Presentación

La visión artificial es una de las áreas de mayor importancia en la ingeniería de sistemas. Los algoritmos de visión artificial se utilizan para todo tipo de problemas, como pueden ser la inspección industrial de productos en busca de defectos, la interacción de un robot con una persona, la conducción autónoma, la vigilancia mediante cámaras o la autenticación de un teléfono móvil.

Hasta hace poco más de una década, la mayoría de los algoritmos que se utilizaban en visión artificial lo hacían aplicando determinadas operaciones y transformaciones fijas sobre las imágenes. Muchos parámetros de los modelos utilizados se obtenían en base a un prueba y error, buscando aquellos parámetros o transformaciones que mejor permitían resolver un problema. Los algoritmos y técnicas utilizados en esta época de la visión artificial clásica son los precursores de las actuales técnicas de aprendizaje automático.

El aprendizaje automático o *machine learning* representa al conjunto de técnicas y algoritmos que permiten resolver problemas aprendiendo por sí mismos a ajustar sus propios parámetros. Es un campo de gran importancia en muchas ramas del conocimiento, desde la propia ingeniería de sistemas o la inteligencia artificial hasta la resolución de problemas de diversa índole en ingeniería, matemáticas, economía o incluso medicina. Dentro de la visión artificial, los algoritmos de aprendizaje automático permiten obtener modelos muy potentes que resuelven multitud de problemas que iremos viendo a lo largo del libro, como son la clasificación, segmentación, detección, comparación, generación sintética de imágenes. Estos algoritmos permiten dotar a nuestros sistemas de cierta inteligencia artificial, ya que en ocasiones algunos de estos modelos pueden llegar a realizar tareas, normalmente limitadas a personas, con el mismo desempeño y a mucha mayor velocidad.

Este libro presenta dos partes de la visión complementarias. En una primera parte del libro se presentan varios aspectos de la visión artificial clásica, como son los fundamentos de las imágenes que serán bases para todo el aprendizaje automático, los componentes de los sistemas de visión artificial y los sensores que se utilizan, así como algunas de las aplicaciones prácticas. En una segunda parte se explicarán los frameworks de trabajo que se utilizarán en el libro, principalmente Tensorflow y Pytorch. A continuación, se explicará cómo funcionan los algoritmos de aprendizaje no supervisado y supervisado. Poco a poco iremos mostrando cómo entrenar redes neuronales para ser capaces de entrenar modelos de clasificación simples, utilizando redes de convolución. Presentaremos

redes de clasificación más avanzadas, como ResNeXt o ConvNeXt. Mostraremos cómo procesar vídeo mediante redes LSTM, cómo segmentar objetos mediante redes como Mask R-CNN, cómo detectar objetos con redes como YOLO, o cómo obtener el esqueleto de personas mediante OpenPose o BlazePose. Abordaremos la comparación de imágenes mediante redes siamesas o la generación de imágenes nuevas mediante redes generativas. Finalmente, mostraremos cómo mejorar los resultados mediante la combinación de modelos de la misma o distinta naturaleza, o cómo utilizar métodos combinados de redes siamesas para resolver el problema de aprendizaje con una única imagen por categoría utilizando CP-CVV.

Este libro está dirigido por una parte a alumnos de asignaturas de visión artificial, aprendizaje automático o inteligencia artificial, de grado, máster. Estas asignaturas se imparten principalmente en ingenierías de la rama industrial, informática y de telecomunicaciones, en sus niveles de grado y máster. Por otra parte, el libro se dirige también a los profesionales e investigadores que trabajan en visión artificial, así como a los interesados en algoritmos de aprendizaje automático e inteligencia artificial. Como requisito previo, el lector debería conocer a un nivel medio el lenguaje de programación Python.

En el libro se explicará cada uno de los códigos mostrados, enlazando el propio libro con un repositorio con el fin de que el lector pueda probar los programas. Se explicarán técnicas de visualización de gráficos de entrenamiento, métricas de evaluación de modelos, tipos de capas, funciones de activación y error de las redes neuronales, y técnicas de explicación de cómo funcionan estos modelos, con el objetivo de evitar que estas redes sean una caja negra con resultados impredecibles.

Esperamos que el libro resulte interesante y enriquecedor para el lector, animándole a experimentar con sus propios modelos a partir de los presentados en el libro.

Sobre los autores

Jaime Duque Domingo es Doctor en Ingeniería de Sistemas y Control por la UNED (2018), tiene un Máster en Ingeniería de Sistemas también por la UNED (2014), un Máster de Profesorado por la Universidad Isabel I (2018) y es Ingeniero en Informática por la Universidad de Valladolid (2011). Durante 18 años trabajó en el desarrollo de complejos proyectos informáticos para el ámbito privado, tanto en España como en el extranjero. En los últimos años se ha centrado en el mundo académico, participando en varios proyectos de investigación, realizando distintas publicaciones e impartiendo docencia universitaria. Ha obtenido seis premios de investigación, incluyendo dos premios INFAIMON otorgados al mejor trabajo de visión artificial presentado en las Jornadas Nacionales de Automática (2015 y 2018) o el Premio Extraordinario de Doctorado de la UNED. Actualmente es profesor en el Departamento de Ingeniería de Sistemas y Automática de la Universidad de Valladolid. Ha publicado 16 artículos en revistas indexadas en SCI-JCR, en el primer y segundo cuartil, 14 artículos en congresos nacionales e internacionales, otro libro sobre visión artificial y un capítulo sobre robótica en el autismo. Su campo de actuación se centra en la visión artificial y robótica, especializándose en técnicas de aprendizaje profundo, robótica social y cognitiva, o sistemas de posicionamiento. Como investigador, ha trabajado en la Universidad de Valladolid y en el Centro Tecnológico CARTIF. Ha sido también profesor en la Universidad Europea Miguel de Cervantes (UEMC), así como profesor visitante en la Carnegie Mellon University (CMU), en Pittsburgh (Estados Unidos).

Jaime Gómez García-Bermejo es Catedrático de Universidad del Departamento de Ingeniería de Sistemas y Automática de la Universidad de Valladolid. Es Doctor en Ingeniería Industrial por la Universidad de Valladolid y tiene un Máster en Procesamiento de Imágenes por l'École Nationale Supérieure de Télécommunications de Paris (Francia). Ha participado en cerca de un centenar de proyectos de investigación competitivos, internacionales, nacionales y regionales, muchos de ellos en cooperación con empresas o entidades públicas y enmarcados en convocatorias públicas tanto de los Programas marco de la Unión Europea como del Plan Nacional (incluyendo Retos, Innpacto, Cien, Avanza, CEDETI, Profit etc). En el aspecto de transferencia de conocimiento al sector productivo, ha participación en unos 125 contratos de investigación con empresas y entidades públicas. También es coautor de diversas patentes licenciadas para su uso por importantes empresas de ingeniería, sector de la construcción y empresas de servicio, así como de diversas licencias software. En el apartado de publicaciones es coautor unas 75 contribuciones científicas de relevancia, la mayoría correspondientes a artículos en revistas indexadas JCR-SCI. También es coautor de más de 125 contribuciones a congresos, muchos de alto nivel internacional, y ha impartido diversas conferencias invitadas. Ha dirigido 6 tesis doctorales en los últimos 10 años, así como más de un centenar de trabajos fin de grado/máster. Es evaluador de proyectos para numerosas agencias (ANEP, AAC, etc.) y ha trabajado como Experto en I+D en numerosas ocasiones para firmas como AENOR y ACIE, para la evaluación y la acreditación de actividades de investigación. Es miembro académico del Instituto de las Tecnologías Avanzadas de la Producción – I.T.A.P. y asesor científico del Área de Visión Artificial en el Centro Tecnológico CARTIF.

Eduardo Zalama es Doctor Ingeniero Industrial por la Universidad de Valladolid desde 1994. Actualmente es Catedrático en la Escuela de Ingenierías Industriales de la Universidad de Valladolid. Ha sido Profesor Visitante en la Universidades de Boston y Carnegie Mellon (Pittsburgh). Su línea de investigación se centra en el ámbito de la robótica y visión artificial con especial énfasis en la transferencia pues en los últimos años ha realizado su investigación en el Centro Tecnológico Cartif donde ha dirigido la División de Sistemas Industriales y Digitales. Es autor de más de un centenar de artículos peer-review en revistas y libros de prestigio internacional y más de cien comunicaciones en congresos nacionales e internacionales en el campo de la robótica y visión artificial. Desde el punto de vista industrial y de investigación ha participado en más de 90 proyectos de investigación competitivos de ámbito internacional, nacional y regional destacando proyectos del Programa Marco y Plan Nacional (incluyendo Cenit, Innpacto, Avanza y Profit). En el aspecto de transferencia de conocimiento al sector productivo, ha participación en más de un centenar de contratos de investigación con empresas, en la mayoría de los cuales ha actuado como investigador principal. También es coautor de varias patentes en explotación que han sido licenciadas para su uso por importantes empresas de ingeniería, sector de la construcción y empresas de servicio. También se han concedido licencias de uso de software a empresas como Telefónica I+D o Renault-Valladolid. Es evaluador de la Comisión Europea (diversas convocatorias), ANEP, ANECA (programa Academia), agencias de evaluación autonómicas, y ha sido contratado como Experto en I+D en diversas ocasiones por la firma AENOR y ACIE, para la evaluación de actividades de investigación. Finalmente se destaca la pertenencia a diferentes comités científicos nacionales e internacionales, pertenencia a la red CEA-GTROB, HISPAROB y Eurobotics aisbl.

Parte I

Introducción a la visión artificial y componentes de los sistemas de visión

Capítulo 1

Introducción y conceptos básicos

1.1. Introducción general

La visión artificial es una disciplina que se ocupa de la captación, el análisis y la comprensión de imágenes y vídeos. Se relaciona estrechamente con otros campos científicos y tecnológicos tales como la inteligencia artificial, el aprendizaje máquina, el procesamiento por computador, el procesamiento de señal, gráficos por computador, robótica e incluso con psicología y neurociencias.

Con frecuencia el término visión artificial se utiliza para referirse a cuatro disciplinas íntimamente relacionadas: visión por computador, visión máquina, tratamiento de imágenes y visión artificial propiamente dicha. La visión por computador (*computer vision* en la literatura anglosajona) se ocupa de la utilización de los computadores para alcanzar un conocimiento de las imágenes adecuado para desarrollar tareas similares a las que realizamos los humanos por medio de nuestro sistema visual. La visión máquina (*machine vision*) se orienta específicamente a la automatización de tareas típicamente industriales por medio de imágenes, como la inspección, el control de procesos y el guiado de robots. El tratamiento de imágenes (*image processing*) se refiere genéricamente al procesamiento de las mismas por medio de computadores de cara su mejora y análisis. Por último, la visión artificial propiamente dicha (*artificial vision*) se relaciona con reproducir el comportamiento del sistema visual humano. Por supuesto estas cuatro disciplinas comparten algoritmos y tecnologías, por lo que es habitual referirse al conjunto de ellas genéricamente como visión por computador o, como lo haremos en este libro, visión artificial.

La visión artificial, entendida en este sentido amplio, comienza su andadura a mediados de los años 60, cuando por primera vez se aborda la conexión de una cámara a un computador (1966, M. Minsky). Surgen también los primeros algoritmos para la detección y seguimiento de bordes. En la década de los 70 asistimos al desarrollo de numerosos algoritmos de visión artificial que

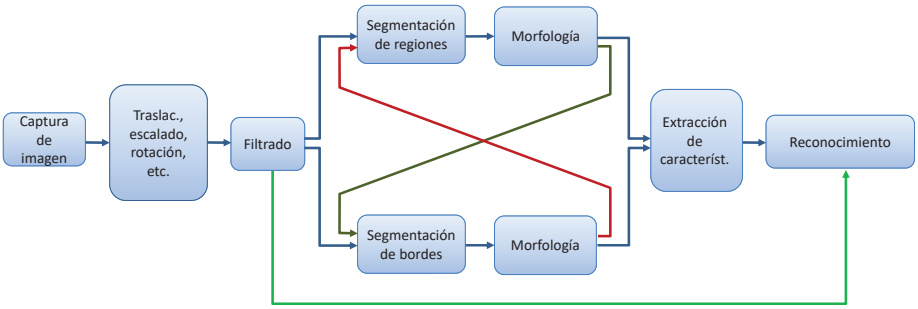


Figura 1.1: Arquitectura basada en convoluciones 3D

continúan vigentes hoy en día: mejora de la extracción de regiones y bordes, reconocimiento de formas geométricas, flujo óptico, estimación del movimiento. Los 80 vienen marcados por una creciente formalización matemática, lo que convierte a la visión artificial en una disciplina científica madura. Con esto, y también gracias al desarrollo de los computadores y la microelectrónica, queda plenamente establecido el *pipeline* clásico de visión artificial. Existen distintas formulaciones de este *pipeline* que básicamente cubren, de una forma u otra, las siguientes etapas (*véase* la figura 1.1):

- Captura de las imágenes. Adquisición de la información visual de la escena por medio de cámaras que proporcionan imágenes individuales o secuencias de imágenes (vídeo).
- Filtrado. Se trata de una etapa fundamental en la que se cubren múltiples aspectos, desde la reducción del ruido de las imágenes hasta el resaltado de características de interés como líneas, bordes de regiones, regiones de distinta morfología o características de más alto nivel. Esta etapa puede ser determinante de cara al éxito de los tratamientos posteriores.
- Segmentación. Consiste en el particionado de la imagen en conjuntos de píxeles o *segmentos* (segmentos de información; no confundir con segmentos lineales) que comparten ciertas propiedades. La etapa previa filtrado determina la mejor forma de operar, sobre la base del análisis de propiedades tales como los niveles de gris, color o textura, y sus variaciones a lo largo de la imagen, en muchos casos combinado con criterios de proximidad en la imagen. Además, la segmentación puede ir orientada a la obtención de regiones o de bordes, y estos últimos pueden ser conectados de modo de que determinen contornos de regiones. Regiones y contornos son, en definitiva, dos formas equivalentes de representar el resultado de la segmentación, y es posible pasar fácilmente de unas a otras y a la inversa.
- Morfología. Operaciones basadas en teoría de conjuntos que se aplican a regiones y bordes o contornos para mejorar el resultado de la segmen-

tación. Son habituales las operaciones de erosión, dilatación, apertura y cierre, entre otras. Muchas de estas operaciones conllevan la interacción con conjuntos de puntos denominados *elemento estructurante* que adaptados a distintas necesidades.

