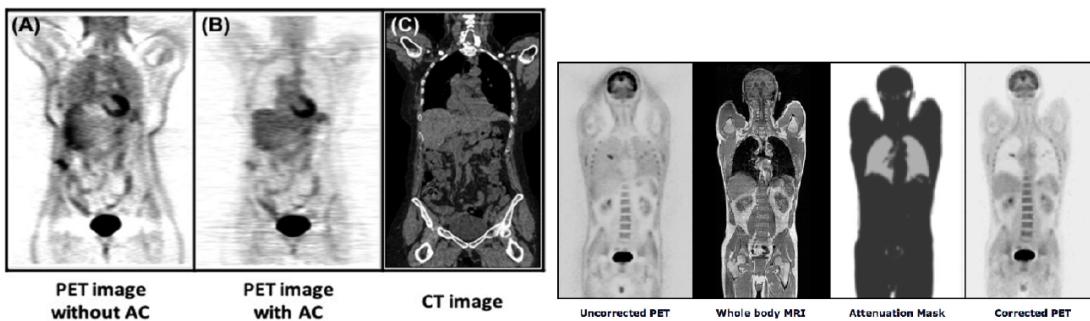


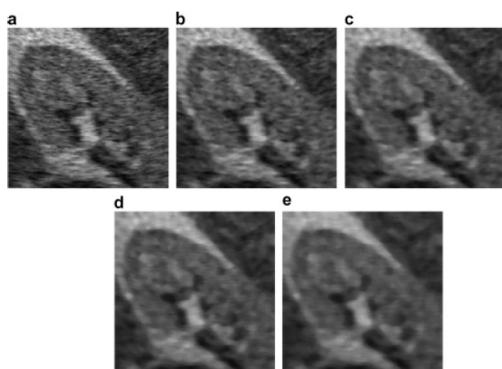
Figura VI.2: Corrección de atenuación (AC por sus siglas en inglés): A medida que los fotones emitidos por el trazador radiactivo viajan a través del cuerpo, algunos fotones son absorbidos o dispersados por tejidos como huesos, músculos o grasa. Esta pérdida de fotones significa que los detectores reciben menos fotones de los tejidos más profundos o densos, lo que hace que estas regiones parezcan artificialmente menos activas en las imágenes PET o SPECT.

En la práctica clínica, la PET se usa en oncología (detección y estadiaje tumoral), neurología (metabolismo cerebral) y cardiología (perfusión miocárdica). La SPECT se aplica en estudios de perfusión cardíaca, imagen ósea y cerebral, mapeo de receptores y diagnóstico de infecciones o inflamación. Aunque la resolución espacial es menor y tenga más artefactos y desventajas, su bajo costo y amplia disponibilidad de trazadores la hacen muy utilizada.

VI.4. Ultrasonido

La ecografía utiliza ondas sonoras de alta frecuencia emitidas por un transductor, un dispositivo con un material que vibra con la electricidad para generar las ondas. Estas ondas viajan a través de los tejidos y se reflejan en las interfaces, siendo luego captadas y reconstruidas en imágenes en tiempo real. El proceso es altamente dependiente del operador y de las propiedades del tejido.

Uno de los principales desafíos del procesamiento de imágenes ecográficas es el ruido de moteado (*speckle*), que surge por interferencia coherente de las ondas y genera un aspecto granular que reduce el contraste y dificulta la visualización de estructuras pequeñas. Se emplean algoritmos de eliminación de ruido especializados que preservan bordes y detalles finos.



La calidad de las imágenes varía considerablemente según el dispositivo, la habilidad del operador y la constitución del paciente. Los equipos portátiles suelen producir imágenes de menor calidad. Además, los artefactos como sombras, reverberaciones y efectos de atenuación dificultan la interpretación, por lo que deben detectarse y corregirse durante el procesamiento. Dado que el ultrasonido es una técnica dinámica y en tiempo real, los algoritmos de procesamiento deben ser rápidos y eficientes, equilibrando calidad de mejora y costo computacional.

