UNIVERSIDAD AUTÓNOMA DE MADRID

Programa de Doctorado:
Ingenieria Informática y Telecomunicación (RD2011)

Tesis Doctoral

ATENCIÓN VISUAL BASADA EN UN ESPACIO PERCEPTUAL CONJUNTO DE COLOR Y BRILLO PARA LA MEJORA DE ALGORITMOS DE SEGUIMIENTO DE OBJETOS EN SECUENCIAS DE VÍDEO

Víctor Fernández-Carbajales Cañete

Director:
Dr. Miguel Ángel García García

Escuela Politécnica Superior - Tecnología Electrónica y de las Comunicaciones
Septiembre de 2017

Abstract

This doctoral thesis performs an in-depth review of current visual attention models based on biological approaches to propose a new visual attention model based on a joint perceptual space of both colour and brightness, and shows that this model is able to extract more discriminant visual features, especially when dealing with objects that are very similar visually. That joint colour and brightness space is based on a biologically-inspired theoretical perceptual model originally proposed by Izmailov and Sokolov in the scope of psychophysics. The present thesis proposes a computational model that allows the application of Izmailov and Sokolov's theoretical model to digital images, since the original model can only be applied to perceptual data directly drawn from psychophysical experiments. Experimental results with real video sequences show that the proposed visual attention model yields significantly more accurate results in the particular application scope of video tracking than well-known visual attention models that process colour and brightness separately.

Keywords: Visual attention models, Perceptual spaces, Tracking, Colour and Brightness.

Resumen

Esta tesis doctoral realiza una revisión en profundidad de los actuales modelos de atención visual basados en enfoques biológicos, con el fin de proponer un nuevo modelo de atención visual basado en un espacio perceptual conjunto de color y brillo. Además, se demuestra que este modelo es capaz de extraer características visuales más discriminantes, especialmente cuando se trata de objetos visualmente muy similares. Ese espacio conjunto de color y brillo se basa en un modelo perceptual teórico de inspiración biológica, originalmente propuesto por Izmailov y Sokolov en el ámbito de la psicofísica. La presente tesis propone un modelo computacional que permite la aplicación del modelo teórico de Izmailov y Sokolov en imágenes digitales, ya que el modelo original sólo puede aplicarse a datos perceptuales directamente extraídos de experimentos psicofísicos. Los resultados experimentales con secuencias de vídeo reales demuestran que el modelo de atención visual propuesto produce resultados significativamente más precisos en el ámbito particular del seguimiento en secuencias de vídeo que los modelos de atención visual actuales, los cuales procesan el color y el brillo por separado.

Palabras clave: Modelos de atención visual, Espacios perceptuales, Seguimiento, Color y Brillo.

Agradecimientos

Al finalizar un trabajo tan arduo y lleno de dificultades como el desarrollo de una tesis doctoral es inevitable que te asalte un muy humano egocentrismo que te lleva a concentrar la mayor parte del mérito en el aporte que has hecho. Sin embargo, el análisis objetivo te muestra inmediatamente que la magnitud de ese aporte hubiese sido imposible sin la participación de personas e instituciones que me han facilitado las cosas para que este trabajo llegue a un feliz término. Por ello, es para mí un verdadero placer utilizar este espacio para ser justo y consecuente con ellas, expresándoles mis agradecimientos.

Debo agradecer de manera especial y sincera al Profesor Miguel Ángel García García por aceptarme para realizar esta tesis doctoral bajo su dirección. Su apoyo y confianza en mi trabajo y su capacidad para guiar mis ideas ha sido un aporte invaluable, no solamente en el desarrollo de esta tesis, sino también en mi formación como investigador. Las ideas propias, siempre enmarcadas en su orientación y rigurosidad, han sido la clave del buen trabajo que hemos realizado juntos, el cual no se puede concebir sin su siempre oportuna participación.

Quiero expresar también mi más sincero agradecimiento al Profesor José María Martínez Sánchez por haberme facilitado siempre los medios suficientes para llevar a cabo todas las actividades propuestas durante el desarrollo de esta tesis. Este agradecimiento también se extiende al resto de miembros del VPULab, por su apoyo como compañeros de fatigas, cafés, ideas y guías que tanto hacen faltan durante este largo camino.

Finalmente, pero no por ello con menor relevancia, a mi esposa y padres, por permitirme centrarme en la realización de esta tesis, apoyarme en los momentos difíciles y sobre todo apreciar el trabajo que realizaba cada día.

Índice general

Índice de figuras

Índice de tablas

Notación

$\triangle C$	Diferencia cromática
$\triangle W$	Diferencia acromática
$\triangle S$	Diferencia perceptual de Izmailov y Sokolov
$\triangle E$	Diferencia perceptual según el CIE76
$\triangle RG$	Diferencia cromática en oposición rojo-verde
$\triangle BY$	Diferencia cromática en oposición azul-amarillo
$\triangle I$	Diferencia cromática de la intensidad
JND_i	Índice JND
$\overline{JND}$	JND normalizado
JND_{max}	Índice JND máximo
JND_{min}	Índice JND mínimo
α	$\overline{JND_i}$ del estímulo
β	$\overline{JND_i}$ del fondo
$\eta(\beta)$	Polígono de primer orden de η sobre β
$\ominus$	Operador de diferencia entre escalas
$\bigoplus$	Operador de adición entre escalas
$\psi_i(x,y)$	Imagen en escala i
$\mathscr{F}_t(x,y)$	Valores de una característica en $\psi_i(x,y)$
$\mathscr{F}_{c,s}(x,y)$	Valor absoluto de la diferencia entre escalas c (centro) y s (envolvente)
$\overline{\mathscr{F}}(x,y)$	Mapa de conspicuidad
$\mathscr{J}_t(x,y)$	Mapa de saliencia cromático-acromático
$\mathscr{J}(x,y)$	Mapa de conspicuidad cromático-acromático
$\mathscr{O}_t(x,y)$	Mapa de saliencia de orientación (Sobel)
$\mathscr{O}(x,y)$	Mapa de conspicuidad de la orientación
$\mathscr{S}(x,y)$	Mapa de saliencia final

INTRODUCCIÓN

Capítulo 1

Introducción

El seguimiento de objetos en secuencias de vídeo tiene como objetivo determinar la trayectoria de objetos en movimiento procesando cada una de las imágenes que componen una secuencia de vídeo (Yilmaz et al., 2006), (Borji and Itti, 2013). Una vez seleccionados uno o varios objetos en la imagen inicial, el proceso de seguimiento identifica esos objetos en imágenes posteriores, es decir, asocia los objetos identificados en una imagen dada con los objetos previamente identificados en imágenes anteriores. El seguimiento de objetos sigue siendo un problema abierto debido a la gran complejidad de los entornos en los que normalmente debe aplicarse. Por ejemplo, para las aplicaciones de vídeo-vigilancia en entornos complejos (por ejemplo, en un centro comercial), debe ser capaz de hacer frente a las siguientes condiciones: **iluminación deficiente** o cambiante, con la presencia de **gran número de objetos de pequeño tamaño** debido a la ubicación de las cámaras de seguridad, además de **muy similares** respecto a sus características visuales, y que **interactúan entre sí** con patrones de comportamiento complejos, rápidos y muy frecuentes.

Por lo tanto, uno de los aspectos clave y principales desafíos del seguimiento de objetos en vídeo es la caracterización adecuada de los objetos elegidos para ser seguidos con el fin de poder identificarlos sin ambigüedad en las siguientes imágenes, siempre que sigan presentes. Esa caracterización debe basarse en rasgos medibles extraídos de los propios objetos, relacionados principalmente con su patrón de movimiento y/o su apariencia visual. Respecto a esta última caracterización, es decir, mediante su apariencia visual, es crucial poder detectar características visuales que puedan ser correlacionadas de manera fiable entre imágenes sucesivas.

Las características visuales que tradicionalmente se han utilizado para el seguimiento son las esquinas, los contornos, así como las regiones de interés. Desafortunadamente, su utilidad puede

estar significativamente comprometida cuando se consideran en escenarios complejos con una calidad de imagen limitada, tal como se indicó anteriormente. Por un lado, se pueden percibir esquinas sin el suficiente contraste o definición para ser detectadas de forma fiable. Incluso si se pueden detectar, nada garantiza que su entorno sea lo suficientemente discriminante como para poder identificarlas visualmente, así como a los objetos asociados en imágenes sucesivas. Además, la presencia de ruido en la imagen puede conducir a la detección de bordes falsos que conlleven fallos de robustez y decremento del rendimiento del algoritmo. A su vez, los contornos pueden ser incluso menos fiables en este contexto, ya que, en general, no garantizan una delimitación correcta, y por tanto, no permiten una correcta identificación de los objetos seguidos en escenas concurridas, con auto-oclusiones y continuas agrupaciones. Algo similar se aplica a las regiones, ya que su apariencia (por lo general, distribuciones de color o textura) suele caracterizarse de manera bastante aproximada y, por tanto, ambigua a través de una amplia variedad de modelos estadísticos (histogramas, espaciogramas, *Kernels*, ...). Por otro lado, las regiones son también muy sensibles a las auto-oclusiones y agrupaciones, debido a su naturaleza descriptiva global, lo que las hace aún menos fiables.

Por lo tanto, es necesario considerar enfoques alternativos para extraer características visuales fiables que permitan una mejor caracterización de los objetos seguidos en función de su apariencia visual. Tal alternativa es la atención visual (Borji and Itti, 2013), (Begum and Karray, 2011). En particular, los modelos de atención visual intentan imitar el proceso cognitivo que permite a los humanos concentrarse en los elementos más relevantes (características) de una escena visual, ignorando a los demás. En otras palabras, estos modelos buscan encontrar rasgos visuales cuya saliencia los hace destacar del resto de la escena visual.

Los llamados modelos de atención visual basados en la saliencia, como el modelo ampliamente conocido propuesto en (Itti et al., 1998), así como sus variaciones posteriores, procesan independientemente diversas características visuales, en particular: el color (cromaticidad), el brillo y la orientaciones (textura). De esta manera, las grandes variaciones locales de color producirán medidas de alta saliencia de color independientemente del brillo asociado a esas regiones. Este comportamiento plantea un problema en el caso de regiones oscuras o mal iluminadas, en las que el color no se puede percibir con precisión debido a un brillo insuficiente. En esas condiciones de bajo brillo, las variaciones de color pueden ser simplemente el resultado del ruido de la imagen, en cuyo caso, los modelos de atención visual mencionados anteriormente tendrán, paradójicamente, tendencia a detectar como características fiables regiones de imagen que no son, en absoluto,

distintivas ni discriminatorias. Por lo tanto, se produce la perdida de la supuesta ventaja de la atención visual para mejorar la caracterización visual de los objetos de una imagen. Desafortunadamente, esas condiciones de bajo brillo son bastante típicas en aplicaciones reales, tales como la vídeo-vigilancia y la monitorización.