- Extracción de características. Cálculo de descriptores geométricos y topológicos de las regiones encontradas tras la segmentación, tales como perímetro, área, compacidad, momentos de inercia de distinto orden, número de agujeros y de regiones conexas, etc.

- Reconocimiento de los objetos presentes en la escena. En esta etapa se emplean desde técnicas sencillas de umbralización directa de descriptores hasta otras más elaboradas, del ámbito de la Inteligencia Artificial.

- Interpretación de la escena. Se analizan objetos y sus interrelaciones para dar sentido a la escena en el contexto de la aplicación deseada. Es la culminación del proceso de visión y de nuevo las técnicas de Inteligencia Artificial pueden jugar un papel fundamental en esta etapa final.

Una aproximación alternativa al problema, utilizada desde antiguo en visión máquina y que obvia algunas de estas etapas, es lo que se conoce como comparación con plantillas o *Template Matching*. Consiste en recorrer la imagen en busca de regiones parecidas a una pequeña plantilla que contiene una imagen del elemento buscado. Además, si se adopta la correlación normalizada como medida de similitud, se consigue cierta insensibilidad ante variaciones de la iluminación. Esta técnica se emplea con frecuencia en visión máquina, pero su robustez es limitada cuando nos enfrentamos a variabilidad en el aspecto visual de los objetos.

Por otra parte, a finales de los 90 se desarrollan los conocidos algoritmos *Scale-Invariant Feature Transform* (SIFT) y Speeded-Up Robust Features (SURF), y sus distintas variantes. Estos algoritmos permiten detectar y caracterizar puntos relevantes de los objetos de forma robusta frente a cambios de iluminación, rotación y escala. Por ello, resultan de interés en tareas como reconocimiento de objetos, incluso en presencia de oclusiones, *stitching* o pegado de imágenes, *matching* o emparejamiento en visión estéreo, *tracking* o seguimiento visual, etc. En 2001 se produce otro avance significativo de la mano del algoritmo de Viola-Jones. Se trata de un algoritmo capaz de detectar rostros u otros objetos en tiempo real, con un coste computacional asumible incluso para dispositivos con prestaciones modestas.

Más recientemente, en los 2000, asistimos al resurgimiento de las redes neuronales convolucionales. La idea original data de los años 80, pero ha sido necesario esperar casi 30 años para verla implementada de manera efectiva gracias a los modernos procesadores gráficos (GPUs). La red AlexNet, de finales de 2012, representa un hito destacable por ser una de las primeras en sacar provecho de este tipo de procesadores. A raíz de AlexNet hemos asistido a un desarrollo vertiginoso del campo de las redes convolucionales, donde nuevas arquitecturas, como los modelos ViT o las redes más avanzadas de convolución (CoAtNet,

ConvNeXt, etc.) han visto la luz. A lo largo del libro ahondaremos en alguno de estos modelos.

El presente libro se inscribe en este contexto general. No pretende profundizar en el *pipeline* clásico o en las técnicas de los años 90, para cuya descripción ya existen excelentes textos. En lugar de ello, tras este capítulo introductorio revisaremos primero los componentes de los sistemas de visión, abarcando desde las cámaras, las ópticas y los sistemas de iluminación, hasta los modernos sistemas de visión 3D. Luego describiremos las aplicaciones de la visión artificial en la Industrial 4.0. Por último nos adentrarnos específicamente en cómo utilizar las redes neuronales para resolver distintos problemas de visión artificial. En todo ello se ha adoptado un enfoque esencialmente práctico con el fin de que sirva de guía de referencia a la que podamos acudir para resolver los distintos problemas de visión artificial que podamos encontrar a lo largo de nuestra vida profesional.

1.2. Imagen y vídeo

En primera aproximación, una imagen digital es una composición de elementos, denominados píxeles, habitualmente organizados en forma de matriz o línea, cuyos valores numéricos representan la luminosidad o el color de los puntos de una imagen. En los sistemas de visión, la imagen es capturada por medio de una cámara con su correspondiente óptica, bajo condiciones de iluminación adecuadas. La cámara incluye un dispositivo sensor, normalmente plano y de naturaleza discreta, con sus elementos fotosensibles distribuidos de manera regular. Este sensor muestrea espacialmente la información óptica y la convierte en señales eléctricas, que son amplificadas y cuantificadas en forma de niveles numéricos para su transferencia hacia el computador. Por su parte, la óptica es la encargada de enfocar la energía luminosa de la escena sobre el sensor. Cámara y óptica son dos elementos bien diferenciados, si bien en ocasiones denominaremos *cámara* al conjunto, por simplicidad.

Los niveles que proporciona la cámara son una forma de representar la intensidad que debería adoptar una fuente de luz, o varias en el caso de color, para producir una sensación visual similar a la imagen óptica de partida. Estos niveles se representan usualmente mediante números enteros sin signo de 8 bits, dado que una diferencia de $1/2^8$ es apenas discernible para el ojo humano en el rango de luminosidades y colores que manejamos habitualmente. No obstante, por supuesto es posible una cuantización más precisa. Por ejemplo, resulta cada vez más frecuente recurrir a 10, 12, 14 o incluso 16 bits (por canal cromático, en su caso). Además, el procesamiento ulterior de las imágenes puede dar como resultado valores que se representan más adecuadamente mediante otros tipos de datos: enteros de otros tamaños, con o sin signo, valores lógicos o números reales. Por todo ello, en general no restringiremos el tipo de datos que puede contener una imagen.

En cuanto a la señal de vídeo, se puede entender como una secuencia de imágenes adquiridas en instantes de tiempo sucesivos. En primera aproxima-

ción, podemos suponer que cada imagen de la secuencia contiene toda la información correspondiente a un cierto instante de tiempo, si bien esto es matizable como veremos en el capítulo dedicado a los componentes de los sistemas de visión. Adicionalmente, asumiremos que el tiempo transcurrido entre dos imágenes consecutivas es siempre el mismo.

Por otra parte, para la transmisión y el almacenamiento de imágenes y vídeo se recurre con frecuencia a distintos contenedores y formatos. Estos vienen caracterizados por la organización de los valores de luminosidad o color que contienen, los metadatos que los acompañan y los algoritmos utilizados para la codificación de la información, con o sin pérdida. Algunos formatos de imagen populares son BMP, PNG , TIFF, EXIF y JPEG. Algunos formatos de vídeo comunes son AVI, MPEG, MOV y MP4. Una amplia discusión de estos y otros formatos puede encontrarse en textos especializados. En nuestro caso asumiremos que tanto las imágenes como las secuencias de vídeo han sido adecuadamente descodificadas hasta su representación en forma de matriz o secuencia temporal de matrices, previamente a abordar su procesamiento.

1.3. El color

Tal como hemos indicado, las cámaras proporcionan, por cada píxel, los niveles de intensidad que deberían adoptar las luces de un cierto dispositivo de visualización para producir en el observador humano una sensación visual análoga a la imagen capturada. Además, la mayoría de los sistemas actuales contempla luces de tres *colores primarios*, rojo, verde y azul, que suponen un compromiso adecuado entre complejidad de las cámaras y los dispositivos de visualización, y la gama de colores que permiten reproducir.

La elección de tres primarios descansa sobre el hecho de que el observador humano estándar, o promedio, utiliza tres tipos de células para percibir el color. Se trata de los denominados *conos*, sensibles a luz de longitudes de onda cortas, medias y largas, según su tipo. Un cuarto tipo de células, los *bastones*, perciben la cantidad total de luz. Las señales eléctricas que generan estos cuatro tipos de células son interpretadas por el cerebro como luz y color.

En los años 30 se realizaron una serie de experimentos encaminados a relacionar cuantitativamente la radiación electromagnética con la percepción visual. Fruto de ello se propuso el denominado espacio color CIE 1931 XYZ, que asigna a cada color 3 números positivos, de forma unívoca. Estos números pueden entenderse como las coordenadas de cada color en un espacio cromático tridimensional denominado XYZ. Por conveniencia, el espacio XYZ fue definido de forma que Y representa la luminosidad, es decir la cantidad total de luz, mientras que X y Z se relacionan con el color propiamente dicho.

Los *diagramas cromáticos* son una forma adecuada de manejar este espacio cromático (véase la Figura 1.2). Se trata de secciones planas del espacio XYZ en las que se representa la gama de colores que puede percibir el observador humano promedio, para un determinado nivel de luminosidad (un cierto valor de Y). Los *colores espectrales*, es decir correspondientes a las radiaciones monocromáticas

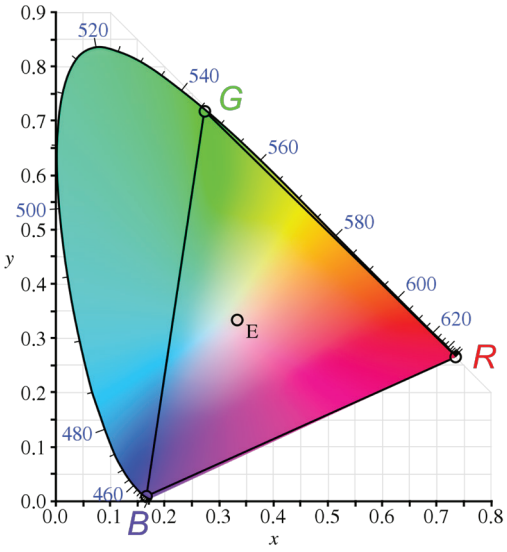


Figura 1.2: Diagrama cromático CIE 1931. Solo se visualizan correctamente los colores reproducibles a partir de las (pocas) tintas empleadas en la impresión de este documento. (Adaptado de Wikipedia, CIE 1931 color space).

(de una única longitud de onda), se sitúan en el contorno de la gama, mientras que la denominada *línea de púrpuras* delimita la región por la parte inferior. Por construcción, todos los colores de la gama pueden generarse combinando luces de los (infinitos) colores espectrales, balanceadas en función del inverso de su distancia al color deseado medida sobre el diagrama. La característica forma de *lengua* o *herradura* de estos diagramas deriva de la fuerte correlación existente entre las respuestas espectrales de los distintos tipos de conos.

En general, un conjunto arbitrario de luces permite generar todos los colores presentes en el interior del polígono que circunscriben dentro del diagrama cromático (pero no los presentes en el exterior). A modo de ejemplo, el espacio color estándar *CIE 1931 RGB* especifica tres luces monocromáticas de longitudes de onda 700 nm (rojo, marcado R en la Figura 1.2), 546.1 nm (verde, G) y 435.8 nm (azul, B) que, adecuadamente balanceadas, permiten producir todos los colores del interior del triángulo que delimitan. Obviamente no todos los colores pueden reproducirse por este medio. Por ejemplo, se pierden todos los verdeazulados situados a la izquierda de la línea BG y los púrpuras bajo la línea RB. Por lo general, la situación empeora en el caso de utilizar otra combinación de primarios monocromáticos, o bien de luces no monocromáticas como es el caso de muchos monitores o dispositivos de proyección. Desde luego, el uso de un mayor número de primarios, convenientemente seleccionados, permite ampliar la gama de colores reproducibles. Por ejemplo, algunos dispositivos de visualización trabajan con cuatro primarios (rojo, verde, azul y amarillo). Sin embargo, su uso suele resultar antieconómico, en especial si se tienen en cuenta

las implicaciones respecto al diseño de cámaras.

En todo caso, lógicamente los niveles de color proporcionados por la cámara deben adaptarse a los primarios que utilice cada dispositivo de visualización. Para ello debe se requiere una transformación que, para tres primarios, adopta la forma

$$\begin{pmatrix} R' \\ G' \\ B' \end{pmatrix} = \mathbf{M} \begin{pmatrix} R \\ G \\ B \end{pmatrix} . \tag{1.1}$$

La naturaleza lineal de la transformación deriva del hecho de que, en definitiva, equivale a un cambio de ejes en el espacio XYZ. Adicionalmente se pueden incorporar otras transformaciones, lineales o no, como la conocida *corrección gamma* que compensa las no linealidades de los dispositivos de visualización así como del propio sistema visual humano, tal como veremos en un capítulo posterior.

Existen formas alternativas de especificar la información cromática, no tanto orientadas a su reproducción mediante luces R, G y B, como a otros fines. Entre ellas destaca el sistema HSV (*hue* o matiz, *saturation* o saturación, *value* o valor), también conocido como HSB (*hue*, *saturation*, *brightness* o brillo). Este sistema se encuentra estrechamente relacionado con la forma en que los humanos percibimos y describimos el color. En este sistema, el matiz puede entenderse como una tonalidad específica dentro del espectro visible, descrita como su posición en un determinado *círculo cromático*. Se especifica habitualmente como un ángulo en el rango $[0, 360^{\text{o}})$. La saturación corresponde a la *cantidad* de ese color y se especifica típicamente en un rango de 0 %, para el gris, a 100 % para los colores puros. Finalmente, el brillo corresponde intuitivamente a la *cantidad de luz* que contiene el color, desde 0 % para negro hasta 100 % para blanco. Estos tres rangos se discretizan convenientemente, muchas veces en forma de enteros cortos o números reales, para su manejo eficiente por el computador. La transformación entre los espacios RGB y HSV es bien conocida, aunque no lineal.

Los espacios descritos, RGB y HSV, resultan adecuados para tareas de visión. No obstante, existen espacios de color mejor adaptados a otros ámbitos. Por ejemplo, el sistema HSL/HSI (*h*ue, *saturation*, *lightness* o *intensity*) es similar al HSV, pero se adapta a la reproducción de los colores mediante mezcla de pinturas en lugar de luces. Otros como el CIEL*a*b* y el propio CIEXYZ se usan en colorimetría por cuanto permiten expresar mediante números positivos todos los colores que puede percibir el observador humano, con independencia del dispositivo de visualización utilizado. Espacios como el CMY(K) (*cyan*, *magenta*, *yellow* (*black*)) se adecúan a la reproducción mediante tintas que reflejan estos tres colores primarios (y absorben, por tanto, todos los demás). Existen también sistemas más orientados a fotografía digital como el YCbCr. Una descripción más exhaustiva escapa al ámbito de este libro.

1.4. Operaciones básicas con imágenes.

Una vez asumido que las imágenes se representan en forma de matrices, se pueden contemplar una amplia variedad de operaciones útiles. Además, estas matrices pueden tener naturaleza meramente bidimensional, por ejemplo en el caso de imágenes de luminosidad, o bien tridimensional, como en el caso de imágenes con varios *canales* o *planos cromáticos*. En el primer caso se suele hablar de imágenes monocanal y en el segundo de imágenes multicanal. En una primera aproximación, cada canal de una imagen se puede procesar independientemente del resto. En otros casos, los resultados obtenidos en los distintos canales se combinan entre sí para dar lugar a una imagen con un número distinto de canales. A continuación, citamos simplemente a modo ilustrativo algunas operaciones típicas.