VI.5. Microscopía

La microscopía óptica captura las interacciones de la luz a escala micro o nanométrica, transformadas por sensores digitales en datos de píxeles. Las tareas de procesamiento incluyen corrección de enfoque automático, eliminación de ruido (especialmente en imágenes de fluorescencia con poca luz), y mejora de contraste o iluminación desigual. Los desafíos incluyen la calibración precisa para mantener la resolución espacial, la segmentación de células superpuestas y el manejo de grandes volúmenes de datos generados por la microscopía de alto rendimiento. En este contexto, se aplican métodos de aprendizaje profundo para la segmentación y eliminación de ruido, así como enfoques computacionales avanzados como la reconstrucción holográfica.

Las aplicaciones abarcan el conteo celular, el análisis morfológico, la patología, el cribado de fármacos y la investigación en biología molecular. Dada la variabilidad entre

imágenes microscópicas, los pasos de preprocesamiento deben priorizar la corrección de iluminación, la normalización del color y la reducción de ruido antes del análisis cuantitativo.

VI.6. Integración y flujo de trabajo clínico

Las modalidades de imagen suelen utilizarse conjuntamente, como en PET/CT o PET/MRI. Los flujos de trabajo automatizados incluyen reducción de ruido, registro y segmentación, apoyándose en estaciones centralizadas de procesamiento de imágenes médicas para la visualización multimodal. El éxito del procesamiento depende de comprender tanto la física de formación de cada modalidad como su contexto clínico.

VI.7. Conclusión

- **CT:** Modalidad basada en rayos X que genera mapas tridimensionales de densidad. Los principales retos incluyen el ruido, artefactos como endurecimiento del haz o rayas metálicas, y la segmentación de tejidos blandos en presencia de estructuras de alto contraste.
- **PET y SPECT:** Técnicas de medicina nuclear que miden la distribución funcional de trazadores. Presentan alto ruido por conteo limitado de fotones, correcciones complejas por atenuación y dispersión, necesidad de registro con CT o MRI, baja resolución y dificultades de análisis cuantitativo.
- **Microscopía:** Imagen óptica de tejidos y células a micro/nanoescala, con retos de alta resolución, iluminación desigual, ruido en baja luz, segmentación compleja y grandes volúmenes de datos.
- **Ultrasonido:** Imagen en tiempo real basada en ondas sonoras, afectada por ruido de moteado, variabilidad dependiente del operador y el equipo, artefactos de sombra y reverberación, y restricciones de procesamiento en tiempo real.

Capítulo VII

Reconocimiento de patrones en imagen biomédica

VII.1. Clasificación de imágenes

VII.1.1. ¿Qué es la clasificación?

La clasificación busca asignar una etiqueta entre varias categorías o clases a una imagen o grupos de éstas. Se pueden asignar múltiples etiquetas a una imagen (multiestancia).

El algoritmo pasa primero por una fase de entrenamiento y una fase posterior de test e inferencia. El dataset es un conjunto de imágenes de entrenamiento (suele ser supervisado), que implica que el dataset viene etiquetado. Se extraen las características y se entrena con los datos de ejemplo y las etiquetas. De esta forma se obtiene un clasificador entrenado.

En el test, se proporciona una imagen no existente en el entrenamiento. Se extraen las mismas características que se extrajeron en el entrenamiento aplicando el clasificador obtenido durante la fase de entrenamiento. Esto permite predecir la clase que hay en esa imagen.

Dataset La base de datos se suele dividir en tres conjuntos para poder realizar validación cruzada. Las imágenes de entrenamiento sirven sólo para entrenar el clasificador. El subset de validación no sirve para entrenar, si no para medir el error y ajustar los parámetros. Las imágenes de validación nunca se deben usar para entrenar. Por último están las imágenes test, que dan una medida del error con datos nuevos que nunca se hayan visto.

Extracción de características Las imágenes se deben modelar para extraer las características. Se pueden modelar los píxeles, histogramas de los niveles de gris, patrones (templates) o descriptores de regiones o puntos de interés entre otros.

Entrenamiento Se genera una función que obtiene una predicción al ser aplicada sobre características de la imagen.

VII.1.2. Evaluación del rendimiento

La precisión de algoritmos de Machine Learning (ML) debe ser evaluada para seleccionar el mejor en cada tarea. Una tarea puede ser, por ejemplo, segmentación de tumores cerebrales en imágenes de resonancia magnética (MRI) e identificar qué píxeles son tumor.