Los problemas asociados con la percepción del color en regiones oscuras (también se produce en regiones sobrexpuestas con colores desaturados) han sido previamente identificados en el contexto de la consistencia computacional de color (Gijsenij et al., 2011), que tiene por objeto modificar el color de los píxeles de una imagen de modo que se perciban igual sin importar los colores de las fuentes de luz que iluminan la escena. En particular, la intensidad de color (Brown et al., 2012) se ha propuesto como una estimación heurística de la fiabilidad del tono de un píxel, con el fin de indicar como poco fiables los tonos de los píxeles oscuros y/o mal saturados. El objetivo es que la consistencia del color sólo tenga en cuenta píxeles fiables. Este problema proviene del hecho de que el tono del color es independiente de su intensidad y saturación, lo que se aplica a todos los modelos invariantes de color (Geusebroek et al., 2001).

Esta tesis propone un nuevo modelo de atención visual basado en el cálculo de la saliencia mediante el procesamiento conjunto del color y del brillo, de acuerdo con un modelo perceptual teórico de inspiración biológica originalmente propuesto por Izmailov y Sokolov en el ámbito de la psicofísica. Dado que este modelo sólo es aplicable a datos perceptuales directamente extraídos de experimentos psicofísicos, el presente trabajo propone un modelo computacional que permite la aplicación directa del modelo teórico de Izmailov y Sokolov a imágenes digitales. Los resultados experimentales con secuencias de vídeo reales muestran que el modelo de atención visual propuesto mejora significativamente la tarea del seguimiento de objetos en secuencias de vídeo al encontrar características visuales más discriminantes, especialmente cuando se trata de objetos con una alta similitud visual, obteniendo resultados más precisos en este ámbito de aplicación que otros modelos de atención visual ampliamente conocidos que procesan el color y el brillo por separado.

1.1. Objetivos

1.1.1. Objetivos Generales

El objetivo principal de esta tesis es proponer un nuevo modelo de atención visual que fusiona color y brillo, lo cual permite una caracterización más discriminatoria y robusta de los objetos, incluso si son visualmente muy similares, que puede ser explotada en aplicaciones como el seguimiento de objetos. En particular, el modelo perceptual teórico de inspiración biológica propuesto por Izmailov y Sokolov en el ámbito de la psicofísica es el fundamento de los desarrollos presentados en esta tesis.

Dos principios impulsan el logro del objetivo anterior: las teorías psicológicas y la evidencia psicofísica sobre cómo los humanos perciben el color y el brillo conjuntamente, así como el método matemático que permite su aplicación sobre imágenes digitales.

1.1.2. Objetivos Específicos

Los objetivos específicos de esta tesis se enumeran a continuación:

1. Estudio del modelo psicofísico de Izmailov y Sokolov.
2. Desarrollo de una adaptación computacional del modelo de Izmailov y Sokolov.
3. Desarrollo de un modelo de atención visual basado en el modelo computacional de Izmailov y Sokolov.
4. Desarrollo de un algoritmo de seguimiento explotando el modelo de atención visual anteriormente indicado.
5. Demostración objetiva de que el modelo propuesto mejora el rendimiento del seguimiento de objetos en secuencias de vídeo.

1.2. Organización de la Tesis

Esta tesis consta de ocho capítulos que se han organizado en cuatro partes, a saber: la introducción, el modelo perceptual de Izmailov y Sokolov, el modelo de atención visual y las conclusiones. La parte introductoria incluye este capítulo, así como el capítulo 2, que presenta una revisión de los temas transversales relacionados con esta tesis. Con el fin de separar los temas transversales de

este trabajo, se han distribuido en secciones sobre los diferentes campos relacionados, justo antes de la discusión de cada método propuesto.

La parte I se dedica a la descripción del modelo perceptual de Izmailov y Sokolov y su adaptación computacional. Así, el capítulo 3 presenta un análisis en profundidad del método psicofísico perceptual, centrándose en su significado perceptual y los estudios realizados para su obtención. Por su parte, el capítulo 4 describe la adaptación computacional del modelo teórico, preparándolo para sus usos posteriores en los modelos de atención visual.

La Parte II propone un modelo de atención visual basado en la arquitectura de IKN (véase la sección 2.2), pero integrando el color y el brillo a través de la adaptación computacional anteriormente presentada, así como su uso para el seguimiento de objetos, y la validación objetiva de la mejora introducida por este modelo en el ámbito de aplicación indicado. En primer lugar, el capítulo 5 describe la extensión del modelo de atención visual de IKN basada en el modelo perceptual de Izmailov y Sokolov. A continuación, se ha propuesto un método robusto de seguimiento de objetos visualmente similares en secuencias de vídeo en el capítulo 6. Finalmente, los resultados experimentales muestran que esta técnica se comporta significativamente mejor que los enfoques previos del estado del arte, tal como se recoge en el capítulo 7.

En la parte final, el capítulo 8 resume las contribuciones de esta tesis y propone futuras líneas de investigación.

Capítulo 2

Estado del Arte

Este capítulo sintetiza el contexto actual relativo a los temas sobre los que pivota el trabajo propuesto en esta tesis doctoral. Con este fin, este capítulo se divide en tres secciones, cada una dedicada a uno de los aspectos tratados en los siguientes capítulos. En concreto, en estas secciones se presenta:

1. Una revisión de los modelos perceptuales de color a través de los espacios de color utilizados para su representación.
2. Los principales modelos de atención visual propuestos por la comunidad científica.
3. Una revisión de las aproximaciones actuales al problema del seguimiento de objetos en secuencias de vídeo.

2.1. Modelos perceptuales de color

La representación del color en las imágenes digitales es una de las piedras angulares del procesamiento de imágenes. Si hablamos de los modelos perceptuales de color, es inevitable hacer una revisión de los espacios de color predominantes en la actualidad. Hay cuatro modelos principales (aunque algunos presentan variaciones) que actualmente son ampliamente utilizados: CIE, RGB, HSL/V y CMYK.

La primera familia son los **modelos CIE**[1]: CIEXYZ, CIELAB y CIELUV. La CIE propuso el **modelo CIEXYZ** en 1931 (Burdick, 1997) como estándar de medida del color. El modelo CIE

[1]Commission Internationale de l'Eclairage

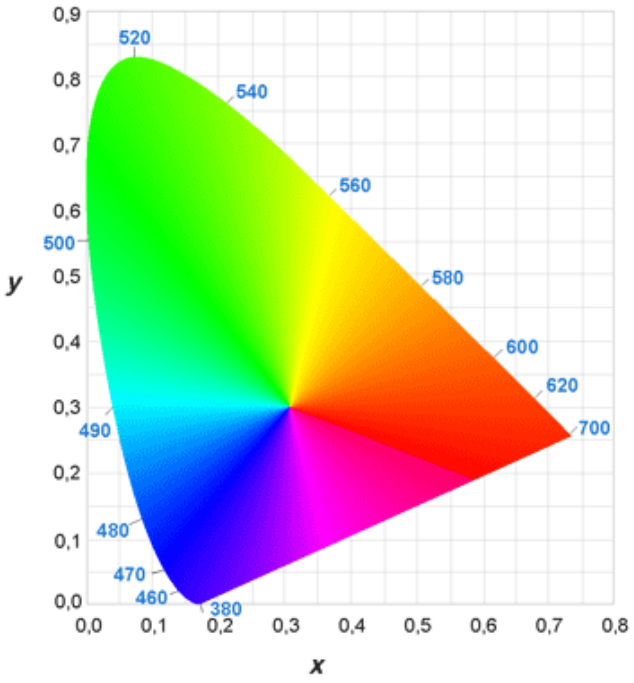

Figura 2.1: Representación del modelo CIEXYZ en el plano XY.

derivó de una serie de experimentos realizados por David W. Wright (Wright, 1928) y John Guild (Guild, 1931). Estos experimentos definieron con precisión los tres colores primarios de la síntesis aditiva de color, a partir de los cuales pueden crearse todos los demás. En este modelo, Y significa luminosidad, Z es aproximadamente igual al estímulo azul (conos S), y X es una mezcla de las curvas de sensibilidad del rojo y del verde (conos L y M). De esta manera, XYZ suele confundirse con las respuestas de los conos en RGB. Sin embargo, en el espacio de color CIEXYZ, los valores de triple estímulo no equivalen a las respuestas S, M y L del ojo humano, incluso teniendo en cuenta que X y Z son aproximadamente iguales al rojo y al azul; pero realmente, deben verse como parámetros 'derivados' de los colores rojo, verde y azul. La representación del modelo CIEXYZ en el plano XY puede verse en la figura 2.1.

La siguiente aportación en la familia CIE es el **modelo CIELAB** en 1976 (Burdick, 1997). CIELAB se plantea con el objetivo de obtener un espacio perceptual más uniforme que CIEXYZ. El propósito es introducir un espacio de color que sea más "perceptivamente lineal"que otros espacios de color. Perceptivamente lineal significa que un cambio del color en una cierta cantidad produce un cambio casi de la misma importancia visual. Los tres parámetros del modelo representan la luminosidad de color (L, L=0 negro y L=100 blanco), su posición entre rojo y verde (valores negativos indican verde mientras valores positivos indican rojo) y su posición entre amari-

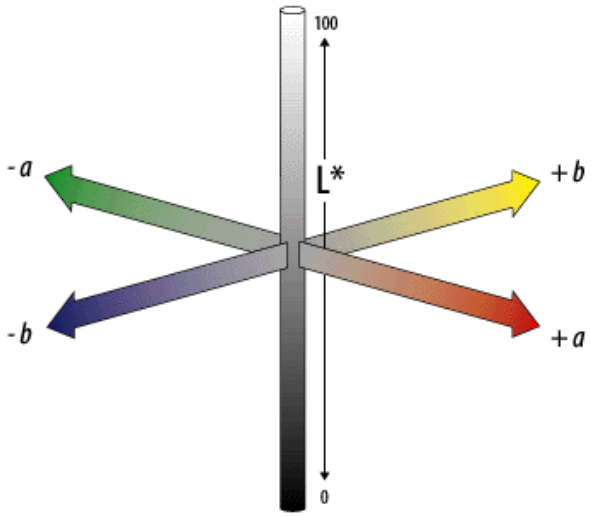

Figura 2.2: Representación del modelo CIELAB.

llo y azul (valores negativos indican azul y valores positivos indican amarillo). El modelo de color CIELAB fue creado para servir como referencia. Por lo tanto, las representaciones visuales de la gama de colores en este modelo nunca son exactas. El modelo de color CIELAB es tridimensional y solo puede ser representado adecuadamente en un espacio de tres dimensiones (figura 2.2).