- Suma de imágenes. Se utiliza por ejemplo para promediar los niveles de varias imágenes sucesivas en el tiempo, con el fin de reducir el ruido.
- Resta de imágenes. Es de uso común en tareas como detectar movimiento en secuencias de imágenes o modelar el fondo estático de una escena con relación a los objetos móviles (aunque para esto existen técnicas más elaboradas, basadas en la caracterización estadística de la variación de los niveles de cada píxel a lo largo del tiempo).
- Multiplicación de imágenes elemento a elemento. Se destina habitualmente a poner a nivel 0 ciertos puntos de una imagen, según un patrón expresado en forma de imagen o *máscara* de naturaleza lógica.
- Operaciones con escalares. La multiplicación con escalares y la suma o resta de un cierto valor a cada punto son útiles para modificar el rango dinámico de los niveles de la imagen, por ejemplo para el desplazamiento y estiramiento del histograma. Combinadas con suma de imágenes se usan para mezclado o *blending* de imágenes.
- Umbralización. Operación para el etiquetado de los píxeles de la imagen por comparación de su valor contra un cierto valor umbral. El umbral puede fijarse *a priori*, calcularse a partir de un análisis global de la imagen u obtenerse a partir del análisis de los píxeles en una cierta vecindad. En este último caso se habla de *umbralización adaptativa*. La idea se generaliza fácilmente a más de un umbral y de dos niveles de salida, en lo que se conoce como umbralización multinivel.
- Operaciones lógicas sobre imágenes binarias. Se emplean por ejemplo para la mejora del resultado tras una umbralización.

A efectos prácticos es importante tener en cuenta que, tal como ya ha sido mencionado, el resultado de las operaciones con imágenes pueden ser matrices de valores lógicos, enteros de distinto tamaño, números reales etc. Esto tiene implicaciones obvias respecto al tipo de variable más adecuado para almacenar

el resultado de la operación. Además, para visualizar el resultado en forma de imagen es preciso traducir los valores a niveles manejables por el dispositivo de visualización: típicamente tres enteros positivos, por ejemplo de un byte, para representar las intensidades de rojo, verde y azul. Esta traducción se realiza mediante tablas de consulta (*look up tables o LUTs*) asociadas al dispositivo. Es importante que estas tablas se configuren adecuadamente en cada caso, con el fin de evitar una mala interpretación del resultado de las operaciones.

1.5. El producto de convolución

Además de las operaciones básicas, el denominado *producto de convolución* o simplemente *convolución* constituye una operación de gran interés en visión, debido a su versatilidad y a sus excelentes propiedades matemáticas. Se suele expresar matemáticamente como una correlación,

$$J(x,y) = W(x,y) \circledast I(x,y) = \sum_{s=-M/2}^{M/2} \sum_{t=-N/2}^{N/2} W(s,t)\ I(x+s, y+t)\ . \quad (1.2)$$

Aquí, I es la imagen de entrada o fuente, J es la imagen de salida o destino y W una matriz N filas por M columnas, habitualmente de pequeño tamaño con relación a imagen, que denominaremos genéricamente *filtro* o, más específicamente según el contexto, *núcleo*, *kernel* o *máscara* de la convolución. Es importante remarcar que esta ecuación corresponde a una correlación, y no a una convolución. La diferencia entre ambas radica en que esta última conlleva un giro de π radianes del núcleo. Sin embargo, en visión resulta frecuente utilizar el término *convolución* como sinónimo de correlación, en particular cuando los valores del núcleo se obtienen por aprendizaje, y así lo haremos en este libro salvo que se indique lo contrario.

La convolución, además de ser lineal, tiene las propiedades conmutativa y asociativa. Esto permite, entre otras cosas, calcular la convolución de una imagen con una secuencia de filtros como su convolución con un único filtro, resultante de convolucionar aquellos entre sí. De esta manera se simplifica la manipulación matemática y se reduce el coste computacional.

$$W_2(x,y) \circledast (W_1(x,y) \circledast I(x,y)) = (W_2(x,y) \circledast W_1(x,y)) \circledast I(x,y)\ . \quad (1.3)$$

El número de operaciones necesario para completar una convolución crece cuadráticamente con el tamaño del filtro. Afortunadamente muchos filtros útiles son *separables*, esto es, resultan de convolucionar dos filtros monodimensionales entre sí. Aquí la propiedad asociativa permite convolucionar la imagen con uno de ellos, y el resultado con el otro, lo que conlleva una reducción cuadrática del número de operaciones.

Por lo demás, los principales aspectos a definir de cara a realizar una convolución son los siguientes:

- Rellenado (o *padding*).
- Paso (o *stride*).
- Los pesos y el tamaño del núcleo.

El rellenado deriva de que los cálculos no pueden ser completados en los puntos cuya distancia a los límites de la imagen sea inferior a la mitad del tamaño del filtro en la dimensión correspondiente. En este caso existen dos alternativas. La primera consiste en asumir que la imagen destino tendrá un tamaño inferior al de la imagen fuente, con la consecuente reducción de las necesidades de cómputo y almacenamiento. Esto puede resultar ventajoso para el tratamiento encadenado de grandes cantidades de datos, como en el caso de las redes convolucionales profundas. La segunda alternativa consiste en rellenar los datos faltantes alrededor de la imagen con un valor prefijado o una copia del dato más próximo en la imagen. De esta forma se pueden completar los cálculos, aun asumiendo que el resultado será erróneo en los puntos de la periferia. Las imágenes fuente y destino tendrán el mismo tamaño lo que simplifica la programación y la electrónica.

En cuanto al paso, es la distancia que avanza el núcleo entre una posición y la siguiente. La definición formal dada por la ecuación 1.2 contempla que el núcleo recorra la imagen de entrada avanzando una posición cada vez, esto es, con un *paso* de 1. Otras veces se prefiere un paso mayor, por ejemplo del mismo tamaño que el núcleo, para acelerar los cálculos y reducir el tamaño de la imagen de salida.

Por lo demás, los coeficientes o pesos del filtro, junto el tamaño de este, determinan el tipo de filtrado que se obtiene con la convolución. Así, existen filtros basa bajos, de realzado, de derivación y un largo etcétera. Generalmente, los filtros cuyos pesos suman más de 1 tienden a aclarar la imagen y los que suman menos tienden a oscurecerla. Cuando los pesos suman 1, la luminosidad general no se ve afectada. Para las operaciones de derivada primera y segunda es habitual recurrir a filtros cuyos pesos suman cero.

El filtro de la media de $3x3$ es un ejemplo clásico de filtro de convolución,

$$\frac{1}{9}\begin{pmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{pmatrix}. \tag{1.4}$$

Se utiliza para reducir el ruido en las imágenes, a costa de disminuir el contraste. La intensidad del filtrado puede aumentarse incrementando el tamaño del filtro. Además, es frecuente recurrir a filtros de forma cuadrada, al menos cuando la relación de aspecto de la distancia entre píxeles está próxima a 1:1. Existen alternativas para el filtrado de ruido que no son de convolución, como el popular filtro de la mediana que se comporta como un excelente pasa bajos. Sin embargo, conlleva un elevado coste computacional y presenta peores propiedades matemáticas que el filtro de la media, por lo que su uso está menos extendido.

Volviendo a la convolución, tradicionalmente se han venido utilizando núcleos de pequeño tamaño para ciertas operaciones básicas, el filtrado de ruido o el

Figura 1.3: Imagen de prueba.

cálculo de gradientes de luminosidad en la imagen. El filtro de Prewit es un buen ejemplo. Utiliza dos núcleos que se obtienen convolucionando el filtro de la media con las aproximaciones centrales de la derivada en las direcciones horizontal y vertical, $(-1, 0, 1)$ y $(-1, 0, 1)^t$:

$$\begin{pmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix} \quad , \quad \begin{pmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix} . \tag{1.5}$$

Otro filtro popular es el de Sobel, que utiliza dos núcleos de convolución cuyos pesos resultan de añadir, a las componentes horizontal y vertical de la derivada en cada punto, la contribución de las derivadas a 45^{o} proyectadas sobre la dirección en cuestión:

$$\begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix} \quad , \quad \begin{pmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{pmatrix} . \tag{1.6}$$

La Figura 1.3 muestra un ejemplo de imagen de prueba, en este caso de un solo canal cromático, y la Figura 1.4 el resultado de aplicar el citado filtro. Los valores mínimos (más negativos) se muestran en negro y los máximos en blanco. El resto se visualizan en tonalidades intermedias de gris.

Los filtros de Prewitt y Sobel permiten obtener fácilmente el módulo del vector gradiente y su dirección en cada punto de la imagen. Los máximos locales del módulo a lo largo de dicha dirección son buenos candidatos a puntos de borde de los objetos presentes en la imagen. El también popular filtro de Canny

(a) Componente horizontal.

(b) Componente vertical.

Figura 1.4: Filtrados de Sobel de la imagen de prueba.

proporciona una conectividad mejorada entre estos puntos a costa de un mayor esfuerzo computacional, sobre la base de conectar puntos de gradiente elevado a través puntos con gradiente más reducido.

Filtros similares pueden usarse para resaltar bordes de regiones en direcciones específicas de la imagen, como por ejemplo:

$$\begin{pmatrix} -1 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \quad y \quad \begin{pmatrix} 0 & 1 & 1 \\ -1 & 0 & 1 \\ -1 & -1 & 0 \end{pmatrix} . \tag{1.7}$$

Los filtros basados en derivada segunda constituyen un interesante complemento a los anteriores. El operador *divergencia del gradiente*, también conocido como *laplaciana*, permite detectar zonas de la imagen donde la luminosidad alcanza máximos o mínimos locales. Algunas discretizaciones comunes de este operador son

$$\begin{pmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{pmatrix} \quad y \quad \begin{pmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{pmatrix} . \tag{1.8}$$

Los pasos por cero de la laplaciana de una imagen en los puntos donde el gradiente es significativo sirven para detectar de forma precisa los puntos de borde. Por su parte los máximos y mínimos corresponden a puntos pertenecientes a líneas de mayor o menor nivel que el entorno, respectivamente.

Los filtros descritos hasta aquí se utilizan en ciertas aplicaciones prácticas por cuanto representan un cierto compromiso entre complejidad, coste computacional y calidad de los resultados. Sin embargo, existen alternativas ventajosas para tratar con las distintas frecuencias espaciales que suelen aparecer en las imágenes. Muchas de estas alternativas se basan en el denominado filtro *gaussiano*, cuyos pesos siguen una distribución gaussiana bidimensional caracterizada por su desviación típica σ,

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \ . \tag{1.9}$$

El filtro gaussiamo proporciona un suavizado más natural que el de la media. Además, el grado de suavizado se controla de forma precisa a través del parámetro σ. Es también separable, como el de la media, pero presenta unas excelentes propiedades matemáticas que facilitan su manipulación analítica. Por otra parte, hay que tener en cuenta que se trata de un filtro de respuesta impulsional infinita (*Infinite Impulse Response*, IIR), es decir, tiene un número infinito de coeficientes no nulos. Sin embargo, a efectos prácticos se puede aproximar satisfactoriamente por un filtro de respuesta impulsional finita (*Finite Impulse Response*, FIR) de radio 3 σ. Un ejemplo de filtro gaussiano, para $\sigma = 1$, es

$$\begin{pmatrix} 0,00 & 0,01 & 0,02 & 0,01 & 0,00 \\ 0,01 & 0,06 & 0,10 & 0,06 & 0,01 \\ 0,02 & 0,10 & 0,16 & 0,10 & 0,02 \\ 0,01 & 0,06 & 0,10 & 0,06 & 0,01 \\ 0,00 & 0,01 & 0,02 & 0,01 & 0,00 \end{pmatrix} \ . \tag{1.10}$$

La figura figura 1.5 representa de forma visual la distribución espacial de estos pesos. Este tipo de representación será utilizada con frecuencia en el presente texto. Además, salvo donde se indique lo contrario, asumiremos que los niveles de visualización se han ajustado de forma lineal entre negro para el valor mínimo del filtro (aquí, 0,00), y blanco para el valor máximo (aquí, 0,16).

La figura 1.6 muestra el resultado de filtrar una imagen con filtros gaussianos de distinta σ. En este caso se trata de una imagen con tres canales cromáticos. Cada uno ha sido procesado por separado, si bien existen otras alternativas que discutiremos más adelante.

El filtro gaussiano se utiliza con frecuencia en combinación con operadores de derivada primera y segunda. La Figura 1.7 muestra un ejemplo de filtro tipo gradiente de gaussiana, que permite resaltar los bordes de regiones más claras o bien más obscuras que el fondo.

La Figura 1.8 muestra un ejemplo de aplicación de este filtro. En comparación con el filtro de Sobel, los bordes se encuentran mejor definidos. Además, el parámetro σ permite ajustar el comportamiento del filtro.

Por supuesto estos filtros pueden ser escalados en forma anisotrópica y rotados, de forma que presenten una mejor respuesta a tipos específicos de bordes, orientados en distintas direcciones. La Figura 1.9 muestra algunos ejemplos.

Los rasgos lineales, es decir grupos de píxeles con diferente valor que los del entorno, alineados en cierta dirección de la imagen, pueden ser resaltados

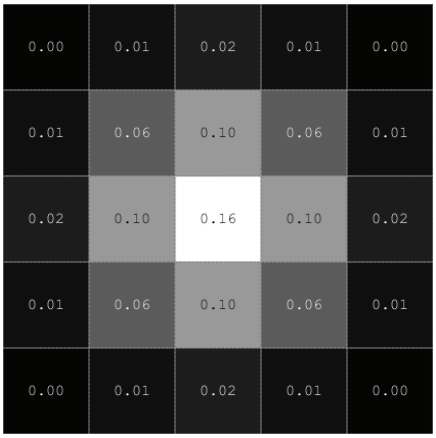


Figura 1.5: Filtro gaussiano de $\sigma = 1$.

mediante filtros de derivada segunda. Una primera alternativa consiste aplicar un nuevo gradiente a cada componente del gradiente obtenido mediante un filtro de primer orden. La Figura 1.10 muestra, en la parte superior, dos filtros de este tipo, orientados a lo largo de las dos direcciones principales de la imagen; y en la parte inferior el gradiente cruzado, de interés para ciertos algoritmos.

El resultado de aplicar los dos primeros filtros mostrados en esta Figura 1.10 a la imagen de prueba se muestra en la Figura 1.11.