Hay tres elementos clave para evaluar la efectividad de algoritmo (de clasificación) de manera sistemática:

- **Ground-truth/dataset:** valor o categoría real de cada dato para la tarea específica.
- **Resultado:** predicción del algoritmo. Aunque se entrenen en otro dataset, deben ser evaluados en común. Esto significa que si entrena un dataset de ayuda a la conducción con datos de Alemania, debe valer también para Suecia.
- **Métrica:** función que calcula la similitud entre los valores o categorías del resultado y del ground-truth.

Hay diferentes modelos de entrenamiento en función de los distintos niveles de anotación. Un entrenamiento sin supervisión no está asociado a la etiqueta. Una supervisión débil define la clase o categoría, y una supervisión completa define los distintos segmentos.

El resultado es la predicción del algoritmo. Tanto la predicción como el ground-truth son binarios, por lo que hay dos tipos de errores en clasificación: false negative y false positive.

- Métricas focalizadas:

$$\text{precisión} = \frac{TP}{TP + FP} = \frac{\text{correctos}}{\text{devueltos}}$$

$$\text{sensitivity, recall, hit rate o true positive rate (TPR)} = \frac{TP}{TP + FN} = \frac{\text{positivos}}{\text{total anotaciones positivas}}$$

- Métricas globales:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{\text{predicciones correctas}}{\text{total predicciones}}$$

$$\text{Error rate} = \frac{FP + FN}{TP + TN + FP + FN} = \frac{\text{predicciones incorrectas}}{\text{total predicciones}}$$

$$\text{F-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Mediante una **matriz de confusión** se puede visualizar los aciertos y errores para clasificaciones multiclas.

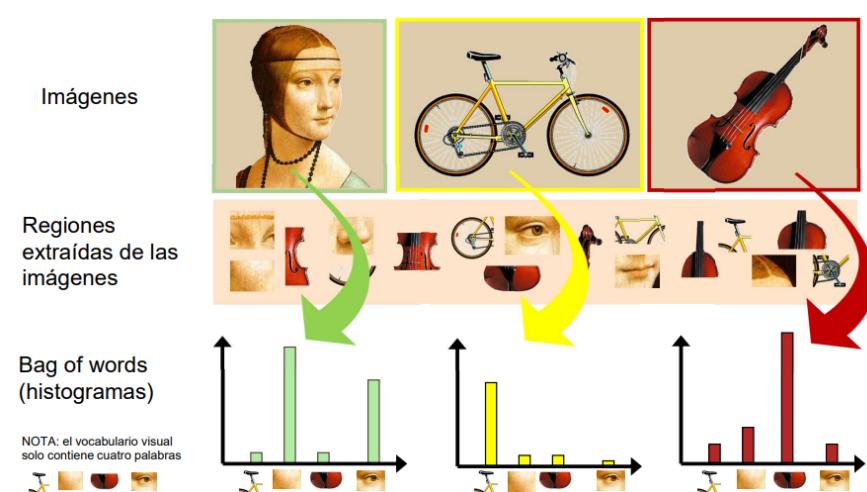
VII.1.3. Clasificación basada en *bag of words*

Se aprenden diferentes características visuales que representen un objeto. El objeto se modela como mezcla de elementos básicos o textones. Dependiendo de su textura se van a agrupar de una forma u otra. La bag of visual words tiende a acumular los patrones más repetidos en un diccionario de textones y caracterizar en un histograma las veces que aparecen los textones en la imagen de entrada. No obstante, no se incluye información de localización espacial ni existe distinción entre frente y fondo de la escena, trata todas las texturas por igual.

Son modelos para describir imágenes (contiene dos ojos, una nariz, una boca), pero no dónde se encontraba originalmente en la imagen (izquierda, derecha, arriba, centro, ...).