Este espacio de color fue modificado en los años 1994 y 2000. En ambos casos, la modificación esta relacionada con la redefinición de la diferencia de color en el espacio CIELAB. La modificación de 1994, denominada CIE94, amplía el modelo CIELAB para tratar las no uniformidades perceptuales, al tiempo que se conservaba la misma definición del espacio mediante la introducción de pesos específicos derivados de pruebas para determinar la tolerancia del color. Por su parte, la modificación del año 2000, denotada CIE2000, fue desarrollada para solucionar el problema de las discrepancias de evaluación en la medición del color entre un colorímetro y el ojo humano. Se define una nueva fórmula mediante la cual, la diferencia calculada por mediciones de color se acerca al umbral de discriminación de color del ojo humano sobre el espacio de color sólido del CIELAB.

Adoptado al mismo tiempo que el CIELAB se encuentra el **modelo CIELUV** (Burdick, 1997). Esta doble definición se debió a la falta de un consenso claro sobre qué espacio de color representaba mejor la percepción humana. CIELUV utiliza la adaptación de punto blanco de Judd, en contraste con la transformación incorrecta de Von Kries usada por CIELAB. Esto produce resultados muy exactos cuando se trabaja con colores con un único punto de iluminación, pe-

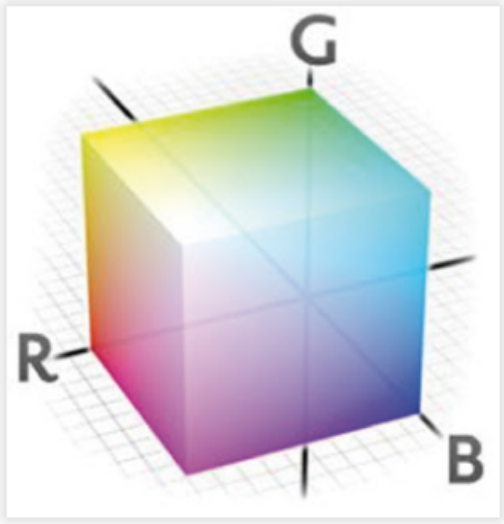

Figura 2.3: Representación del modelo RGB.

ro en contrapartida, produce colores imaginarios (es decir, fuera de la percepción humana). Este problema ha conllevado la relegación de este espacio de color en la actualidad.

Fuera de la familia CIE, debemos destacar inicialmente el **espacio RGB**[2]. A mediados del siglo XIX, Thomas Young y Hermann Helmholtz (Young, 1802) propusieron una teoría de visión tricromática del color que se convirtió en la base para el modelo de color RGB (rojo, verde, azul). Este es un modelo de color aditivo tridimensional (figura 2.3), en el cual las tres luces de colores se suman para producir diferentes tonalidades. La intensidad de luz determina el color percibido. Sin intensidad, cada uno de los tres colores se percibe como negro, mientras que la intensidad máxima lleva a la percepción del blanco. Las diferentes intensidades producen el matiz del color, mientras que la diferencia entre la mayor y menor intensidad define cómo de saturado está ese color. Adicionalmente, es un espacio perceptualmente no-lineal, es decir, las distancias perceptuales entre dos colores definidos en el espacio RGB no son linealmente equiparables en este espacio de color, lo que hace difícil discriminar la diferencia cromática. Aun así, es uno de los espacios de color aditivo más utilizado incluso con variantes como sRGB y Adobe RGB, desarrollados respectivamente por Hewlett-Packard/Microsoft Corporation y Adobe Systems.

En contraposición al modelo aditivo de color propuesto en RGB, se propuso el **modelo CMYK**[3]. CMYK es un modelo de color sustractivo que se utiliza principalmente en impresión.

[2]Red, Green, Blue

[3]Cyan, Magenta, Yellow y Key

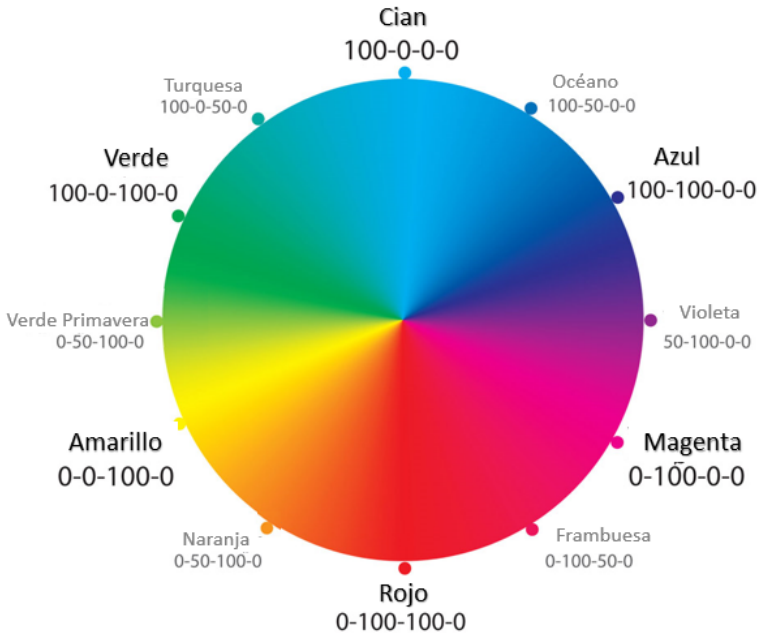

Figura 2.4: Representación del modelo CMYK.

Es la versión moderna y más precisa del antiguo modelo tradicional de coloración (RYB), que se utiliza aún en pintura y artes plásticas. Permite representar una gama de colores más amplia que este último, y tiene una mejor adaptación a los medios industriales. Este modelo se basa en la mezcla de pigmentos y sus capacidades de absorción de luz. El color que presenta un objeto corresponde a la parte de luz que incide sobre éste y que no es absorbida por el objeto. En este modelo, el cían es el opuesto al rojo, lo que significa que actúa como un filtro que absorbe dicho color (-R +G +B), el magenta es el opuesto al verde (+R -G +B) y el amarillo el opuesto al azul (+R +G -B), mientras que Key o negro es la representación de la absorción absoluta de la luz incidente. La figura 2.4 es una ilustración de la rueda cromática sustractiva del modelo CYMK.

Finalmente, se pasa a describir los **modelos HSL**[4] **y HSV**[5]. HSL y HSV son las dos representaciones de coordenadas cilíndricas más comunes de los puntos del modelo RGB. Las dos representaciones reordenan la geometría de RGB en un intento de ser más intuitivas y perceptualmente relevantes que la representación cartesiana clásica (figura 2.3).

HSV fue propuesto en 1978 por Alvy Ray Smith (Smith, 1978) y es un espacio tridimensional compuesto por el matiz (H), la saturación (S) y el valor o brillo (V). El matiz se representa como un

[4]Hue, Saturation, Lightness

[5]Hue, Saturation, Value

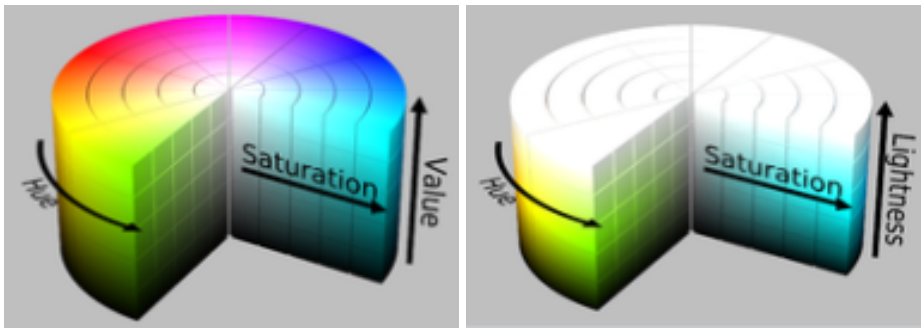

Figura 2.5: Representación de los modelos HSV (izquierda) y HSL (derecha).

ángulo cuyos valores posibles van de 0° a 360° (aunque para algunas aplicaciones se normalizan del 0 al 100). Cada valor corresponde a un color. Por ejemplo, 0 es rojo, 60 es amarillo y 120 es verde. Por su parte, el valor o brillo representa la altura en el eje blanco-negro. Los valores posibles van del 0 al 100, donde 0 siempre es negro, mientras que 100 podría ser blanco o un color más o menos saturado, dependiendo de la saturación del mismo. Finalmente, la saturación se representa como la distancia al eje de brillo negro-blanco. Los valores posibles van del 0 al 100. A este parámetro también se le suele llamar "pureza"por analogía con la pureza de excitación y la pureza colorimétrica. Cuanto menor sea la saturación de un color, mayor tonalidad grisácea habrá y más decolorado estará.

Por su parte, HSL (Burdick, 1997) también es un espacio tridimensional compuesto por el matiz (H), la saturación (S) y la luminosidad o tono (L). El matiz tiene la misma representación y significado que en el modelo HSV. La saturación difiere drásticamente del modelo HSV. En este caso, la saturación también representa la distancia al eje central del cilindro, con valores de 0 a 100, pero cuanto mayor sea la saturación de un color más puro será, mientras que valores cercanos a 0 indican mayor tonalidad grisácea, siendo 0 el gris equivalente de ese color. Finalmente, el tono o luminosidad representa un eje cromático desde el negro (valores cercanos al 0) al blanco (valores cercanos al 100). Las diferencias entre HSV y HSL son claramente visibles en la figura 2.5, donde se ve la representación cilíndrica de ambos modelos.