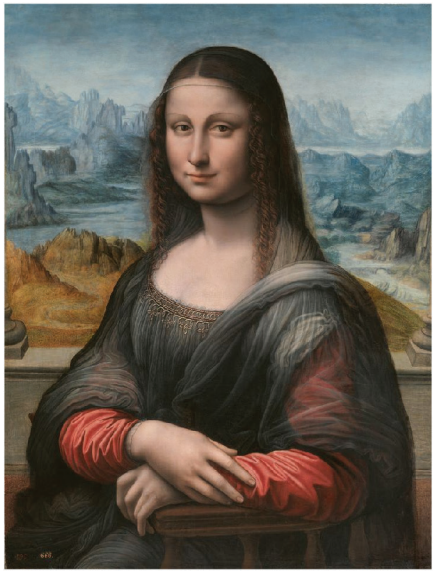
Como en el caso de los filtros de derivada primera, estos filtros también pueden ser escalados y rotados, como en la Figura 1.12, con el fin de resaltar líneas de cierto espesor y con una orientación específica

Una segunda alternativa de segundo orden consiste en calcular la divergencia del gradiente de gaussiana, más conocida como *laplaciana de gaussiana* (LOG). La Figura 1.13 muestra un ejemplo de este filtro junto con su representación en forma de gráfico tridimensional. En esta se aprecia la distribución espacial de pesos, descrita habitualmente como de *sombrero mejicano*.

El filtro LOG es más popular que el gradiente de gradiente de gaussiana por resultar más simple y proporcionar resultados comparables, como se ve en la Figura 1.14, si bien se pierde la información de direccionalidad.

Por otra parte, una propiedad interesante de la LOG es que, debido a su mencionada forma de sombrero mejicano, presenta una fuerte respuesta a *blobs*, entendidos como grupos de píxeles próximos, más oscuros o claros que el entorno. El parámetro σ controla el tamaño de los *blobs* que se detectarán. En concreto, la respuesta máxima se obtiene para *blobs* con un radio en torno a $\sqrt{2}\sigma$. La Figura 1.15 muestra un ejemplo. Además, el filtro LOG también puede ser escalado anisotrópicamente y rotado, como en el caso de los filtros discutidos anteriormente, para detectar blobs de distinto tamaño, forma y orientación.

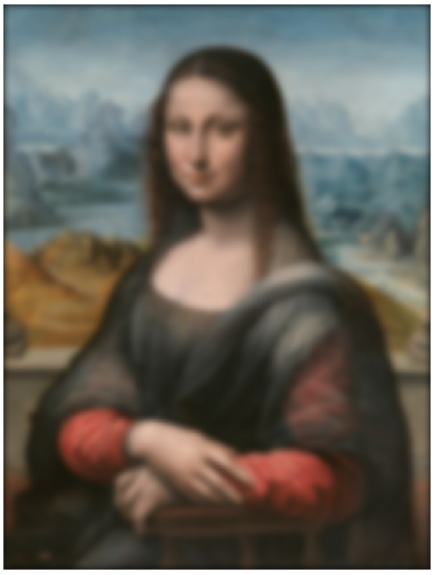
Por otra parte, una LOG se puede aproximar por la Diferencia entre dos Gaussianas (DOG) de distinta σ, o incluso por la diferencia entre una imagen y su filtrado gaussiano. De hecho, las células ganglionares de la retina realizan esta



(a) Imagen original, sin aplicar filtrado.

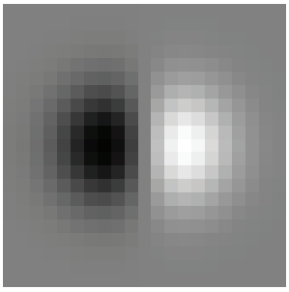
(b) Filtrado con una gaussiana de $\sigma = 2$.

(c) Filtrado con una gaussiana de $\sigma = 4$.

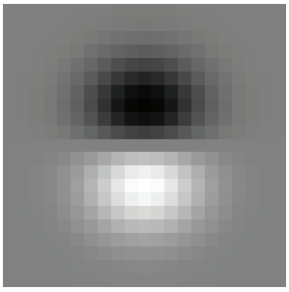


(d) Filtrado con una gaussiana de $\sigma = 6$.

Figura 1.6: Imagen de 954 x 720 píxeles filtrada con filtros gaussianos de distinta σ.



(a) Componente horizontal.



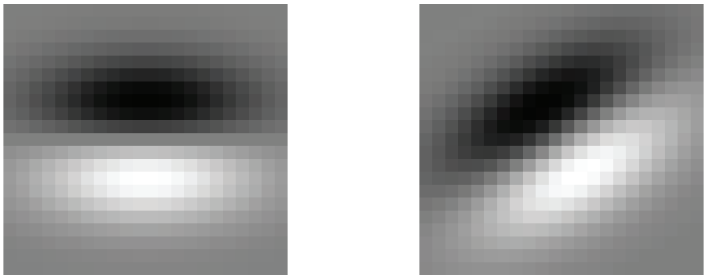
(b) Componente vertical.

Figura 1.7: Filtro de gradiente de gaussiana de $\sigma = 3$.

(a) Componente horizontal.

(b) Componente vertical.

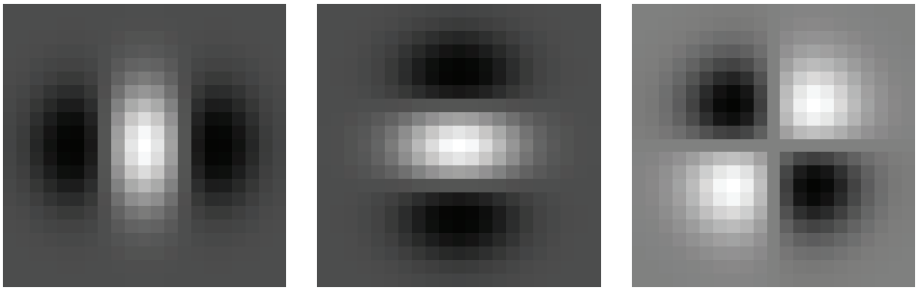
Figura 1.8: Resultado de un filtrado de gradiente de gaussiana de $\sigma = 3$.



(a) Gradiente vertical escalado anisotrópicamente.

(b) El mismo gradiente, rotado.

Figura 1.9: Filtro de gradiente de gaussiana.



(a) Gradiente horizontal de segundo orden.

(b) Gradiente vertical de segundo orden.

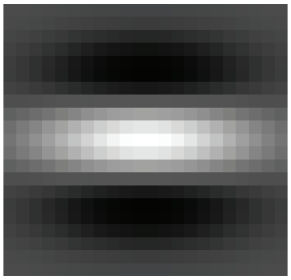
(c) Gradiente cruzado.

Figura 1.10: Filtros de gradiente de un gradiente de gaussiana: (a) gradiente horizontal de la componente horizontal, (b) gradiente vertical de la componente vertical, y (c) un gradiente cruzado (igual al otro por la conmutatividad de la convolución).

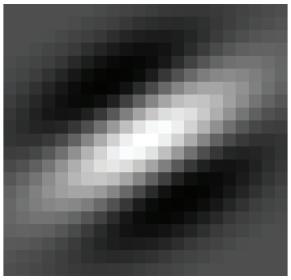
(a) Gradiente horizontal de segundo orden.

(b) Gradiente vertical de segundo orden.

Figura 1.11: Resultados de un gradiente de gradiente de gaussiana.

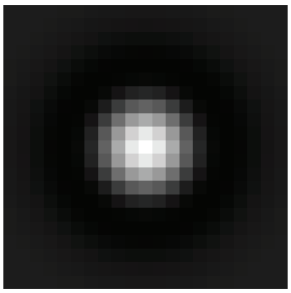


(a) Gradiente vertical de segundo orden, escalado.

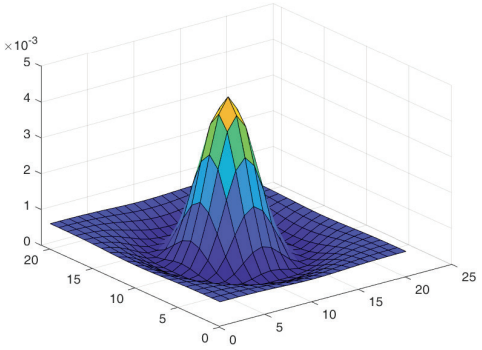


(b) El mismo gradiente, rotado.

Figura 1.12: Filtro de gradiente de un gradiente de gaussiana, escalado anisotrópicamente y rotado.



(a) Laplaciana de gaussiana.



(b) Visualización gráfica de los pesos.

Figura 1.13: Laplaciana de una gaussiana de $\sigma = 3$.

Figura 1.14: Resultado de aplicar una laplaciana de gaussiana de $\sigma = 3$.

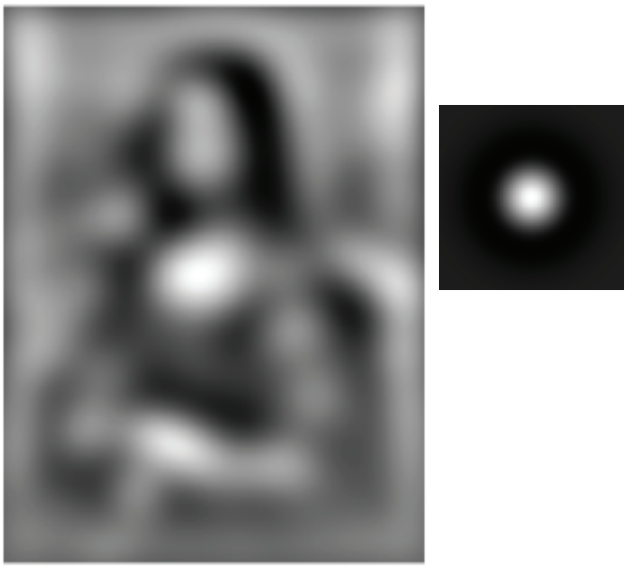


Figura 1.15: Resultado de aplicar una laplaciana de una gaussiana de $\sigma = 45$ (mostrada a la derecha).

operación para favorecer la detección de los objetos. Así operan también ciertos algoritmos de procesamiento de imágenes como los populares SIFT y SURF, para detectar y caracterizar zonas de interés de la imagen a distintas escalas, así como ciertos algoritmos para la construcción de pirámides multiescala utilizados en trasmisión de imágenes y reconocimiento de objetos.

Sobre ideas análogas a las anteriores se construyen otros tipos de filtros, como filtros de derivada de orden mayor o, por ejemplo, los filtros de Gabor. Estos se basan en senoides de frecuencia y fase ajustable, moduladas por gaussianas centradas en las zonas de la imagen bajo análisis, como ilustra la Figura 1.16.

En todo caso, sí es importante notar que, más allá de estos operadores generales, y teniendo en cuenta que en definitiva estamos calculando correlaciones, se

Figura 1.16: Un ejemplo de filtro de Gabor.
Fuente: Chabacano, CC BY-SA 3.0, via Wikimedia Commons

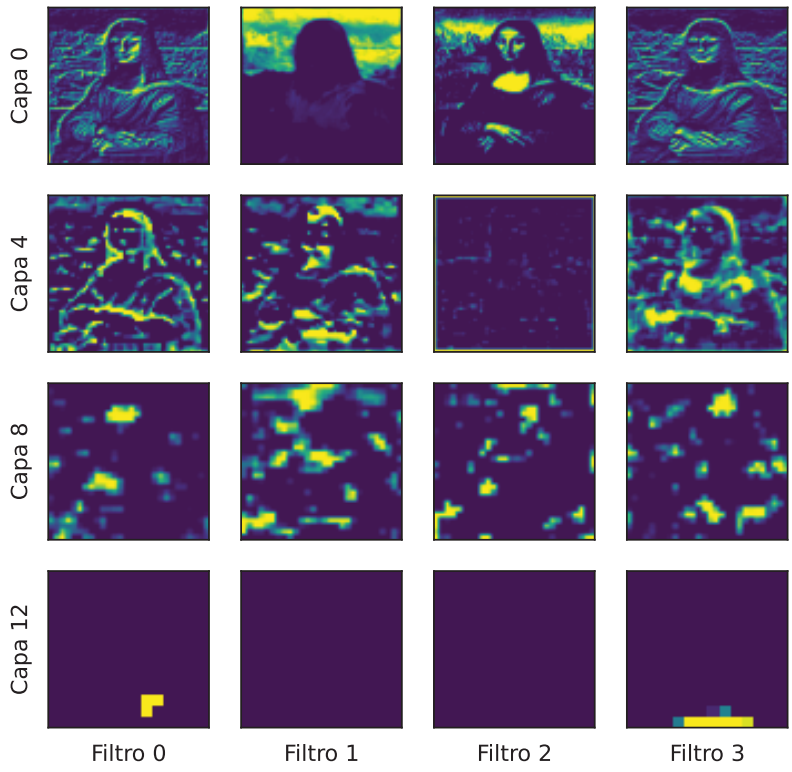


Figura 1.17: Filtros de la red convolucional VGG-16.

pueden diseñar filtros específicos para detectar rasgos particulares. En la Figura 1.17 se muestran algunos ejemplos de filtros calculados por la red convolucional VGG-16 [1] en distintas capas del modelo. Estos filtros se obtienen durante el entrenamiento y permiten la detección de líneas, bordes de regiones, círculos, patrones repetitivos, etc. en distintas orientaciones del plano de la imagen.

El problema a la hora de aplicar todos estos filtros en la práctica consiste en dar con el filtro adecuado en cada caso y ajustar sus pesos convenientemente. Esto puede abordarse sobre la base de hipótesis formuladas *a priori* acerca de los modelos matemáticos de los rasgos que se desea resaltar en las imágenes. Sin embargo, habitualmente se precisará un ajuste posterior por ensayo y error. En todo caso, por lo general estos filtros proveen un mecanismo para resaltar primitivas de bajo nivel, pero se requiere un esfuerzo adicional considerable para detectar formas más complejas. Esto puede abordarse por agrupamiento de las

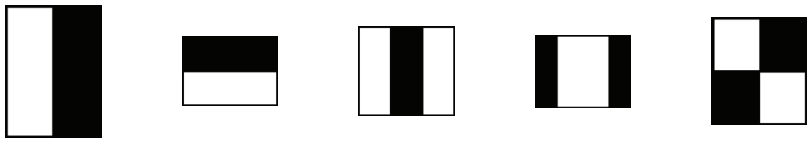


Figura 1.18: Ejemplos de características de Haar.

primitivas de bajo nivel, pero la forma de hacerlo dista de ser evidente. Las redes neuronales convolucionales, que constituyen el objeto de gran parte del presente libro, representan una solución adecuada a ambos problemas: proveen un mecanismo para el aprendizaje de los filtros y sus pesos, y permiten agrupar primitivas de bajo nivel sobre la base de un esquema piramidal de resolución decreciente.