Las etapas del modelo son las siguientes:

1. Extracción de características de las imágenes
2. Aprendizaje del vocabulario visual
3. Cuantificación de características (palabras visuales) a partir del vocabulario visual; se puede limitar a un cierto número, como las 4 palabras visuales más clave, más frecuentes, más distribuidas, etc.
4. Representar imágenes como histogramas con palabras visuales



Este proceso se repite para cada imagen nueva que se encuentre. Todas las imágenes de entrada se matchean con el diccionario visual creado y se busca a la que más se parece. Una vez que se tienen los histogramas, se suele aprender un clasificador para distinguirlas.

Ventajas del modelo Bag-of-words	Desventajas del modelo Bag-of-words
Flexible a la geometría, deformaciones y puntos de vista	El fondo y el primer plano se mezclan al representar toda la imagen
Resumen compacto del contenido de la imagen	La formación óptima del vocabulario
Proporciona representación vectorial dimensional fija para conjuntos	El modelo básico ignora la geometría; debe verificarse después o codificarse mediante funciones
Muy buenos resultados en la práctica	

VII.2. Detección de objetos

VII.2.1. ¿Qué es la detección de objetos?

La detección de objetos permite clasificar y localizar objetos espacialmente en la imagen que discriminan la información del objeto de la información del fondo. La detección de objetos no es la identificación de objetos concretos, si no la clasificación de distintas clases de objeto genéricas.

Presenta muchos desafíos. En el caso de las personas, cada objeto tiene mucha variabilidad: la postura, altura, edad, escala, relación de aspecto, punto de vista más lejano, multitudes, iluminación, ropa. Esto dificulta identificarlos entre ellos.

Entre los criterios de diseño está la búsqueda eficiente de los objetos más probables. Incluso los modelos simples requieren buscar cientos de miles de posiciones y escalas. También se deben diseñar características generalistas y robustas para generar puntuaciones útiles. Para tratar diferentes puntos de vista, a menudo se entrena diferentes modelos para los puntos de vista existentes en la base de datos.

Las etapas generales para la detección de objetos son:

1. Definir modelo objeto: patrón estadístico. El objeto es un rectángulo en la imagen y se buscan características definidas en el rectángulo (color, textura, forma, iluminación). Se pueden utilizar detección de bordes como Prewitt.
2. Generar hipótesis, proponer un alineamiento del objeto a la imagen: la más utilizada es la ventana deslizante. Se selecciona una ventana (cuadradito) en cada ubicación y escala y se realiza un desplazamiento definido por el parámetro paso (stride). Entre una ventana y otra se puede solapar un 99 %. Tradicionalmente, la ventana se mantiene de tamaño y cambia la resolución de la imagen a un tamaño menor. Así, el objeto que se encuentra debería ser más grande que en la imagen original. Esto permite lidiar con multi-escala o ventanas con diferentes tamaños. Cada ventana obtenida se clasifica independientemente para ver si contiene el objeto a detectar.

En la nueva imagen, se obtienen las características asociando los nuevos elementos al diccionario visual. Teniendo una cabeza, se sabe que está en la parte superior del objeto. Así se realiza un votado probabilístico para obtener el centro. Se genera un espacio tridimensional de votación, que es un espacio continuo. La profundidad, el eje z, de ese espacio es la escala de la imagen, si se encontró con el tamaño original o en una menor resolución.

3. Puntuar hipótesis basado en correlación con el modelo. Se realiza una suma de puntuaciones de características en posiciones fijas y se establece un threshold a partir del cual establecer si es un objeto de interés.
4. Refinar detecciones, re-puntuar las detecciones usando todos los resultados disponibles para eliminar las detecciones solapadas. Una técnica suprime los valores no máximos (*non-max suppression*; NMS). Teniendo varios candidatos para un mismo objeto, sólo se va a mantener un único candidato por objeto de interés. Se mezclan y combinan los candidatos comparándolos dos a dos para quedarse con el de mejor confianza siempre y cuando el solape se encuentre por encima de un umbral. Como medida de solape se utiliza la intersección sobre unión (IOU):

$$IOU = \frac{A \cap B}{A \cup B}$$

Otras características adicionales permiten refinar las detecciones, como el contexto o conocimiento a priori. Por ejemplo, en la ayuda a la conducción, se establece una línea horizontal a una determinada altura a partir de la cual no debe intentar clasificar el modelo, ya que sería una altura inviable para la aplicación.