Ambas representaciones se usan ampliamente en el procesamiento de imágenes, ya que a menudo es más conveniente que RGB, pero ambos espacios también son criticados por no separar adecuadamente los atributos de color o por su falta de uniformidad perceptual.

2.2. Modelos de atención visual

Cada segundo entra en el ojo un enorme flujo de datos visuales. Por lo tanto, el procesamiento inmediato de todos estos datos, sin ningún mecanismo para reducir esta ingente cantidad de información, sería inviable. La respuesta de la evolución humana fue el desarrollo de la atención, es decir, la capacidad del ojo humano (mediante la retina) de difuminar la mayor parte de la escena visual sin perder información de las zonas realmente importantes.

En las últimas décadas, muchos ámbitos de la ciencia han intentado conocer cómo funciona la atención visual humana. Los psicólogos han estudiado la correlación conductual de la atención visual como la ceguera producida ante los cambios, por la falta de atención y por el enfoque de la propia atención. Por su parte, los fisiólogos neuronales han mostrado cómo las neuronas se ajustan para enfocar mejor los objetos. Los expertos en neurociencia han hecho un modelo de redes neuronales reales que simula y explica modelos de conducta de la atención. Inspirados en estos estudios, la Informática y la Robótica tratan de crear modelos computacionales para simular este mecanismo.

Existen dos aproximaciones muy diferentes respecto a cómo la comunidad científica ha abordado la simulación de la atención humana. Dentro de la literatura, encontramos los modelos dirigidos por la escena, denominados modelos *Bottom-Up* (BU), y los modelos dirigidos por los objetivos o modelos *Top-Down* (TD).

La aproximación TD se refiere a la asignación voluntaria de atención a ciertas características, objetos o regiones en el espacio. Por ejemplo, un sujeto puede decidir prestar especial atención a una pequeña región en la esquina superior izquierda o sobre todos los elementos rojos de una escena. En sí, la atención estaría dirigida por un objetivo u objetivos predefinidos. Por su lado, los modelos BU defienden que la atención no se dirige voluntariamente. Los estímulos destacados atraen la atención, aunque el sujeto no tenga ninguna intención de atender a estos estímulos. Por ejemplo, si una persona está involucrada en una conversación, pero se produce un destello o brillo, este puede atraer la atención del sujeto de forma involuntaria. A esta capacidad de sobresalir y de atraer la atención de un elemento o área de la imagen, respecto a su entorno, se la denomina saliencia.

Aunque los modelos TD pueden ser muy interesantes, como los experimentos llevados a cabo por Yarbus (Yarbus, 1967), que demostraron que el movimiento ocular de los sujetos de su estudio dependía de la tarea que estaban ejecutando, así como de la experiencia en realizarla, suelen ser

difícilmente extrapolables a modelos computacionales, por ser muy dependientes del entorno y/o tareas concretas a simular. Por esta razón, los modelos BU son los que más ampliamente se han modelado y aplicado en Visión por Computador.

Dentro de los modelos BU debemos destacar las aproximaciones de Kotch (Koch and Ullman, 1985), Itti (Itti et al., 1998) y Frintrop (Frintrop, 2006). Koch (Koch and Ullman, 1985) fue el primero en proponer una aproximación basada en la saliencia visual a través del comportamiento de los grupos neuronales y de la corteza visual del Núcleo Geniculado Lateral (NGL). Koch especificó su modelo sobre la base de tres premisas:

1. La existencia de una serie de características elementales, como el color, la orientación, la dirección del movimiento, la disparidad, etc., que se pueden representar en paralelo y en diferentes mapas topográficos, llamados representación temprana.

2. Se puede definir una cartografía selectiva a partir de la representación topográfica temprana en una representación no topográfica más central, de manera que, en cualquier instante, la representación central contiene las propiedades de una sola ubicación en la escena visual: la ubicación seleccionada. Este mapeo es la principal expresión de la atención visual selectiva temprana. Así, la función de la atención selectiva es fusionar la información de diferentes mapas en un todo coherente.

3. Se pueden describir ciertas reglas de selección que determinan qué ubicaciones se asignarán a la representación central. La regla principal, usando la saliencia de las ubicaciones en la representación temprana, se implementa mediante una red llamada Winner-Take-All (WTA). En esta red WTA se inhibe la ubicación seleccionada, lo que provoca un cambio automático hacia la siguiente ubicación más destacada. Adicionalmente, existen dos reglas que describen la preferencia del sistema visual humano sobre los elementos próximos (proximidad) y similares (similitud).

Como vemos, Koch planteó la existencia de mapas de topografía o de saliencia asociados a ciertos elementos que atraen la atención de forma individual, pero que deben combinarse para formar una representación de qué partes de la escena visual aglutinan una mayor carga de atención o saliencia. Adicionalmente, se describió el mecanismo (red WTA, proximidad y similitud) que permite definir por qué los seres humanos cambiamos nuestro foco de atención.

Estos estudios fueron ampliados con la colaboración de Itti y Niebur (Itti et al., 1998) en un

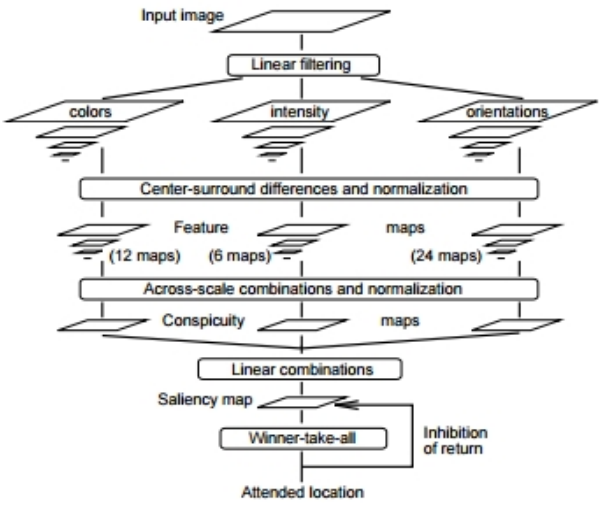

Figura 2.6: Representación gráfica del modelo de IKN.

modelo computacional (modelo IKN) que llevaba el modelo teórico de Koch al ámbito de aplicación del procesamiento de imágenes, como se resume en la figura 2.6. En este modelo, un punto de imagen se considera saliente o destacable cuando su color, brillo u orientación se diferencia significativamente de su vecindario. Esto se denomina antagonismo de centro-envolvente. Para calcular este antagonismo se genera una pirámide gaussiana de 9 niveles para la imagen dada. Los diferentes niveles de esta pirámide se obtienen al submuestrear la imagen del nivel anterior tras aplicar un filtro gaussiano. De esta forma, la pirámide recoge la misma imagen a diferentes niveles de detalle, desde los más finos (nivel 0) a los más bastos (nivel 8). Mediante una operación de sustracción entre los niveles altos y bajos de la pirámide, se caracteriza cómo de diferente es un píxel de su entorno respecto a una característica dada, y por lo tanto, su saliencia temprana.

Posteriormente, mediante procesos de aglutinación de estos mapas de saliencia, se obtendrá el mapa de saliencia global, preparado para indicar qué zonas de la imagen atraen la atención visual con mayor intensidad. Sobre este mapa de saliencia global se aplicará la aproximación WTA anteriormente definida por Koch, que permitirá simular cómo los seres humanos van fijándose en partes alternativas de la imagen, así como el orden en el que lo realizan. La explicación de este modelo se realizará en profundidad en el capítulo 5, ya que su arquitectura es la base del modelo de atención visual propuesto en esta tesis.

Basado en IKN, se propuso una alternativa híbrida, VOCUS[6] (Frintrop, 2006). Al igual que

[6]Visual Object detection with a CompUtational attention System

IKN, VOCUS también se basa en el modelo de atención, mapas de saliencia, creación de la red WAT, así como en la inhibición del retorno. Sin embargo, VOCUS realiza una serie de mejoras sobre IKN, al introducir un procesamiento de saliencia paralelo. Este nuevo mapa de saliencia conjunta no se basa en las aproximación BU, sino que se modela según una estrategia TU. Este mapa de saliencia representa la intención del observador de detectar objetos concretos en la escena visual. Esta aproximación permitió una mayor precisión en la identificación del foco de atención en una imagen dada. La figura 2.7 recoge la arquitectura híbrida de VOCUS.

Otro ejemplo de arquitectura híbrida es la aproximación de Judd, Ehinger, Durand y Torralba (Judd et al., 2009). Estos autores detectaron que la mayoría de algoritmos de atención visual no explotaban los movimientos oculares reales de los seres humanos. Para abordar este problema, se definió un nuevo modelo para el cálculo de la saliencia sobre características en tres niveles de detalle: bajo, medio y alto. La caracterización de bajo nivel considera la intensidad, el color y la orientación presente en la imagen, empleando el modelo IKN para definir el mapa de saliencia resultante. La caracterización de medio nivel se basa en la detección del horizonte, debido a que la mayoría de los objetos descansan sobre la superficie de la Tierra, lo que convierte al horizonte en un lugar donde los seres humanos buscan naturalmente objetos salientes. Esta caracterización se obtiene mediante un algoritmo de detección del horizonte. Por último, la caracterización de alto nivel explota la fijación natural del ser humano en las personas y sus rostros, la cual se implementa mediante el algoritmo de detección de caras de Viola y Jones (Viola and Jones, 2004) y el detector de personas de Felzenszwalb (Felzenszwalb et al., 2008). Estos mapas de saliencia (un total de 33) se normalizan y ponderan para obtener el mapa de saliencia final del modelo.

También caben destacar una variación más reciente del modelo IKN, la aproximación de Won, Lee y Son (Won et al., 2008), la cual modifica la caracterización de bordes propuesta por Itti. En este caso, se realiza la extracción de los bordes más sobresalientes mediante el operador de Sobel, tanto en orientación vertical como horizontal. Adicionalmente, la imagen donde se aplica este operador no es la de escala de grises propuesta por IKN, sino que se define una nueva imagen basada en las caracterización de la oposición cromática (RG y BY).