Por otra parte, existen aproximaciones alternativas a las planteadas en esta sección tanto sobre la base de filtros con pesos predefinidos como aprendidos. Un ejemplo notable son las denominadas *características de Haar*, núcleos rectangulares estructurados como agrupación de núcleos más pequeños con pesos 1 y -1. El número y la forma de las características que resultan útiles en cada zona de la imagen se aprenden mediante un algoritmo de *boosting* (Figura 1.18). Esta solución se caracteriza por presentar un coste computacional de operación reducido cuando se implementa sobre la base de *imágenes integrales*, por lo que se aplica en electrónica de consumo. Sin embargo, por lo general se asume que presenta un campo de aplicación más restringido.

Por último, cabe notar que el carácter espacial de los filtros descritos permite su manejo directo e intuitivo en muchas aplicaciones prácticas. No obstante, cuando la imagen presenta patrones aproximadamente repetitivos (p.ej. texturas o ruido periódico) puede resultar ventajoso realizar el filtrado en el dominio frecuencial ($pixeles^{-1}$). En este dominio, las distintas frecuencias espaciales de la imagen se localizan en zonas específicas del plano frecuencial. Además, los productos de convolución se convierten, en este plano, en productos punto a punto entre las transformadas de Fourier de la imagen y de la máscara espacial de convolución. Esta circunstancia puede ser aprovechada para acelerar la convolución con máscaras de tamaño elevado, mediante la transformación previa al dominio frecuencial, convolución en este y antitransformación del resultado.

1.6. Convolución multicanal

La discusión precedente se ha centrado en la convolución monocanal. La operación se puede aplicar también a imágenes multicanal, sin más que procesar cada canal por separado, tal como hicimos en el ejemplo de la Figura 1.6. El resultado es una imagen con el mismo número de canales que la imagen de partida.

Un tipo diferente de convolución es lo que se conoce como *convolución multicanal* y corresponde a generalizar la idea a tres dimensiones. Existen distintas formas de abordar esta generalización y la diferencia entre ellas es sutil. Además,

la terminología no es estándar ni siempre consistente. Todo ello obliga a cierta reflexión previa que planteamos a continuación.

Partiremos de un concepto generalizado de imagen formada por N filas, M columnas y C *capas*. Cada capa puede estar estructurada, a su vez, como un conjunto de planos o canales P. Por ejemplo, una imagen de grises constará de una sola capa con un solo canal. Una imagen color RGB estará estructurada como una matriz de una capa con tres canales, R, G, y B. Un fragmento de vídeo con n imágenes RGB consistirá en una matriz de n capas, cada una integrada por tres canales. Por otra parte, no asumiremos ninguna restricción *a priori* sobre el tipo valores que contendrá una imagen. Estos podrán estar relacionados directamente con luminosidad o color, o bien ser el resultado de filtrados previos encaminados a resaltar ciertas características, tales como líneas, bordes, blobs etc. Cuando queramos referirnos específicamente a uno u otro tipo de imágenes utilizaremos el término *imagen óptica* o *imagen de características*, según corresponda. En lo que concierne a los filtros, habitualmente usaremos los términos *filtro* para referirnos a una matriz tridimensional y *núcleo* para cada uno de los planos que conforman un filtro. Sin embargo, tal como se ha indicado no se trata de una terminología estándar, por lo será utilizada con cierta flexibilidad.

Volviendo al tema de discusión, una primera forma de generalizar la convolución monocanal es lo que denominaremos *convolución bidimensional multicanal*. Se parte de una imagen con una o más capas de P canales cada una, y de un filtro integrado por P núcleos. Cada canal de una capa de la imagen se convoluciona con uno de tales núcleos, y los resultados obtenidos se suman entre sí. De ello resulta, por cada capa de entrada de P canales, una capa de salida con un único canal. El mismo proceso se repite para las demás capas. Debe notarse que el núcleo recorre cada capa de partida en forma bidimensional, es decir únicamente a lo largo de filas y columnas, de ahí la denominación convolución bidimensional multicanal.

Un ejemplo sencillo, útil en no pocas ocasiones, consiste en convolucionar una imagen de P canales con un núcleo de tamaño 1 x 1 x P. Por ejemplo, la convolución de una imagen RGB con un núcleo de tamaño 1 x 1 x 3 con pesos 0,2126, 0,7152 y 0,0722 proporciona una imagen de grises que preserva la sensación perceptual de luminosidad de la imagen color original (para las tres fuentes de luz definidas por el estándar sRGB y asumiendo que no se aplica corrección gamma). La Figura 1.3 que venimos utilizando como ejemplo fue obtenida de esta manera.

La Figura 1.19 muestra otro ejemplo, esta vez correspondiente a aplicar un filtro conformado por tres núcleos de 5 x 5, con sus 25 valores igualados a 0,2126, 0,7152 y 0,0722 respectivamente. Su aspecto es similar al de la Figura 1.3 pero suavizada por efecto del filtrado de la media. Por otra parte, por supuesto nada impide utilizar núcleos distintos para cada canal y veremos ejemplos de ello en capítulos posteriores.

Una segunda generalización posible de la convolución monocanal es lo que se conoce como *convolución tridimensional*, propiamente dicha. En este caso se asume que la imagen de partida es una matriz tridimensional, como por ejemplo

Figura 1.19: Convolución multicanal con tres máscaras de $5x5x3$, con sus 25 pesos iguales a 0,2126, 0,7152 y 0,0722 respectivamente.

las que proporcionan los equipos de tomografía o resonancia magnética. El filtro es también una matriz tridimensional, habitualmente con un número de filas, columnas y planos reducido en comparación con el de la imagen de partida. La operación es la generalización directa de la descrita en la Ecuación (1.2), a saber:

$$J(x,y,z) = \sum_{s=-M/2}^{M/2} \sum_{t=-N/2}^{N/2} \sum_{u=-P/2}^{P/2} W(s,t,u)\ I(x+s, y+t, z+u)\ . \quad (1.11)$$

El filtro recorre la imagen fuente a lo largo de filas, columnas y planos, dando como resultado una imagen tridimensional de tamaño comparable a la original. Este tipo de procesamiento se utiliza por ejemplo para la mejora y segmentación de imágenes de resonancia, tomografía, rayos X y similares.

1.7. Procesamiento del resultado del filtrado

Una vez completados los filtrados previos, el *pipeline* tradicional de procesamiento continúa con la segmentación. El método más simple consiste en umbralizar directamente el resultado del filtrado mediante $u(J(x,y) - th)$, donde u es la función escalón (Figura 1.20), $J(x,y)$ el nivel en el punto (x,y) resultante de los filtrados precedentes y th el umbral elegido.

Por supuesto esta técnica sencilla admite un sin fin de perfeccionamientos, tales como la umbralización multinivel, el cálculo adaptativo de umbrales, la consideración de criterios de vecindad y conectividad, consideraciones semánticas de más alto nivel, etc. Algunas ideas se han apuntado ya en la sección

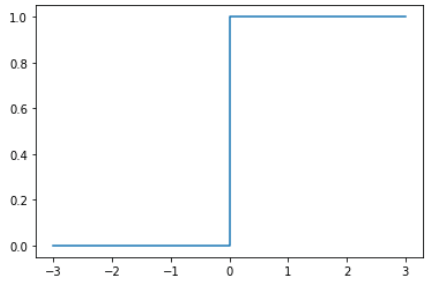


Figura 1.20: Función escalón.

de operaciones con imágenes, o cuando se mencionaron los filtros de Sobel y Canny, aunque existen muchas otras técnicas cuya discusión escapa al ámbito del presente libro.

Por otra parte, en el contexto del aprendizaje profundo se recurre también a alternativas a la función escalón tales como la función sigmoidea (o la similar, tangente hiperbólica) o la función *ReLu*. La primera puede entenderse como una aproximación derivable de la función escalón. Esto resulta de trascendencia cuando se busca optimizar automáticamente los parámetros del filtrado mediante técnicas de derivada, por ejemplo de gradiente descendente como veremos en capítulos posteriores. Por otra parte, cuando el filtrado transcurre a través de varias etapas en cascada, suele resultar ventajoso utilizar la mencionada función *ReLu* o alguna aproximación derivable. Esta función proporciona 0 como valor de salida donde el resultado de una convolución ha sido negativo (es decir, donde se ha encontrado una correlación negativa con el filtro), o el propio valor de la entrada en otro caso. Ello permite centrar las etapas subsiguientes del procesamiento en los rasgos para los que se ha detectado cierta correlación, aún pequeña, entre la imagen y los filtros empleados. Todos estos temas se relacionan íntimamente con las redes neuronales, las redes convolucionales y el aprendizaje profundo, por lo cual serán tratados con detalle en capítulos posteriores.

Por último, las etapas finales de extracción de características y reconocimiento de objetos para la interpretación de las escenas quedan también fuera del ámbito del presente libro y para su estudio nos remitimos a los textos especializados sobre el tema. En todo caso, las técnicas que describimos en los siguientes capítulos permiten obviar gran parte de estas etapas, constituyendo así un camino alternativo hacia la comprensión de las información visual.

1.8. Transformaciones geométricas

Con independencia de las operaciones descritas hasta ahora, numerosas tareas de visión comportan la realización transformaciones geométricas de las imágenes. Las más comunes son las denominadas genéricamente *transformaciones afines*, que conservan el paralelismo. Las más comunes son la traslación, el escalado, la rotación y la inclinación o cizalladura. Se suelen expresar en *forma*

homogénea como:

$$\begin{pmatrix} x_T \\ y_T \\ 1 \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \tag{1.12}$$

donde (x, y) son las coordenadas originales y (x_T, y_T) las resultantes de la transformación. En general, asumiremos que todas las coordenadas vienen dadas en unidades de píxel, es decir, como (*columna*, *línea*). La tercera ecuación no aporta información adicional pero facilita el encadenamiento y la inversión de transformaciones.

Distintos valores de los parámetros c_{ij} dan lugar a transformaciones diferentes. Por ejemplo,

$$\begin{pmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} S_x & 0 & 0 \\ 0 & S_y & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & -c_x & 0 \\ -c_y & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{1.13}$$

expresan, respectivamente, una traslación de vector (t_x, t_y), un escalado (isotrópico en el caso $S_x = S_y$), un giro de ángulo θ en torno al origen (que asumiremos en la esquina superior izquierda de la imagen), y una inclinación de parámetros c_x, c_y. Desde un punto de vista práctico es importante notar que suele resultar ventajoso programar estas transformaciones mediante bucles que recorren la imagen transformada en lugar de la imagen original. Así, el nivel en cada posición (x_T, y_T) de la imagen transformada se obtienen a partir del nivel en la posición (x, y) de la imagen original mediante:

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ 0 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} x_T \\ y_T \\ 1 \end{pmatrix} . \tag{1.14}$$

Además, por lo general las coordenadas obtenidas de esta forma, (x, y), no adoptan valores enteros. Por ello, el nivel que corresponderá a (x_T, y_T) puede obtenerse a partir de las coordenadas redondeadas al vecino más próximo a (x, y) o bien, de forma más elaborada, mediante una interpolación bilineal a partir de los 4 vecinos más próximos a (x, y) o bicúbica a partir de 16 vecinos, con el fin de reducir el *pixelado* de la imagen transformada.

Las transformaciones geométricas descritas constituyen un caso particular de las denominadas genéricamente *transformaciones homográficas*,

$$\begin{pmatrix} x_T\ n \\ y_T\ n \\ n \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} , \tag{1.15}$$

que representan una proyección entre dos planos. Aquí, n es el *parámetro homogéneo* cuyo valor, no nulo, viene dado por la tercera ecuación. Las coordenadas transformadas (x_T, y_T) se calculan dividiendo $(x_T\ n,\ y_T\ n)$ entre n.

(a) Fotografía original del objeto. (b) Transformación homográfica.

Figura 1.21: Resultado de aplicar una homografía, en este caso encaminada a obtener una vista frontal (a invertir la homografía correspondiente a la captura de la imagen).

La Figura 1.21 muestra un ejemplo en el que se ha invertido la transformación homográfica correspondiente a una captura de una imagen con el fin de obtener una perspectiva frontal. Los parámetros se han determinado mediante un ajuste de mínimos cuadrados a partir de cuatro puntos (las esquinas de la hoja de papel).

La transformación homográfica es, a su vez, un caso particular de la *transformación perspectiva* entre el espacio tridimensional y un plano de proyección,

$$\begin{pmatrix} x_T\ n \\ y_T\ n \\ n \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & c_{1z} & c_{13} \\ c_{21} & c_{22} & c_{2z} & c_{23} \\ c_{31} & c_{32} & c_{3z} & c_{33} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} . \tag{1.16}$$

La captación de imágenes de una escena tridimensional por medio de una cámara es un buen ejemplo de transformación perspectiva.

1.9. Remuestreo y pirámides de imágenes

El denominado *remuestreo* es una operación común en visión. Permite modificar el tamaño de las imágenes de cara a su manejo ulterior. Este tamaño se especifica habitualmente en forma de *resolución*, entendida como número de

píxeles que conforman la imagen. El remuestreo puede aplicarse en dos sentidos: submuestreo (*subsampling*) y sobremuestreo (*upsampling*).