VII.2.2. Evaluación del rendimiento

Al igual que en el caso anterior, se evalúa el rendimiento de las predicciones del algoritmo con las anotaciones o ground-truth. Entre las métricas se encuentran:

- Positivos correctos: para cada anotación se busca si existe una predicción cuyo IoU excede un umbral (típicamente $\tau = 0.5$)
- Falsos positivos: detecciones para las que no existe una anotación cuyo $IOU > \tau$
- Falsos negativos: anotaciones para las que no existe una detección cuyo $IOU > \tau$

Debido a la dependencia con el umbral τ en *precision* y *recall*, se calculan curvas para cada valor de τ . Cada punto de la curva corresponde con un valor de τ concreto. Finalmente, como medida de evaluación se utiliza el área bajo la curva (*Average Precision*).

Los datasets deben contener ejemplos de las categorías a detectar y de los posibles casos donde "no detectar". Por ejemplo, para la clasificación de personas, la base de datos de entrenamiento debe tener miles de imágenes con el objeto de interés y millones de imágenes sin el objeto de interés.

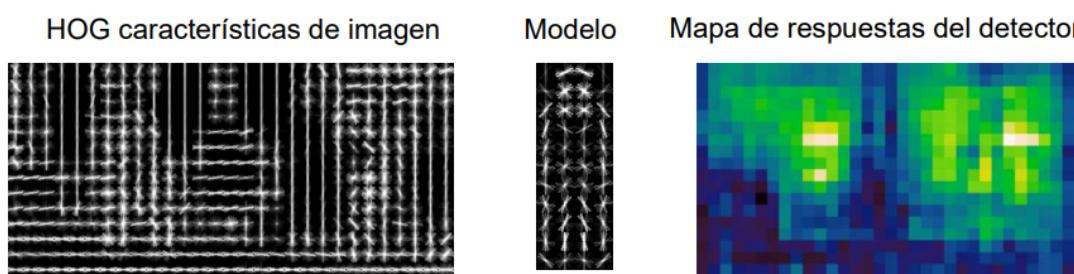
VII.2.3. Detector Dalal-Triggs

Este algoritmo fue el primer detector de personas con un buen rendimiento. Funciona de la siguiente forma:

1. Extraer una ventana de tamaño fijo 64x128 píxeles en cada posición y escala.

2. Normalización gamma y color: se suele probar con diferentes espacios de color (RGB, LAB, escala de grises) y un ajuste por raíz cuadrada o logarítmico.
3. Cálculo de gradientes en x, y y el módulo del gradiente para obtener los bordes.
4. Calcular descriptor HOG (histograma de gradientes orientados) dentro de cada ventana extraída: cálculo de orientación del gradiente en $[0,180]$ y división de la imagen en 8×16 celdas de 8×8 píxeles. Con esto se crean los histogramas de gradientes orientados
5. Normalización de bloques solapados: normaliza celdas 2×2 para obtener macrobloques 16×16 y concatenar histogramas macrobloque en vector v para normalizar dicho vector.
6. Obtención descriptor: se concatenan todos los descriptores normalizados.
7. Clasificar la ventana con un clasificador lineal SVM binaria: clase positiva (objeto) y clase negativa (no-objeto)
8. Realizar supresión no-máximos para eliminar detecciones redundantes con puntuaciones más bajas

Se asume el aprendizaje de un patrón o modelo con SVMs. Con una imagen nueva, se extraen las mismas características y el HOG. A continuación se compara cada ventana de la imagen nueva con el modelo usando ventanas deslizantes, como una convolución. Esto genera un mapa de respuestas del detector que muestra dónde es más probable que se encuentre el objeto en el dominio. Sobre ese mapa se realiza una búsqueda de máximos locales para quedarse sólo con los mejores candidatos en cada región. Esto se hace multiescala y se suprimen las ventanas solapadas con baja puntuación.



Ventajas del detector Dalal-Triggs	Desventajas del detector Dalal-Triggs
Funciona aceptablemente para objetos no deformables con orientaciones canónicas: caras, coches, personas Detección rápida	Requiere grandes cantidades de datos para entrenamiento No es robusto a occlusiones Rendimiento bajo para objetos altamente deformables