Fuera del modelo IKN pero dentro de los modelos BU, podemos destacar las aproximaciones de Hou y Zhang (Hou and Zhang, 2007) y de Maruta, Isshi y Sato (Maruta et al., 2010). En 2007, Hou y Zhang propusieron un modelo independiente de la caracterización, clasificación, así como de cualquier conocimiento previo de los objetos, mediante la generación de un mapa de saliencia en el dominio espectral. Este saliencia es calculada usando el residuo espectral de la

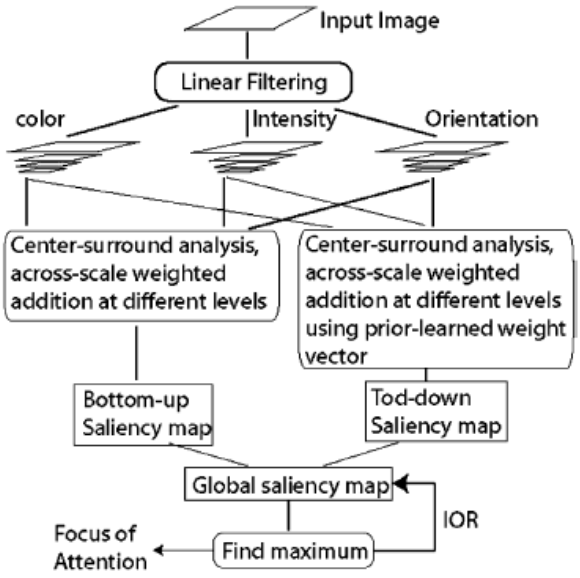

Figura 2.7: Representación gráfica del modelo VOCUS.

imagen obtenida del análisis del espectro logarítmico de la misma. Por su parte, Maruta, Isshi y Sato se basan en el modelado periférico de la visión humana. Este modelado es la clave para crear un nuevo mapa de saliencia donde las regiones salientes se obtienen por el comportamiento de la distribución de los extremos locales, ya que estos extremos detectan regiones de la imagen que son más brillantes o más oscuras que su entorno.

Finalmente, Avraham y Lindenbaum (Avraham and Lindenbaum, 2010) proponen un nuevo mecanismo de atención visual dentro del enfoque BU, pero mediante un modelo estocástico validado que estima la probabilidad de que una parte de la imagen sea de interés, es decir, la saliencia. Este modelo cuantifica varias observaciones intuitivas, tales como la mayor probabilidad de correspondencia entre regiones de la imagen visualmente similares o la probabilidad de que sólo unos pocos objetos de interés estén presentes en la escena. Esta aproximación comienza con una segmentación muy burda, para posteriormente explotar modelos gráficos que la vayan aproximando a los segmentos de la imagen más salientes.

2.3. Seguimiento de objetos en secuencias de vídeo

Con seguimiento de objetos en vídeo nos referimos a la detección, extracción, reconocimiento y seguimiento de objetos en movimiento en las secuencias de imágenes que componen un vídeo,

con el fin de obtener parámetros precisos de su movimiento (tales como posición, velocidad, etc.). El seguimiento de objetos en secuencias de vídeo es una tarea importante en el campo de la Visión por Computador. Es una tecnología fronteriza interdisciplinar, ya que se combina el procesamiento de imágenes, el reconocimiento de patrones, la inteligencia artificial, el control automático, así como otras áreas dentro de la representación del conocimiento. El seguimiento de objetos en vídeo tiene una amplia perspectiva de aplicación, tal como la vídeo-vigilancia, la interacción humano-ordenador, la gestión del tráfico inteligente y la navegación en Robótica, entre otras muchas.

Dentro de este amplio campo podemos identificar diferentes familias de soluciones, dependiendo de la aproximación que se haga a este problema. La primera familia que cabe destacar es el **seguimiento basado en concordancia**. Los algoritmos de seguimiento basados en concordancia establecen un modelo para los objetos presentes en la secuencia de vídeo basado en sus características, para realizar la asignación de los objetos entre imágenes mediante la maximización de la relación de concordancia entre estas características. En (Cannons and Wildes, 2014; He et al., 2015), la caracterización se realiza basándose en regiones, donde la plantilla del objeto es la región del objeto inicial. Esta región se busca en las siguientes imágenes por maximización de la concordancia, calculada por medio de la suma de diferencias al cuadrado (SSD). En estas aproximaciones se usa la información global de los objetos, como las texturas o sus niveles de gris, en múltiples vistas de los mismos.

Una evolución lógica de esta aproximación es basarse en plantillas deformables y no en regiones prefijadas. En estas aproximaciones, se utiliza una superficie o curva que tiene unas propiedades de deformación y elasticidad de sus contornos, lo que permite su deformación para asemejarse a los cambios sufridos por los objetos que se están siguiendo. En (Yilmaz et al., 2004) se propone un algoritmo de seguimiento de contornos que utiliza la estimación de la densidad del *kernel* y el modelo de ondícula (*wavelet*) de Gabor para guiar la evolución del contorno a través de la caracterización del color y la textura, respectivamente. Por su lado, en (Ning et al., 2013) se presenta un nuevo marco de evolución basado en la formación de conjuntos y la segmentación mediante contornos activos (JRACS[7]) que le da una gran robustez frente a objetos no rígidos.

Otro camino para determinar la concordancia es mediante la descripción de modelos. En estos casos, se establece un modelo geométrico del objeto de acuerdo con el conocimiento a priori

[7]Joint Registration and Active Contour Segmentation

que se tenga de ellos, usando estos modelos para su comparación y asignación. Estos modelos pueden ser muy variados, siendo los más normales los jerárquicos (Karaulova et al., 2002), los bidimensionales (S. et al., 2004) o los tridimensionales (Yang et al., 2001). Por otra parte, los parámetros de caracterización más utilizados son la pose y movimiento de los objetos. Por ejemplo, en (Yang et al., 2001) se presenta un nuevo algoritmo de localización de vehículos basado en un modelado tridimensional de los mismos. El modelo utiliza puntos de contorno y se compara con un modelo proyectado del vehículo para determinar su nueva posición.

Finalmente, algoritmos como el de Kanade, Lucas y Tomasi (Lucas and Kanade, 1981; Tomasi and Kanade, 1991), normalmente conocido como KLT, modela la imagen en su conjunto para su posterior registro mediante algunos puntos de control. KLT se basa en el cálculo de mínimos locales de la suma de las diferencias al cuadrado. Es una técnica que utiliza el gradiente espacial de intensidad para dirigir la búsqueda de los puntos homólogos, por lo que es capaz de encontrar la mejor posición para los puntos de control en un número reducido de operaciones, y por lo tanto, de indicar el movimiento de los objetos donde están estos puntos de control.

La otra gran familia de algoritmos de seguimiento son los **algoritmos basados en filtrado**. Estos algoritmos consideran el problema del seguimiento como un problema de estimación del estado de los objetos en las siguientes imágenes. El estado de un objeto puede ser cualquier cosa, desde su apariencia a su movimiento, pero las aproximaciones que intentan caracterizar el movimiento y sus posibles evoluciones son las más comunes. En sí, la clave del seguimiento en estas aproximaciones es inferir la densidad de probabilidad a posteriori del estado del objeto respecto a los datos observados. Dentro de esta familia podemos destacar el filtro de Kalman y el filtro de partículas.

El filtro de Kalman (KF[8]) (Kalman, 1960) es un método eficaz para estimar los estados del objeto mediante la definición de dos modelos: uno que predice los estados (modelo de estado) y otro que estima la función de densidad a posteriori (modelo de observación). Para el seguimiento de los objetos, el filtro de Kalman usualmente usa el ruido gaussiano para representar la incertidumbre de los modelos de estado y observación. Esta incertidumbre es usada para equilibrar automáticamente el efecto de las diferentes observaciones del objeto, así como de las predicciones resultantes del algoritmo. Este modelado solo se utiliza cuando los cambios en el estado del objeto son suaves. Con el fin de resolver esta limitación, se plantea el filtro de Kalman extendido

[8]Kalman Filter

(EKF[9]) (Wen et al., 2015). EKF proporciona una linealización para sistemas no-lineales para posteriormente aplicar la aproximación KF estándar. Esta aproximación permite que el filtro pueda acoplarse a cambios muy fuertes y/o repentinos, ya que elimina los múltiples picos en la función de densidad que se producirían si se aplicara directamente KF en sistemas no lineales.

Por su parte, el filtro de partículas (PF[10]) (Gordon et al., 1993; Choi et al., 2013; Kumar and Dick, 2013) es un filtro de Monte Carlo secuencial, que se utiliza para resolver el problema de estimación bayesiana bajo condiciones no lineales y no gaussianas. El principio fundamental de PF es evaluar la función de densidad de probabilidad del estado mediante un conjunto de partículas de muestreo ponderadas, utilizando el método de Monte Carlo para simular la propagación de la densidad de probabilidad buscada. Si la cantidad de muestras es lo suficientemente grande, el muestreo puede ser una buena aproximación a la función de densidad de probabilidad. La media, la covarianza, así como otros estadísticos pueden ser fácilmente calculados en función de los puntos de muestreo. Gran cantidad de estudios muestran que el seguimiento mediante filtros de partículas es más apropiado que el uso de filtros de Kalman en entornos complejos. Sin embargo, PF presenta problemas debidos a la degeneración de las partículas y las oclusiones.

Finalmente, cabe destacar un conjunto de **aproximaciones por fusión**. Estas aproximaciones combinan una variedad de algoritmos de seguimiento, así como de fuentes de información, para mejorar la precisión del seguimiento. Inicialmente, la fusión de múltiples características suele ser lo más común. Normalmente, no se puede obtener un seguimiento estable si los objetos solo se caracterizan mediante una única fuente de información. Por lo tanto, muchos estudios usan tecnología de fusión de información de múltiples características para mejorar el rendimiento de seguimiento de objetos en secuencias de vídeo. Por ejemplo, (Zhou and Aggarwal, 2006) presenta un método que integra la posición espacial, la forma y el color para mejorar el rendimiento de EKF.

Otras aproximaciones se basan en la fusión de múltiples modelos en un meta-modelo. Estos métodos integran los modelos de los objetos en diferentes instantes temporales o combinan modelos de objetos en diversas vistas para conformar este meta-modelo. Esto mejora la robustez del seguimiento, ya que permite una mayor adaptación a los cambios del objeto. Un ejemplo puede encontrarse en (Khan and Shah, 2009), que presenta un enfoque multivista para resolver los

[9]Extended Kalman Filter

[10]Particle Filter

problemas de seguimiento en escenas concurridas.