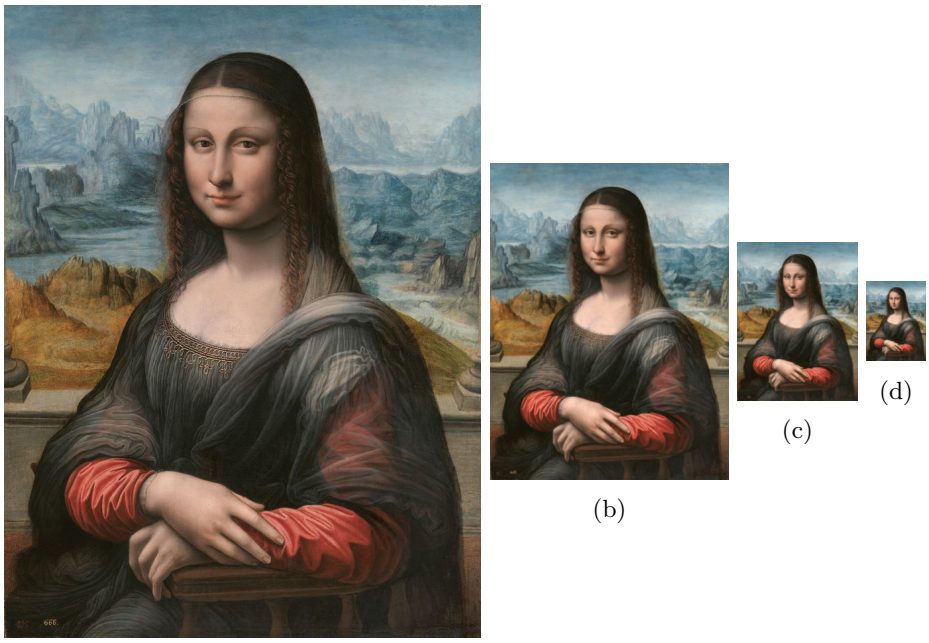
El submuestreo consiste en remuestrear la imagen a una frecuencia espacial inferior, es decir un periodo de muestreo superior. Esto resulta útil para reducir los requerimientos computacionales en tareas como el reconocimiento de objetos o el almacenamiento y la transmisión de imágenes. La mecánica es análoga a la convolución: se define una ventana de análisis que recorre la imagen a lo largo de sus filas y columnas, con un cierto paso (mayor que 1). Para cada posición se calcula el valor que se asignará al píxel correspondiente de la imagen de salida, siguiendo alguna de las alternativas siguientes:

- Selección de la media (*average pooling*)
- Selección gaussiana (*gaussian pooling*)
- Selección del máximo (*max pooling*):

En el primer caso la ventana de análisis implementa un filtro de la media y en el segundo un filtro gaussiano. Se trata de filtros pasa-bajos que permiten reducir los eventuales fenómenos de *aliassing*. La Figura 1.22 muestra un ejemplo: el resultado de un submuestrear una imagen a frecuencias decrecientes en un factor 0,5 , para construir lo que denomina una pirámide de imágenes. (Por supuesto podría haberse empleado otro factor). Como filtro pasa bajos se ha recurrido a una aproximación de una gaussiana de $\sigma \approx 1.082$, de uso común en muchas bibliotecas de procesamiento de imágenes.

En otras ocasiones, en particular cuando se procesan imágenes de características, se persigue específicamente que la imagen remuestreada conserve los rasgos más sobresalientes de la imagen de partida. En tales condiciones resulta preferible recurrir a una selección del máximo valor en la vecindad definida por la ventana de análisis. Esta técnica es de uso común en redes convolucionales por lo que será retomada en capítulos posteriores.

El sobremuestreo, por su parte, consiste en remuestrear la imagen de partida a una frecuencia espacial superior a la de entrada, es decir con un periodo de muestreo inferior. La resolución de salida será superior a la de entrada. Esto conlleva realizar interpolaciones bilineales o bicúbicas, como las mencionadas anteriormente, o incluso, en ocasiones, a interpolaciones guiadas por aprendizaje. La Figura 1.23 muestra un ejemplo, en el que se ha partido de la imagen correspondiente al nivel 1/8 de la Figura 1.22. En general, la operación de submuestreo no es reversible dado que comporta pérdida de información. Por ello, las imágenes de las Figuras 1.22(a) y 1.23(d) no son iguales (la segunda, de 954 x 720 píxeles, se ha generado a partir de la primera, de tan solo 120 x 90 píxeles). Por otra parte, la combinación submuestreo-sobremuestreo se utiliza con frecuencia en arquitecturas tipo codificador-decodificador como veremos en capítulos posteriores, por ejemplo para reducir el coste computacional de reconocer un objeto en una imagen y reubicarlo luego, de forma aproximada, en la imagen original.



(a) (b) (c) (d)

Figura 1.22: Ejemplo de pirámide decreciente imágenes. De izquierda a derecha: imagen original y resultado de su remuestreo gaussiano a octavas decrecientes sucesivas (frecuencias 1/2, 1/4 y 1/8 $pixeles^{-1}$).

(a) (b) (c) (d)

Figura 1.23: Ejemplo de pirámide creciente de imágenes. De izquierda a derecha: imagen de partida (tomada de la figura 1.22(d)), y resultado de su remuestreo a octavas crecientes sucesivas (frecuencias 2, 4 y 8 $pixeles^{-1}$).

Capítulo 2

Componentes de un sistema de visión artificial

2.1. Introducción

El desarrollo de un sistema de visión artificial requiere una elección cuidadosa de los componentes involucrados en la captación de las imágenes. Estos componentes son fundamentalmente la cámara, óptica y la iluminación. A ellos dedicamos los primeros apartados de este capítulo. También se hace una mención expresa de los sistemas 3D, de creciente difusión tanto en la industria como en la vida cotidiana. Sobre la base de todos estos principios se plantean finalmente algunas consideraciones metodológicas relativas a la selección del *hardware* de los sistemas de visión.

2.2. Cámaras

La captación de imágenes es el punto de partida de todo sistema de visión. Para esta labor se necesita una cámara, entendida como un sensor de imagen más su electrónica asociada, y un sistema óptico para concentrar y enfocar la energía luminosa. En la presente sección resumimos algunos aspectos o características destacables de las cámaras, en tanto que las ópticas se estudiarán específicamente en la sección 2.3.

2.2.1. Características de las cámaras

A la hora de elegir o analizar una cámara deben tenerse presentes diferentes aspectos. A continuación, se revisan los siguientes:

- Resolución
- Tamaño del sensor, tamaño del píxel y relaciones de aspecto

- Imágenes por segundo
- Ganancia, *offset* y corrección *gamma*
- Tecnología del sensor

Resolución

En el ámbito de las cámaras, la resolución se entiende como el número total de fotosensores, es decir *píxeles*, que contiene el sensor. Por ejemplo, un sensor de 1028 (columnas) x 720 (filas) tendrá una resolución próxima a los 0.75 megapíxeles. La resolución suele elegirse de manera que, para las condiciones previstas de captura de las imágenes, el menor elemento a detectar cubra un mínimo de 2 x 2 píxeles para evitar *aliassing*). En la práctica, es más recomendable llegar a los 5 x 5 o 10 x 10 píxeles.

En todo caso, es importante notar que la resolución que proporciona la cámara puede ser diferente a la del sensor (y, de hecho, suele serlo). Esto se debe a que la electrónica puede remuestrear la señal a número diferente de columnas y filas. A su vez el propio software de procesamiento de las imágenes también puede remuestrearlas. Desde luego, el remuestreo introduce errores y *artefactos* por lo que, en lo posible, es preferible trabajar con la resolución *nativa* del sensor o bien submúltiplos enteros de esta.

Tamaño del sensor, tamaño de píxel y relaciones de aspecto

El tamaño del sensor de las cámaras más utilizadas en visión suele estar entre 0.5" y 2" (pulgadas), medido en diagonal. A igualdad de otros factores, un tamaño mayor se traduce en píxeles con mayor superficie y, por tanto, más sensibles a la luz. Además, los sensores de *píxel contiguo*, es decir sin espacio muerto entre los píxeles, aprovechan menor la superficie disponible, a costa de una mayor complejidad microelectrónica.

También es importante considerar la relación o *ratio* de aspecto ancho/alto del propio sensor (por ejemplo, 4:3 o 16:9) y, más especialmente, de las distancias entre sus píxeles. Una cámara de píxel cuadrado, es decir con píxeles distanciados en un ratio 1:1, resulta generalmente preferible pues simplifica el software de procesamiento (por ejemplo, un círculo frontal tendrá el mismo diámetro en horizontal y vertical, medido en píxeles). Por supuesto la relación de aspecto puede ser modificada por el remuestreo subsiguiente de la cámara o el software. De nuevo trabajar con valores nativos suele ser preferible.

Imágenes por segundo

El sensor de imagen integra la energía luminosa que incide en él durante un cierto *tiempo de exposición* (*exposure time*). Transcurrido este tiempo, la integración se bloquea y la información eléctrica generada se transfiere a la electrónica (*data readout*). Tras esto, la cámara queda lista para iniciar una nueva captura de imagen (*ready*), tal como se esquematiza en la figura 2.1. El número de veces por segundo que se repite este ciclo es lo que se conoce como imágenes por segundo, *fps* (*frames per second*) o *frame rate*). Además, el proceso

puede realizarse simultáneamente para todas las líneas de la imagen en lo que se conoce como *modo progresivo*, o bien primero para las impares (trama impar) y luego para las pares (trama par) en el *modo entrelazado*. Esto último permite un vídeo más fluido desde el punto de vista visual, pero complica el análisis de escenas en movimiento por el distinto instante de captura de las tramas. Adicionalmente, algunas cámaras permiten solapar el tiempo de exposición con el de transferencia para incrementar el número de imágenes por segundo.

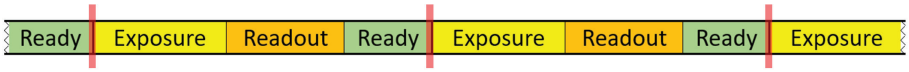


Figura 2.1: Esquema general del ciclo del ciclo de captura de imágenes

Por otra parte, muchas veces la captura se reinicia de forma cíclica, síncronamente con una señal de reloj (por ejemplo, cada 1/60 segs). Se dice que la cámara está en vídeo vivo (*freerun*). Cuando el sistema de visión solicita una imagen, esta estará disponible al cabo de un cierto tiempo de retraso, variable en función de en qué momento del ciclo llegó la solicitud (figura 2.2). Algunas cámaras soportan además la operación asíncrona: el proceso de captura se reinicia a la llegada de una señal externa de disparo ({it trigger}) con lo cual la imagen estará disponible al cabo de un tiempo fijo, predeterminado (figura 2.3).

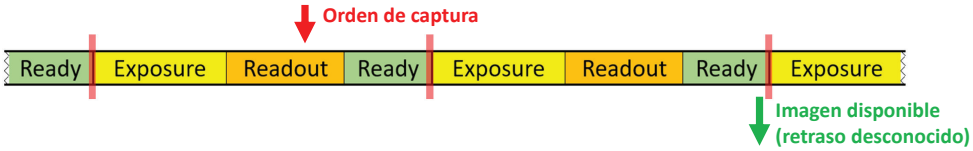


Figura 2.2: Captura en modo de disparo síncrono

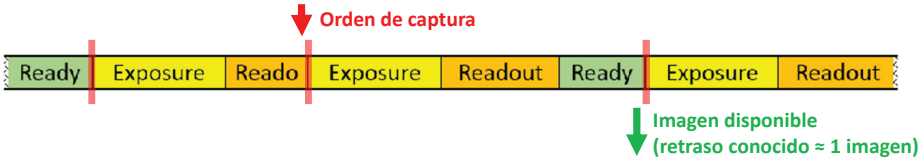


Figura 2.3: Captura en modo de disparo asíncrono

Los principales factores que limitan el número de imágenes por segundo son el tiempo de exposición, la propia electrónica de la cámara y los tiempos de transferencia de la imagen al ordenador. Por ejemplo, no es extraño encontrar cámaras de, digamos, 20 MPixels que apenas alcanzan los 5 fps. En cambio, resulta fácil encontrar cámaras de 1 MPixel que operan a 30 o 60 fps.

Ganancia, *offset* **y corrección** *gamma*

La señal analógica procedente del sensor se amplificada y convierte a valores

numéricos por medio de un convertidor A/D, habitualmente integrado en la propia cámara. Muchas veces el *offset* y la ganancia del convertidor pueden ser ajustados vía software. El mejor aprovechamiento del rango dinámico de la cámara se consigue ajustando el *offset* de forma tal que el nivel digital de salida esté próxima a 0 para la mínima excitación luminosa prevista y a 255 para la mayor, en ambos casos evitando la saturación, y asumiendo que los valores se codifican en 1 byte. De forma adicional, algunas cámaras permiten amplificar los niveles bajos de señal en mayor medida que los altos, para compensar la no linealidad del sistema visual humano y de algunos dispositivos de visualización. Esto se suele modelar a través del exponente γ en la expresión

$$n = A\ S^{\gamma} + n_0 \tag{2.1}$$

donde n es el valor numérico proporcionado por el convertidor tras su conversión a entero, S la señal eléctrica procedente del sensor, A la ganancia, n_0 el offset (en niveles) y γ la citada corrección que suele ser del orden de 0.45. Muchas cámaras utilizadas en visión permiten ajustar gamma a distintos valores, entre ellos la unidad que está indicada cuando se desea que los niveles guarden una relación más sencilla y directa con la cantidad de luz incidente en el sensor.

Tecnología del sensor

Las dos tecnologías microelectrónicas más utilizadas para la fabricación de sensores de imagen son las conocidas CCD (*Charge-Copuled Device*) y CMOS (*Complementary metal-oxide-semiconductor*). En el primer caso la incidencia de luz produce electrones que se acumulan en un condensador MOS (*metal-oxide-semiconductor*) por cada píxel. Las cargas acumuladas se transfieren luego, a través de los píxeles vecinos, hacia una estructura común que las conduce a la electrónica. En el segundo caso se utilizan transistores CMOS que se referencian y leen de forma similar a las posiciones de una memoria CMOS. Tradicionalmente los CCD proporcionaban una mayor calidad de imagen, con mejor relación señal a ruido, aunque el enorme desarrollo de la tecnología CMOS en los últimos años, derivado de su utilización masiva en microelectrónica, está cambiando la situación. Los modernos sensores CMOS ofrecen más resolución a un menor precio y permiten el acceso individualizado a regiones de interés (ROIs, *Regions of Interest*). Además, permiten la exposición simultánea a la luz de todos los píxeles de la imagen (*shutter* común), en lugar de la exposición línea a línea (*shutter* rotativo) que requieren los CMOS más antiguos y conlleva la aparición de artefactos ante escenas en movimiento. Los CMOS también se comportan mejor que los CCD ante iluminación intensa, pues en estos últimos puede producirse *blooming*, es decir desbordamiento de la carga de un píxel hacia píxeles vecinos. Por último, cabe mencionar que la sensibilidad de los sensores CMOS es una función logarítmica con la excitación luminosa: a mayor excitación, menor sensibilidad. Sin embargo, esta no linealidad puede ser corregida fácilmente por la electrónica.

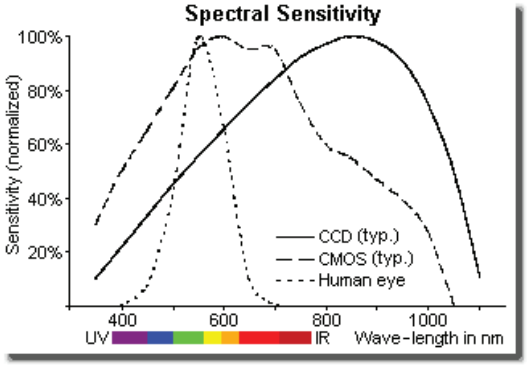


Figura 2.4: Sensibilidad espectral. Fuente: [2]

2.2.2. Cámaras monocromas y color

Por supuesto la sensibilidad de los fotosensores a la luz depende de su la longitud de onda de esta. La *sensibilidad espectral* caracteriza este comportamiento (*véase* la figura 2.4).