Por último, se puede plantear la fusión de diferentes algoritmos de seguimiento. Cada algoritmo de seguimiento tiene sus propias ventajas para ciertas escenas. La fusión de los mismos intenta fortalecer el resultado final al superar las debilidades de cada algoritmo individual. En (Shan et al., 2007) se propone un algoritmo que fusiona Mean Shift (Comaniciu and Meer, 2002) con el filtro de partículas (MSEPF[11]) que mejora considerablemente la eficiencia del muestreo mediante la incorporación de Mean Shift en el filtro de partículas.

Otro ejemplo de estas aproximaciones se propone en (Kwon and Lee, 2011). Kwon y Lee definen un nuevo marco de seguimiento conocido como VLT. Este nuevo marco sigue un objeto de forma robusta mediante la determinación de los algoritmos de seguimiento más adecuados en cada momento. Dado que el entorno del objeto puede variar drásticamente con el tiempo, este método aglutina diferentes formas de representar el estado del objeto (modelos de apariencia, modelos de movimiento, etc.) y ejecuta en paralelo diferente algoritmos de seguimiento para cubrir diversas variaciones de manera eficiente, obteniendo de esta forma un seguimiento óptimo mediante alguno de los pares algoritmo-representación ejecutados.

[11]Mean Shift Embedded Particle Filter

Parte I

MODELO PERCEPTUAL DE IZMAILOV Y SOKOLOV

Capítulo 3

Modelo perceptual de Izmailov y Sokolov

El modelo teórico perceptual propuesto por Izmailov y Sokolov en (Izmailov and Sokolov, 1991) define un espacio métrico de color donde cada punto representa un color específico y las distancias euclidianas entre ellos son proporcionales a las diferencias de color percibidas entre estos colores por las personas. Este modelo se derivó a través de experimentos psicofísicos con personas y técnicas de análisis de escalado multidimensional. En realidad, se llevaron a cabo tres series de experimentos como se describe a continuación.

En el primer experimento, se proyectaron simultáneamente pares de colores, uno ocupando el centro de la imagen, denominado *estímulo*, y el otro ocupando el resto de la imagen, denominado *fondo*. Tanto la luminancia como la longitud de onda del fondo se mantuvieron constantes considerando una luz neutra. En cambio, la longitud de onda del estímulo se varió en 16 valores discretos de 425 nm a 675 nm, más el color blanco (es decir, un total de 17 colores), manteniendo un brillo constante para todos los colores. Se pidió a los sujetos del estudio que indicaran la diferencia de color percibida entre el estímulo y el fondo, en una escala de 0 (mismo color) a 9 (máxima diferencia cromática).

Una implementación del algoritmo de Sherpard-Kruscal para el análisis de escala multidimensional se aplicó a las diferencias de color subjetivas obtenidas del estudio anterior. Como resultado, se dedujo un espacio semiesférico tridimensional con ejes X_1, X_2 y X_3. Las figuras 3.1 y 3.2 muestran la asignación de los 17 colores en igualdad de brillo a los subespacios X_1X_2 y X_1X_3, respectivamente.

La conclusión principal de estos resultados es que la diferencia cromática perceptual ($\triangle C_{ij}$) entre dos colores con brillo equivalente $({}^iX_1, {}^iX_2, {}^iX_3)^T$ y $({}^jX_1, {}^jX_2, {}^jX_3)^T$ en este espacio se puede

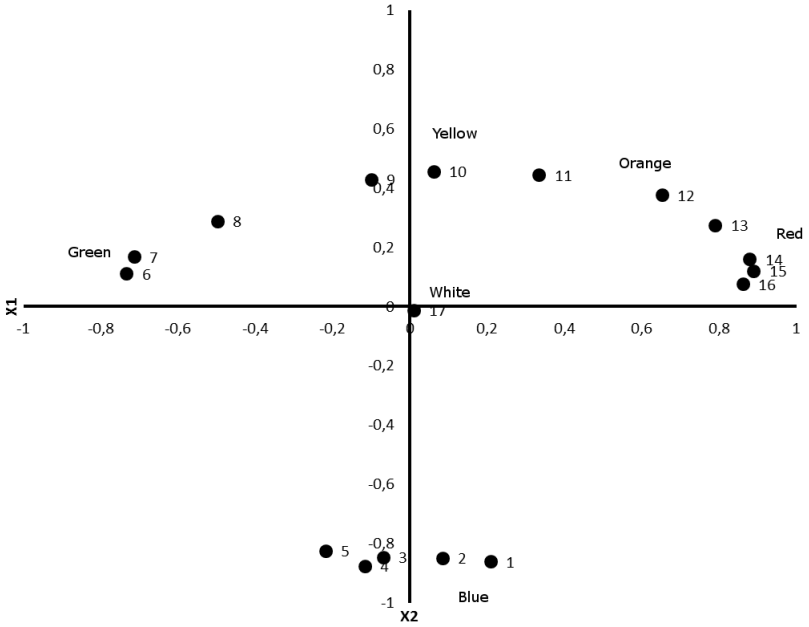

Figura 3.1: Proyección de los 17 colores (en igualdad de brillo) en el plano X_1X_2.

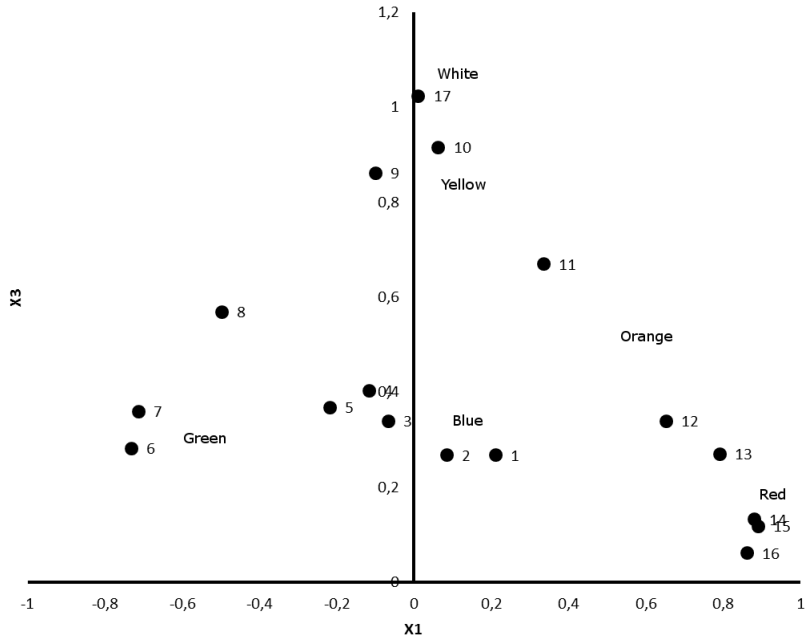

Figura 3.2: Proyección de los 17 colores (en igualdad de brillo) en el plano X_1X_3.

estimar por medio de la distancia euclidiana entre estos dos puntos, donde $\triangle X_\chi = {}^i X_\chi - {}^j X_\chi$:

$$(\triangle C_{ij})^2 = (\triangle X_1)^2 + (\triangle X_2)^2 + (\triangle X_3)^2. \tag{3.1}$$

En el segundo experimento, se proyectaron dos niveles de luminancia acromática para el estímulo y el fondo, respectivamente. Se combinaron siete niveles de luminancia del estímulo (0.2, 1, 2, 10, 20, 100 y 200 cd/m^2) con tres niveles de luminancia del fondo (1, 10, 100 cd/m^2). Los sujetos del experimento tenían que indicar la diferencia en el brillo percibido entre el estímulo y el fondo en una escala de 0 (mismo brillo) a 9 (máxima diferencia de brillo).

El algoritmo de escalado multidimensional de Shepard-Kruscal también se aplicó a las diferencias subjetivas del brillo, dando lugar a un espacio bidimensional descrito como una semicircunferencia con ejes Y_1 e Y_2. Como en el experimento anterior, la diferencia acromática perceptual ($\triangle W_{ij}$) entre dos niveles de luminancia $({}^i Y_1, {}^i Y_2)^T$ y $({}^j Y_1, {}^j Y_2)^T$ puede estimarse a través de la distancia euclidiana entre ambos puntos, donde $\triangle Y_\chi = {}^i Y_\chi - {}^j Y_\chi$:

$$(\triangle W_{ij})^2 = (\triangle Y_1)^2 + (\triangle Y_2)^2. \tag{3.2}$$

Finalmente, en el tercer experimento, la longitud de onda del estímulo se varió en 25 valores discretos entre 425 nm a 675 nm más el color blanco (es decir, 26 colores), mientras que el fondo se mantuvo oscuro. A su vez, se consideraron seis niveles de luminancia para el estímulo de 0.2 cd/m^2 a 200 cd/m^2. Se pidió a cada sujeto del experimento que describiera el color percibido del estímulo, correspondiente a un destello de luz monocromática, con una combinación específica de longitud de onda y luminancia usando uno, dos o tres colores básicos entre cinco posibles alternativas: rojo, amarillo, verde, azul y blanco. Se generó un vector pentadimensional con las respuestas de cada sujeto y combinación. Estos vectores, obtenidos a partir de las respuestas de todos los sujetos para una misma combinación de color y luminosidad, se aglutinaron, produciendo un único vector en este espacio pentadimensional definido. El algoritmo de Shepard-Kruscal fue aplicado de nuevo, dando lugar a un espacio de color y luminancia hiperesférico de cuatro dimensiones, con coordenadas Z_1, Z_2, Z_3 y Z_4.

Como en los experimentos previos, la diferencia perceptual ($\triangle S_{ij}$) entre dos puntos de color/luminancia $({}^i Z_1, {}^i Z_2, {}^i Z_3, {}^i Z_4,)^T$ y $({}^j Z_1, {}^j Z_2, {}^j Z_3, {}^j Z_4,)^T$ puede ser estimada a través de la dis-

tancia euclidiana entre ambos puntos:

$$(\triangle S_{ij})^2 = (\triangle Z_1)^2 + (\triangle Z_2)^2 + (\triangle Z_3)^2 + (\triangle Z_4)^2, \tag{3.3}$$

donde $\triangle Z_\chi = {}^i Z_\chi - {}^j Z_\chi$.