Así, la señal generada por cada fotosensor viene dada en primera aproximación por

$$S = A_p \int_{\lambda} V(\lambda) E_s(\lambda) d\lambda \tag{2.2}$$

donde λ es la longitud de onda de la radiación considerada, mientras que S es la señal eléctrica que genera un fotosensor de área A_p, con sensibilidad espectral $V(\lambda)$, cuando sobre él incide una irradiancia $E_s(\lambda)$. Esta última es la energía que incide en el sensor en forma de radiación electromagnética por unidad de tiempo y de superficie, tal como detallaremos más adelante. Además, tal como se ha mencionado, en el caso de los sensores CMOS la señal es una función logarítmica de esta integral, que por simplicidad no ha sido incluida explícitamente en la expresión. La señal eléctrica S es luego amplificada y convertida a valores numéricos como se expresó en la ecuación 2.1.

En todo caso, con independencia de que se trate de un sensor CCD o CMOS, la sensibilidad espectral se extiende más allá del espectro visible por un ojo humano promedio, en especial hacia el rango infrarrojo cercano. Por ello las cámaras suelen incluir un filtro óptico antepuesto al fotosensor, que bloquea esta radiación. El efecto de este filtro puede suponerse incorporado a la función de sensibilidad $V(\lambda)$ de la ecuación 2.2 por lo que no explicitaremos aquí.

Las cámaras color siguen estos mismos principios, pero utilizan tres funciones de sensibilidad espectral diferentes, $V_R(\lambda)$, $V_G(\lambda)$ y $V_B(\lambda)$, entorno a las longitudes de onda cortas, medias y largas, respectivamente (en lugar de una única $V(\lambda)$ extendida a todo el espectro visible). Estas tres funciones se diseñan de manera que la cámara proporcione la intensidad que hay que dar a, típicamente, tres fuentes de luz de colores primarios rojo, verde y azul, para reproducir la excitación luminosa incidente en el sensor.

Las tres funciones de sensibilidad espectral se implementan por medio de filtros ópticos. En las *cámaras trisensor* se usan tres sensores de imagen diferentes, cada uno dotado de su propio filtro. Un divisor óptico de haz permite dividir la luz incidente en tres haces, cada uno de los cuales se encamina hacia el correspondiente sensor con su filtro. Las *cámaras* color *monosensor* constituyen una alternativa más económica. En este caso, durante la fabricación se integra sobre cada píxel un filtro óptico individual, microscópico, que implementa la función de sensibilidad deseada. Estos píxeles se suelen agrupar de tres en tres o de cuatro en cuatro (en este caso uno para el rojo, dos para el verde y uno para el azul), a lo largo y ancho del sensor. La información obtenida es luego remuestreada de forma que para cada píxel se obtienen tres valores: el realmente capturado por el píxel en cuestión (rojo, verde o azul), y los valores de los otros dos colores obtenidos por interpolación a partir de los píxeles vecinos. Lógicamente este proceso de remuestreo conlleva sacrificar calidad de la imagen por lo que las cámaras color monosensor están más indicadas cuando no se requiere una calidad de imagen muy elevada.

Por último, cabe notar que en algunas aplicaciones prácticas puede ser conveniente usar *cámaras multiespectrales*, que cuentan con entre 4 y 10 canales cromáticos; o hiperespectrales, con más de 10. Esto permite una caracterización más precisa de la distribución espectral de la luz procedente de la escena, a costa de una menor relación de señal a ruido (por la menor energía luminosa que se capta al restringir el rango de longitudes de onda). Se utilizan comúnmente en campos como agroalimentación, teledetección, geología, cartografía, detección de contaminantes o vigilancia.

2.2.3. Cámaras lineales

Las cámaras lineales comparten las características de las convencionales, de estructura matricial, pero sus píxeles se disponen a lo largo de una línea (o tres para cámaras lineales trisensor). Esto simplifica notablemente la integración del sensor y su electrónica asociada, al tiempo que permite resoluciones y velocidades de adquisición elevadas. Están particularmente indicadas para captar imágenes de escenas en movimiento, en las que una de las dimensiones es muy grande o se encuentra poco definida. Algunas aplicaciones típicas son la inspección de rollos de material (papel, caucho, plástico, chapa), vigas, firme de carreteras, cereales o legumbres y un largo etcétera.

2.3. Ópticas

La óptica u objetivo es un conjunto de elementos que permiten enfocar la energía luminosa procedente de la escena sobre el sensor, con el fin de formar imagen. En la presente sección se revisan los principios básicos de este proceso, así como los ajustes y las características principales de las ópticas más habituales

en visión. También se presentan algunos tipos de ópticas y filtros ópticos.

Por lo demás, desde un punto de vista constructivo la óptica puede ser solidaria a la cámara o, más frecuentemente, constituir un elemento separado e intercambiable (mediante un anclaje de tipo rosca o bayoneta). Esto resulta poco relevante desde un punto de vista funcional por lo cual, en general, no haremos diferencia entre ambas soluciones.

2.3.1. La cámara *pinhole*

La luz es una radiación electromagnética y, como tal, se propaga en forma de ondas. Cuando las ondas emitidas por una fuente de luz inciden sobre la superficie de un objeto, cada punto de esta (entendido como un elemento diferencial de superficie) refleja parte de la energía incidente. Esta energía reflejada se emite en forma de ondas esféricas centradas en el punto en cuestión (con mayor o menor amplitud según la dirección considerada). Así, disponer un sensor frente a la escena no basta para formar imagen, puesto que sobre cada punto de este sensor incidiría luz reflejada por *todos* los puntos visibles de la escena, tal como ilustra la figura 2.5.

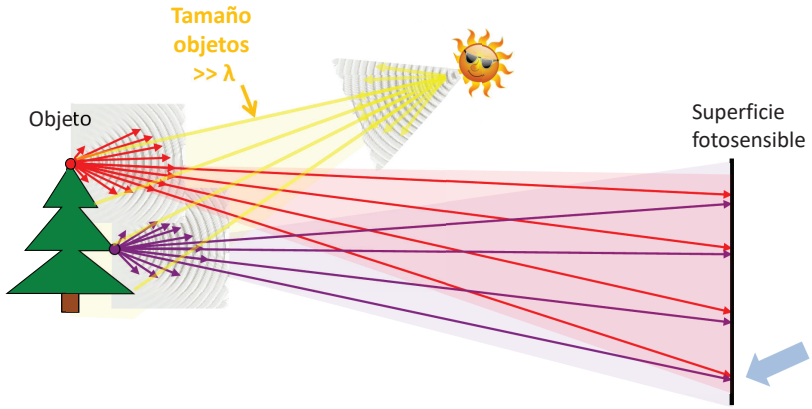


Figura 2.5: Ondas reflejadas que inciden en (toda) la superficie del sensor

La situación cambia en el caso de disponer, frente al sensor, una barrera con un pequeño orificio o *pinhole*, como muestra la figura 2.6. Ello permite que la luz incidente en cada punto del sensor provenga de una única dirección y, en definitiva, de un solo punto de la escena: se forma una imagen. Este arreglo es lo que se conoce como *cámara oscura* y es conocido desde antiguo. Existen cámaras comerciales que siguen el mismo principio, denominadas *cámaras pinhole*. Constan simplemente de una carcasa con un pequeño orificio en la parte frontal, protegido por un cristal, y un sensor de imagen en la parte posterior.

Las cámaras pinhole son muy económicas y pueden usarse en aplicaciones donde se dispone de iluminación intensa. Sin embargo, adolecen de dos limitaciones importantes. En primer lugar, el orificio debe tener un diámetro pequeño

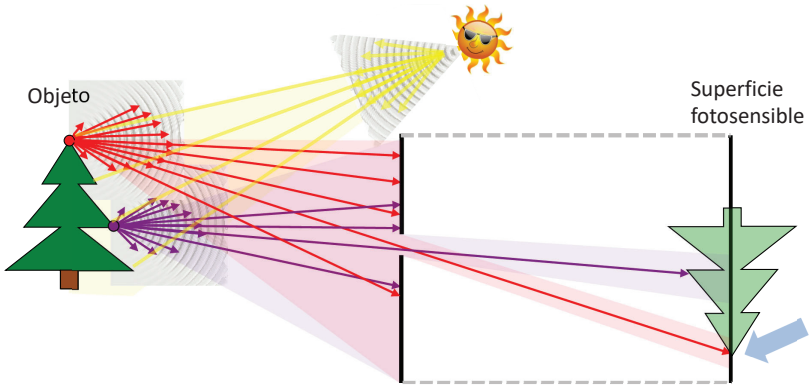


Figura 2.6: Cámara *pinhole*

para que la imagen resulte nítida. Sin embargo, cuando este diámetro es del orden la décima de milímetro se producen fenómenos de *difracción*: la luz emerge del orificio en forma de onda esférica, no de haz rectilíneo, y la imagen pierde nitidez. Además, por supuesto sería deseable captar una mayor cantidad de luz procedente de cada punto de la escena, y no solo aquélla que se dirige directamente al *pequeño* orificio. Estas dos limitaciones se soslayan con la utilización de lentes, como veremos a continuación.

2.3.2. Lentes

Una lente biconvexa con caras de geometría esférica es capaz de enfocar en un punto las ondas que inciden en ella, tal como se muestra en la figura 2.7, siempre que se asuma incidencia *paraxial*, es decir próxima al *eje óptico* de la lente, y que la luz proceda de un punto lejanode forma que el frente de onda sea aproximadamente plano. En lo que sigue asumiremos estas dos condiciones, así como que la lente es delgada y que todas las longitudes de onda convergen en un mismo punto, y no a una distancia ligeramente diferente según su longitud de onda (como de hecho ocurre en realidad).

Bajo las asunciones citadas, la formación de imágenes queda descrita mediante la *ecuación de las lentes delgadas*, que se deduce fácilmente por semejanza de triángulos (2.8):

$$\frac{1}{d_0} + \frac{1}{d_i} = \frac{1}{f} \ . \tag{2.3}$$

En esta ecuación, d_0 es la distancia de la escena a la lente, d_i la distancia de esta al plano sensor, y f la denominada *distancia focal* o, simplemente, focal. Esta última es la distancia entre el *centro óptico* de la lente y el denominado *foco*, un punto matemático donde convergen los rayos de luz que inciden paralelamente al eje óptico. La distancia focal es fundamental porque condiciona dónde se formará la imagen y con qué tamaño. Ademas, para el caso habitual de escenas distantes (en comparación con la focal) podemos asumir $1/d0 \approx 0$, con

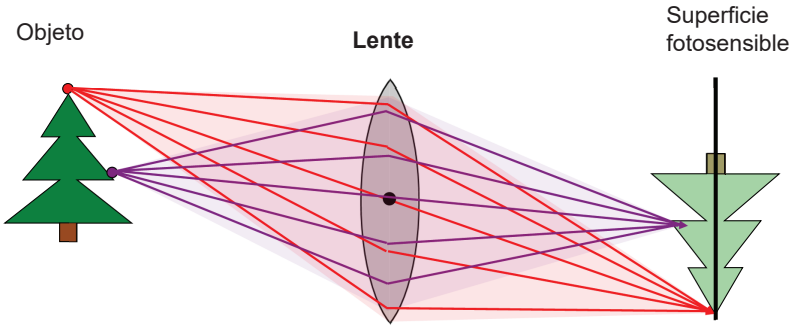


Figura 2.7: Formación de imagen mediante una lente

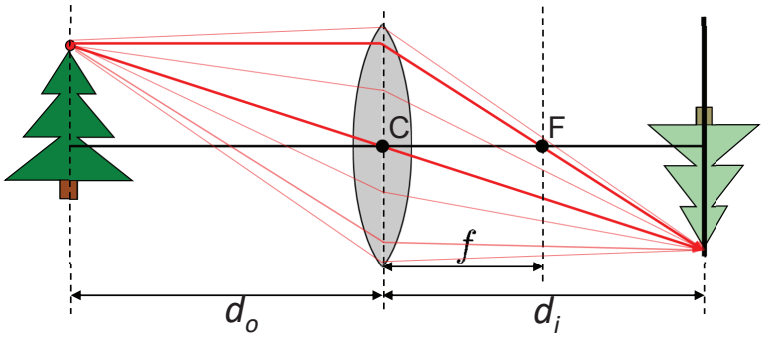


Figura 2.8: Modelo para la ecuación de las lentes delgadas

lo cual la imagen se formará a una distancia $d_i \approx f$. Esto conduce a un modelo geométrico simplificado de la formación de imágenes conocido como *modelo pinhole* por su comportamiento geométrico similar al de las cámaras pinhole descritas anteriormente. De acuerdo con el modelo pinhole, cada punto de la escena se proyecta sobre el plano sensor situado a una distancia $d_i \approx f$ del *centro óptico*, según una perspectiva cónica con vértice en dicho centro. De forma añadida, en este modelo es habitual considerar que el plano sensor se encuentra entre el objeto y el centro óptico. Por supuesto esto no corresponde a su ubicación real, pero esta imagen hipotética no aparece invertida respecto a la escena, lo cual resulta ventajoso tanto conceptual como matemáticamente. De todo ello resulta el esquema de la figura 2.9.

2.3.3. Distancia focal

La distancia focal es el parámetro determinante de la óptica por cuanto está relacionado directamente con el tamaño de la imagen que se proyectará sobre el plano sensor. Una focal corta producirá una imagen más pequeña, tal como se aprecia en la figura 2.10; y a la inversa. Sin embargo, el tamaño del

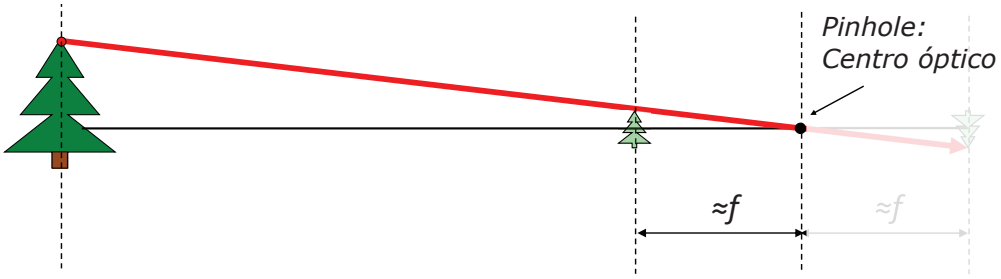


Figura 2.9: Modelo *pinhole*

sensor y su resolución también resultan determinantes. En efecto, la misma figura 2.10 muestra cómo, para una focal y una resolución dadas, la imagen que proporciona el sensor 2 será mayor que la del sensor 1, en el sentido de que los objetos presentes en ella cubrirán un mayor número de píxeles.

Así pues, resulta más útil analizar la cuestión en términos de la denominada *apertura angular*, esto es el ángulo visual que cubre la cámara considerando los límites del sensor. Para un conjunto dado de lente más sensor, se deduce fácilmente que la apertura angular η viene dada por

$$\eta = 2 \arctan \frac{T_{max}}{2f} \tag{2.4}$$

donde T_{max} es el tamaño del sensor en la dimensión considerada (ancho, alto o diagonal). Una focal menor conllevará un campo visual mayor, en el que los objetos se verán más pequeños para una resolución dada; y a la inversa.

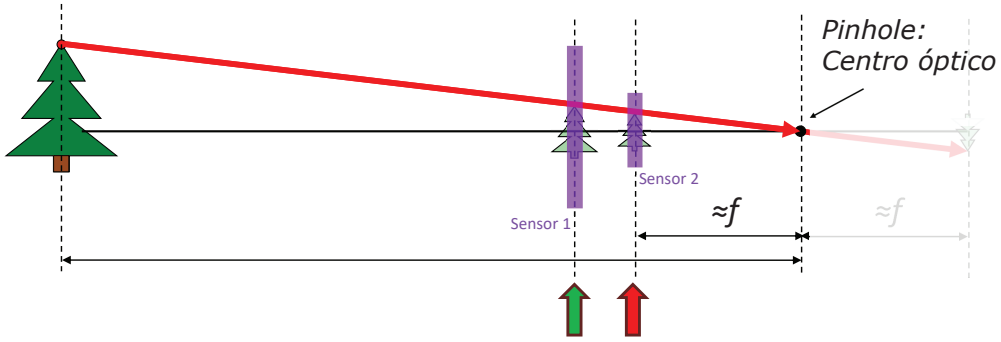


Figura 2.10: El tamaño de los objetos (en píxeles) depende de la distancia focal de la lente y del tamaño y la resolución del sensor

Por ejemplo, para el caso de un sensor de altura 35 mm y relación ancho/alto de 3/2 (que corresponde a un formato tradicional en fotografía analógica) se obtienen los campos visuales mostrados en la figura 2.11. En el caso de otros formatos, lo más práctico suele ser acudir a las tablas o *calculadoras de ópticas* que proporcionan los fabricantes de ópticas.

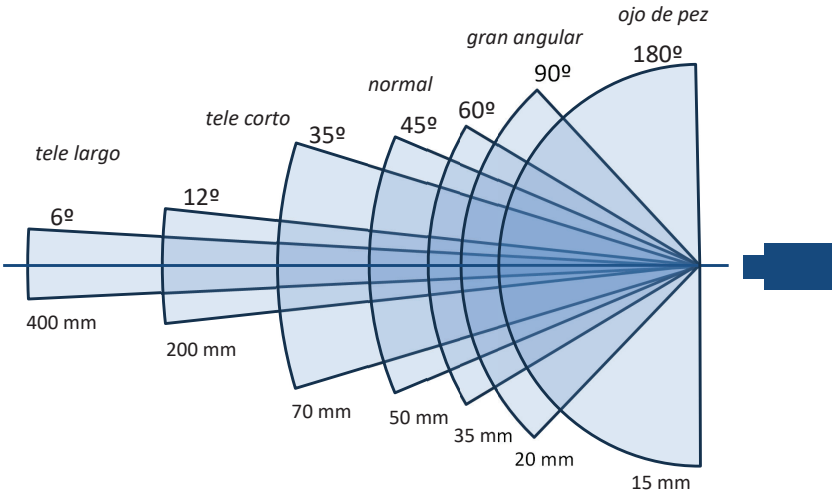


Figura 2.11: Relación entre apertura angular horizontal y distancia focal para el caso particular de un sensor de formato 35mm

En la misma figura se ha anotado también la denominación habitual de los objetivos en función de su apertura angular horizontal: tele largo (teleobjetivo largo), tele corto, normal (unos 45º), gran angular y ojo de pez (180º). Merece la pena insistir en que esta relación entre focal y ángulo rige únicamente para sensores con el formato especificado. Si el sensor tiene otro tamaño, esta relación será diferente.

Por otra parte, la distancia focal de una lente viene determinada por su geometría (el radio de su superficie) y el material en que está fabricada (más específicamente de su *índice de refracción* que cuantifica la capacidad de ese material para desviar la luz). Así pues, en principio no sería posible modificar la distancia focal de una lente una vez fabricada. Por suerte se pueden construir arreglos de varias lentes que, al desplazarse a lo largo del eje común, consiguen el efecto equivalente a un cambio de focal. Esto se explota en las denominadas *ópticas zoom*. Desde luego, estas ópticas zoom, de focal variable, son más voluminosas, caras y frágiles que sus equivalentes de focal fija. Además, es difícil garantizar una buena calidad de imagen a lo largo de todo su rango de focales. Por ello, en visión suele ser preferible trabajar con ópticas fijas, si acaso contando con ópticas zoom únicamente en la fase de prototipado para determinar experimentalmente la focal más conveniente para cada aplicación.

2.3.4. Enfoque

Las ópticas contienen, pues, una lente biconvexa o un arreglo de lentes equivalente a la misma, caracterizada por su distancia focal. La misión que cumple es enfocar, sobre cada punto del plano sensor, la energía luminosa procedente de un área de la escena. Además, el enfoque se producirá a una distancia dada del

centro óptico, que como determina la ecuación (2.3) depende de la distancia a la que se encuentra el objeto. Si esta distancia es mayor, la imagen se formará a una distancia menor del centro óptico, y a la inversa. Por ello, las ópticas deben contar con un mecanismo de *enfoque* que permita ajustar la distancia entre la lente y el plano sensor, de forma tal que el objeto bajo estudio se proyecte en el sensor con la debida nitidez. Bien entendido, la nitidez se consigue para un objeto situado a una distancia específica, que suele venir marcada sobre la propia óptica (en metros). Sin embargo, el enfoque puede considerarse correcto no ya a una distancia exacta, sino en un cierto rango de distancias que se conoce como *profundizad de campo*. La profundidad de campo depende de la distancia al objeto y de la distancia focal, como acabamos de ver, así como de la denominada *apertura focal* que analizamos en la siguiente subsección.

2.3.5. Apertura focal

La otra gran misión de la óptica es regular la cantidad de luz que incide en el sensor. Esto se consigue mediante un mecanismo de tipo *diafragma* que conforma una apertura aproximadamente circular por la que la luz llega hasta el sensor. La relación entre la luminosidad de la escena y la luminosidad de la imagen capturada por la cámara viene dada por

$$E_{sensor} = \frac{\pi}{4}\left(\frac{D}{f}\right)^2 cos^4(\alpha)\ L_{escena}(\lambda) \quad . \tag{2.5}$$

Esta ecuación, cuya deducción puede encontrarse fácilmente en los libros de óptica, vincula la *irradiancia* incidente en el sensor, E_{sensor}, con la *radiancia* emitida por el punto en cuestión de la superficie observada, L_{escena} (*véase* la figura 2.12). La irradiancia y la radiancia serán definidas con precisión en la sección correspondiente al estudio de la iluminacióm (sección 2.4). Por el momento, baste apuntar que E_{sensor} es la energía que incide en el sensor por unidad de tiempo y de superficie. Por su parte, α es el ángulo de incidencia de la luz sobre el sensor, habitualmente próximo a 90^{o} (al menos en la zona central de la imagen). Por último, D/f representa el diámetro de la apertura (circular) dividido entre la distancia focal de la lente.

El cociente f/D es lo que conoce como *Número f*, *razón focal*, *apertura focal* o, simplemente, *apertura* (no confundir con la apertura entendida como el orificio de entrada de la luz). Por ejemplo, un número f 22 significará que el diámetro de la focal es 22 veces mayor que la apertura seleccionada. Este número es fundamental porque determina la cantidad de luz que alcanzará el sensor: un número f menor se traduce en una mayor cantidad de luz, y la inversa. Por ello, el número f se encuentra claramente marcado en las ópticas, por ejemplo en la forma f/22, 1:22, 1:22D, f 22 o simplemente 22. A modo de ejemplo, un ajuste f/16 dejará pasar aproximadamente el doble de luz que un f/22, dado que $(22/16)^2 aprox 2$. Por lo demás, el número f mínimo de una óptica es un indicativo de la cantidad máxima de luz que dejará pasar. Una óptica de calidad suele tener un número f mínimo del orden de, digamos, al menos 2. Cabe remarcar

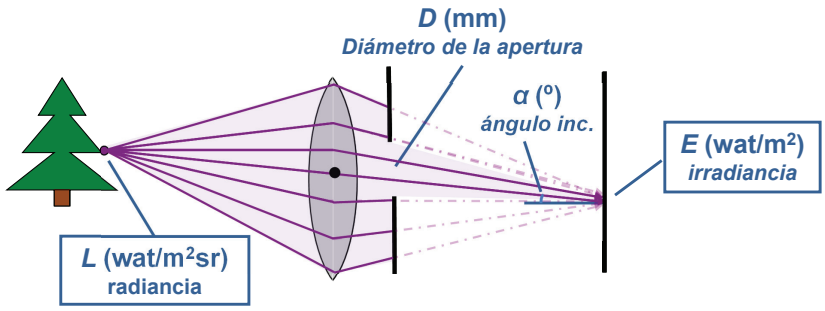


Figura 2.12: Captación de la energía luminosa

que una distancia focal larga conllevará un oscurecimiento de la imagen como se aprecia claramente en la fórmula; en definitiva, la misma energía luminosa debe *repartirse* entre más píxeles. Lógicamente la forma de compensar la pérdida de luminosidad con focales largas es permitir aperturas mayores, lo que eventualmente requerirá lentes de mayor diámetro. Esto se traduce en más peso y coste de las lentes, así como mayor fragilidad de la óptica y dificultades crecientes para enfocar en la periferia de la imagen (donde se cumple peor la aproximación paraxial).

Un último aspecto de interés es que ajustar la óptica a números f bajos proporcionará imágenes más luminosas, pero suele acarrear una menor profundidad de campo. Así, en visión artificial, resulta recomendable trabajar con valores elevados del número f y compensar el oscurecimiento de la imagen con una iluminación más intensa de la escena, siempre que ello sea posible.

2.3.6. Distorsiones, aberraciones y otros fenómenos

El comportamiento de las ópticas reales difiere, en ocasiones significativamente, del ideal. Se habla de distorsiones y aberraciones.

Un caso frecuente es la distorsión geométrica tipo barril, en que la imagen aparece comprimida hacia la periferia del sensor. De hecho, la denominación *tipo barril* deriva de la forma característica que adoptan los rectángulos en presencia de esta distorsión. La distorsión de barril suele ser notoria en las ópticas con focal corta (y, de hecho, suele introducirse *ex profeso* en este tipo de ópticas para evitar que los objetos en la zona central de la imagen aparezcan excesivamente pequeños). Otras veces se produce la distorsión contraria, denominada tipo cojín, normalmente en ópticas zoom por una mala compensación de la distorsión tipo barril para ciertas focales. La figura 2.13 ilustra ambos tipos de distorsiones. El propio software de procesamiento de imágenes puede corregir estas distorsiones tras una adecuada calibración, si bien a costa de modificar la apertura angular con que se verá la imagen.

La aberración esférica es otro fenómeno frecuente. Consiste en diferencias de enfoque entre la parte central de la imagen y la periferia. Las ópticas de calidad

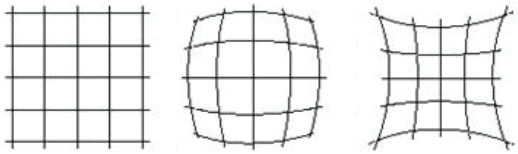


Figura 2.13: Imagen ideal y con distorsiones de tipo barril y cojín

corrigen este fenómeno por medio de lentes con una geometría particular, más costosas de fabricar. También surge la aberración cromática. Consiste en que la luz que atraviesa la lente queda enfoca a distancias ligeramente distintas según cuál sea su longitud de onda (debido que el índice de refracción varía con esta: las longitudes de onda largas se desvían menos que las cortas). El problema se corrige incorporando en la óptica parejas de lentes con distinto índice de refracción.

En general, estos y otros defectos pueden evitarse empleando ópticas de calidad, especialmente si se desea trabajar con lentes de gran diámetro para captar mayor cantidad de luz.

Existen otros fenómenos de etiología diversa que también afectan a la calidad de las imágenes y que son bien conocidos en el campo de la fotografía. Un ejemplo es el *viñeteo*, que consiste en un oscurecimiento de la periferia de la imagen, más notorio con focales cortas y diafragmas abiertos. Se debe, entre otras causas, a que el objetivo bloquea parte de la luz que incide con ángulos grandes. También puede mencionarse el conocido *efecto estrella*, provocado por la difracción de la luz en los vértices de las hojas del diafragma. Pueden también mencionarse las vibraciones de la cámara durante la captura de vídeos, que pueden compensarse mediante estabilizadores de imagen, es decir sistemas electromecánicos que modifican la orientación de las lentes en función de la realimentación proveída por un conjunto de giróscopos. En general, disponer de ópticas de calidad y trabajar con buenas condiciones de captura de imagen es la mejor recomendación contra todos estos fenómenos.

2.3.7. Ópticas especiales

Existen distintos tipos de ópticas adaptadas a condiciones particulares de captura de imágenes, que describimos aquí brevemente.

Las ópticas *zoom*, es decir con focal variable, han sido ya comentadas por lo que no procede extenderse más. La recomendación general es reservar este tipo de ópticas para el desarrollo de prototipos, por los motivos ya apuntados. Las ópticas con un campo angular amplio se construyen a veces usando arreglos con espejos parabólicos o incluso sistemas multicámara (con posterior pegado de las imágenes).

Otras ópticas de uso particular son las denominadas *telecéntricas*, capaces de producir imágenes sin efecto de perspectiva. Se trata de una solución cara y voluminosa (la óptica debe tener un diámetro similar a la del objeto a medir), y el efecto de telecentricidad solo se consigue a una distancia muy concreta de