Al analizar las proyecciones de todos los puntos cuatridimensionales en cada una de sus dimensiones, Izmailov y Sokolov establecieron (Izmailov and Sokolov, 1991) que existía una relación entre estas cuatro dimensiones y las obtenidas previamente en los dos experimentos anteriores:

$$Z_1 = X_1, \quad Z_2 = X_2, \quad Z_3 = Y_2 X_3, \quad Z_4 = Y_1 X_3, \tag{3.4}$$

siendo X_1, X_2 y X_3 las coordenadas del espacio cromático derivadas del primer experimento, e Y_1 e Y_2 las coordenadas del espacio acromático derivadas del segundo experimento. Después, tras manipular las expresión en (3.3), tomando en cuenta las equivalencias en (3.1), (3.2) y (3.4), se obtuvo (Izmailov and Sokolov, 1991) la siguiente expresión para estimar la diferencia perceptual entre dos colores:

$$(\triangle S_{ij})^2 = (\triangle C_{ij})^2 + {}^i X_3 {}^j X_3 (\triangle W_{ij})^2, \tag{3.5}$$

donde $\triangle C_{ij}$ es la diferencia cromática definida en (3.1), $\triangle W_{ij}$ la diferencia acromática definida en (3.2) y X_3 la tercera componente del espacio cromático, que está directamente relacionada con la claridad del color, como se muestra en el capítulo 4.

Este modelado de la diferencia perceptual del color definido en (3.5) difiere de otros modelos de diferencia de color que tratan separadamente los componentes cromáticos y acromáticos. Por ejemplo, en el modelo CIELAB, la diferencia entre dos colores (L_i^*, a_i^*, b_i^*) y (L_j^*, a_j^*, b_j^*), donde L^* es la componente acromática (claridad) y a^*, b^* las componentes cromáticas, es aproximada según la formulación de CIE76 (Sharma, 2002):

$$(\triangle E_{ij})^2 = (\triangle L_{ij}^*)^2 + (\triangle a_{ij}^*)^2 + (\triangle b_{ij}^*)^2, \tag{3.6}$$

donde $\triangle L_{ij}^* = L_i^* - L_j^*$, $\triangle a_{ij}^* = a_i^* - a_j^*$ y $\triangle b_{ij}^* = b_i^* - b_j^*$.

La ecuación (3.6) se asemeja a (3.1), ya que como se muestra a continuación, X_3 está fuertemente correlada con la claridad del color, de forma similar a L^*, mientras que X_1 y X_2 son canales de oposición cromática, al igual que a^* y b^* respectivamente. Por lo tanto, el segundo término

acromático descrito en (3.5) enriquece el modelo agregando información perceptual no considerada en otros modelos de diferencia de color. En particular, $\triangle W_{ij}$ introduce información sobre la diferencias entre el brillo de una región (estímulo) y su vecindario (fondo). Esto es una novedad con respecto a los modelos de diferencias de color anteriores, que solo procesan el color de los puntos comparados, obviando su entorno. Esto hace que el modelo de Izmailov y Sokolov sea particularmente atractivo para la atención visual, donde el significado de los puntos no sólo depende de su apariencia local o individual, sino también de la apariencia general de las regiones en las que se encuentran.

El espacio de color original y el modelo de diferencia de color derivado propuestos por Izmailov y Sokolov están fuertemente ligados a los experimentos psicofísicos de los cuales se obtuvieron, implicando parámetros como longitudes de onda, luminancias y/o pares de luminancias. Esto hace que este modelo de percepción humana sea puramente teórico, y por lo tanto, no pueda aplicarse directamente a imágenes digitales. El siguiente capítulo propone un modelo computacional que permite la aplicación del modelo teórico de Izmailov y Sokolov a imágenes digitales en color.

Capítulo 4

Adaptación computacional del modelo perceptual

En este capítulo se propone un modelo computacional que relaciona el espacio de color RGB con las variables X_1, X_2, X_3, Y_1 e Y_2 que caracterizan el modelo perceptual teórico de Izmailov y Sokolov (Izmailov and Sokolov, 1991) resumido en el capítulo anterior. Esto permite el cálculo de las diferencias perceptuales de color definidas en (3.5) directamente a partir de imágenes digitales en color. Las secciones siguientes describen, respectivamente, la asignación de los subespacios cromáticos (X_1, X_2, X_3) y acromáticos (Y_1, Y_2).

4.1. Mapeo computacional del subespacio cromático

El subespacio cromático derivado por Izmailov y Sokolov en su primer experimento es un espacio euclidiano tridimensional, en donde cada longitud de onda está asociada con un vector (X_1, X_2, X_3). Con el objetivo de definir un mapeo entre el espacio de color RGB[1] y este subespacio cromático, en primer lugar, es necesario asignar cada vector RGB a su correspondiente longitud de onda aproximada. Hemos realizado este mapeo con el algoritmo propuesto por Burton (Burton, 1996). De acuerdo a esta aproximación, la equivalencia entre las 17 longitudes de onda utilizadas por Izmailov y Sokolov con sus vectores RGB queda recogida en la tabla 4.1.

La distribución de puntos en el subespacio cromático sugiere una fuerte correlación entre las dos primeras dimensiones (X_1, X_2) y los dos canales de oposición cromática RG e YB, con una co-

[1]En este trabajo, los canales R, G y B se suponen normalizados entre cero y uno, con el fin de hacer el método independiente de la profundidad del color

N	Longitud de onda (nm)	*R*	*G*	*B*
1	425 (azul)	0.165	0	1
2	440	0.08	0	1
3	450	0	0.2	1
4	460	0	0.4	1
5	466 (azul claro)	0	0.521	1
6	520 (verde)	0.141	1	0
7	525	0.215	1	0
8	554	0.627	1	0
9	570 (amarillo)	0.859	1	0
10	575	0.929	1	0
11	600 (naranja)	1	0.694	0
12	613	1	0.494	0
13	625	1	0.306	0
14	635	1	0.153	0
15	650	1	0.039	0
16	675 (rojo)	1	0	0
17	Blanco	1	1	1

Cuadro 4.1: Conversión de longitud de onda a *RGB* para los 17 colores considerados por Izmailov y Sokolov (Izmailov and Sokolov, 1991).

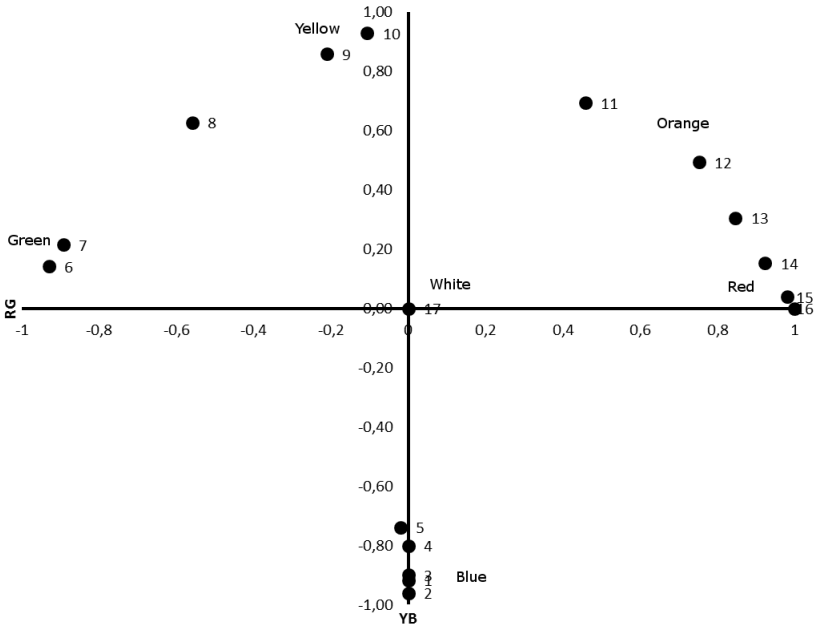

Figura 4.1: Proyección de los 17 colores (en igualdad de brillo) en el plano *RG*-vs-*BY* basado en Izmailov y Sokolov (Izmailov and Sokolov, 1991).

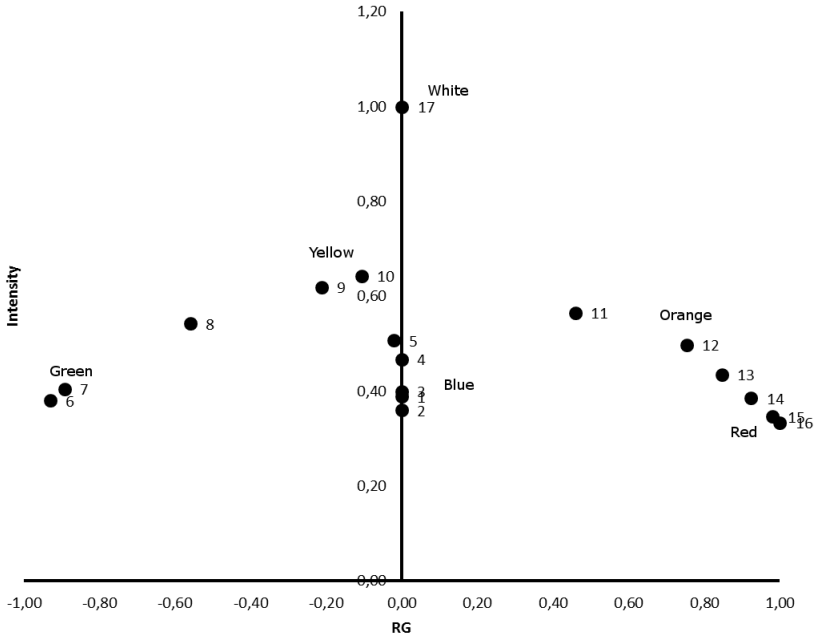

Figura 4.2: Proyección de los 17 colores (en igualdad de brillo) en el plano *RG*-vs-*I* basado en Izmailov y Sokolov (Izmailov and Sokolov, 1991).

rrespondencia con las rutas neuronales presentes en la retina y el núcleo geniculado lateral (NGL) de los primates (Engel et al., 1997). En particular, RG indica cómo de rojo (valores positivos) o verde (valores negativos) es un color, mientras que YB es su equivalente para el amarillo (valores positivos) y el azul (valores negativos). A su vez, la tercera dimensión X_3 está fuertemente correlacionada con la intensidad de color *I*, que también corresponde a una tercera ruta neuronal presente en la retina y en el NGL de los primates (Engel et al., 1997).

Un posible mapeo entre el espacio de color *RGB* y el espacio *(RG,YB,I)* se puede definir de acuerdo con la formulación propuesta en (Itti et al., 1998):

$$RG = \mathscr{R} - \mathscr{G}, \quad YB = \mathscr{Y} - \mathscr{B}, \quad I = \frac{(R+G+B)}{3}, \tag{4.1}$$

con $\mathscr{R}$, $\mathscr{G}$, $\mathscr{B}$ y $\mathscr{Y}$ definidas como:

$$\begin{aligned} \mathscr{R} &= \frac{R-(G+B)}{2}, \quad \mathscr{G} = \frac{G-(R+B)}{2}, \\ \mathscr{B} &= \frac{B-(R+G)}{2}, \quad \mathscr{Y} = \frac{(R+G)}{2} - \frac{|R-G|}{2-B}, \end{aligned} \tag{4.2}$$

que son fijadas a cero si presentan valores negativos. Las figuras 4.1 y 4.2 muestran el mapeo de los 17 colores, con igual brillo, en los subespacios *RG-vs-YB* y *RG-vs-I* definidos de acuerdo a (4.1).

Sin embargo, hay diferencias significativas entre este mapeo y el espacio perceptual derivado en (Izmailov and Sokolov, 1991), como se muestra en las figuras 4.1 y 4.2. Por ejemplo, restablecer los valores negativos en (4.2) no es coherente con el espacio perceptual anteriormente mencionado. Así, las longitudes de onda azules (del 1 al 5) no deberían tener ningún componente RG según (4.1), mientras que sí lo tienen en el espacio perceptual. Esta contradicción entre ambos espacios de color es aún más evidente con respecto a la intensidad. En particular, mientras que los rojos, verdes y azules producen una intensidad similar según (4.2), esto no ocurre en el espacio perceptual, en el cual los azules y los verdes producen intensidades perceptuales notablemente superiores a las de los rojos. Esto puede explicarse por la sensibilidad espectral de las células del tipo cono (Sharma, 2002). Por tanto, mientras que la sensibilidad espectral máxima de los conos *S* y *M* se alcanza respectivamente en las longitudes de onda azules y verdes, las longitudes de onda rojas caen en la cola derecha de la curva de sensibilidad de los conos *L*. Por lo tanto, el ojo humano es más sensible a las longitudes de onda azules-verdes puras. Además, mientras que la intensidad asociada con el amarillo en el espacio perceptual es significativamente más cercana a la intensidad del blanco que a los otros colores básicos, esto no se produce según (4.2). Esto puede atribuirse al

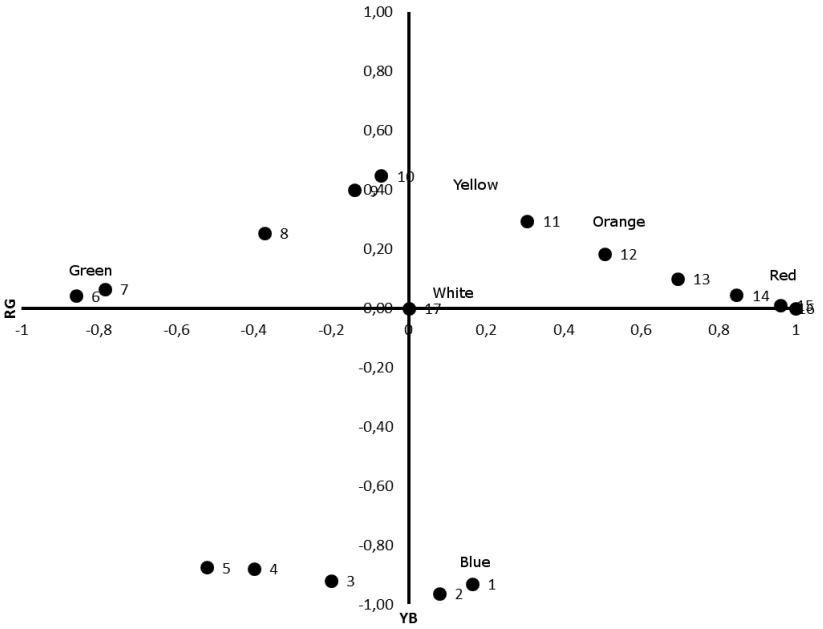

Figura 4.3: Proyección de los 17 colores (en igualdad de brillo) en el plano *RG*-vs-*BY* usando el mapeo propuesto.

hecho de que las longitudes de onda amarillas producen la respuesta máxima de los conos *L*, las cuales también caen en la cola derecha de la curva de sensibilidad de los conos *M*. Por lo tanto, el ojo humano es muy sensible a los amarillos.

Teniendo en cuenta las consideraciones anteriores, hemos derivado la siguiente asignación alternativa entre los espacios de color *RGB* y *(RG,YB,I)* para producir una distribución de puntos más cercana a la del espacio perceptual teórico de Izmailov y Sokolov (Izmailov and Sokolov, 1991) que el propuesto en (Itti et al., 1998):

$$\begin{aligned} RG &= R-G, \quad I = \frac{3Y+(R+3G+3B)/7}{4}, \\ YB &= \begin{cases} (Y-B)/2 & Y \geq B \\ Y-B & Y < B, \end{cases} \end{aligned} \tag{4.3}$$

con el canal amarillo Y definido como se describe en (4.4) para lograr una respuesta más nítida que con la definición de (4.2):

$$Y = \frac{R+G}{2}(1-|R-G|) \tag{4.4}$$

Las figuras 4.3 y 4.4 muestran el mapeo de los 17 colores, con igualdad de brillo, en los nuevos subespacios *RG-vs-YB* y *RG-vs-I* de acuerdo a (4.3).

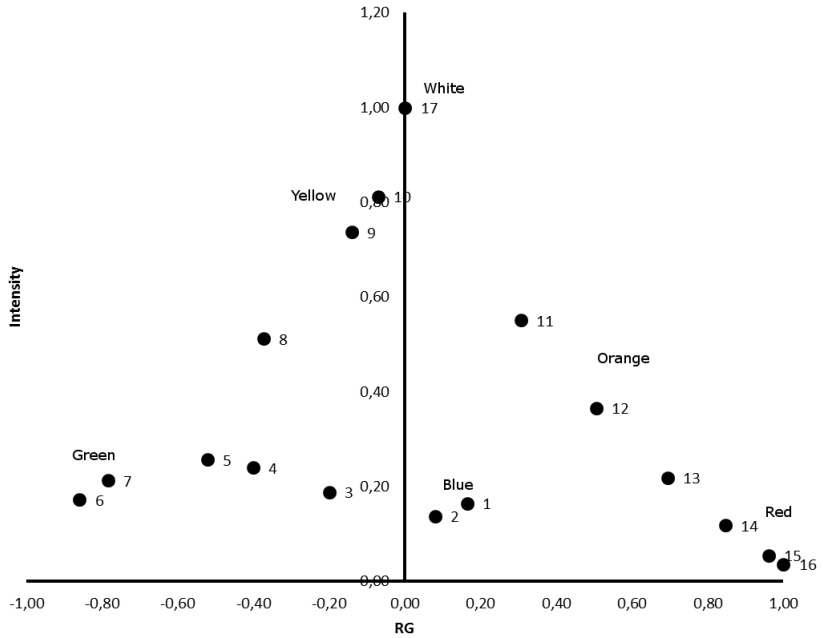

Figura 4.4: Proyección de los 17 colores (en igualdad de brillo) en el plano *RG*-vs-*I* usando el mapeo propuesto.

La correlación entre ambos espacios de color tridimensionales, es decir (X_1, X_2, X_3) y *(RG,YB,I)*, puede ser evaluada numéricamente mediante la determinación de la transformación tridimensional entre los 17 puntos definidos en el espacio de color *(RG,YB,I)* frente a los puntos en el espacio (X_1, X_2, X_3). La transformación de siete valores de Helmert entre ambas nubes de puntos tridimensionales produce un valor de escala de 1.02, un vector de rotación *(-0.02,0.06,0.08)* (en radianes) y un vector de traslación *(-0.03,-0.06,-0.10)*, con un error cuadrático medio (RMSE[2]) de 0.12. Este error es atribuible al hecho de que las posiciones de los 17 puntos en el espacio de color original (X_1, X_2, X_3) son el resultado de las diferencias subjetivas percibidas por los sujetos del experimento, mientras que los puntos definidos en el espacio *(RG,YB,I)* son extraídos a partir de valores exactos de las diferentes longitudes de onda de acuerdo con las asignaciones definidas en (4.3). A su vez, si se utiliza el espacio *(RG,YB,I)* de acuerdo a (4.1), la transformación de Helmert entre ambas nubes de puntos tridimensionales produce un factor de escala de 1.09, un vector de rotación *(0.17,0.15,0.02)* (en radianes) y un vector de traslación *(0.06,0.02,0)*, con un RMSE de 0.22, casi el doble del error que con el mapeo alternativo propuesto.

El espacio de color CIELAB también se consideró como una alternativa al espacio propuesto *(RG,YB,I)* con el fin de aproximarlo al espacio original (X_1, X_2, X_3). En particular, el componente

[2]Root-Mean-Square Error

acromático L^*, normalizado entre cero y uno, podría considerarse equivalente a X_3. A su vez, el componente cromático a^*, que modela el componente rojo/azul, y b^*, que modela el componente amarillo/azul, son aparentemente similares a X_1 y X_2, respectivamente. Sin embargo, cuando los 17 puntos de color utilizados por Izmailov y Sokolov se mapean al espacio CIELAB, sus posiciones relativas tienen una diferencia significativamente mayor con respecto a las del espacio (X_1, X_2, X_3) que las posiciones equivalentes en *(RG,YB,I)*. Esto se corroboró calculando de nuevo la transformación Helmert entre los 17 puntos CIELAB y (X_1, X_2, X_3), lo que da un RMSE de 0.34, casi tres veces el error obtenido con el espacio *(RG,YB,I)* propuesto.

Por lo tanto, se considera que el espacio de color *(RG,YB,I)* propuesto en (4.3) es comparable con el subespacio cromático (X_1, X_2, X_3) original, con la ventaja de que los puntos en el espacio propuesto pueden obtenerse directamente a partir de imágenes digitales en color:

$$X_1 \equiv RG, \quad X_2 \equiv BY, \quad X_3 \equiv I. \tag{4.5}$$

Así, dados dos puntos de color $(RG_i, YB_i, I_i)^T$ y $(RG_j, YB_j, I_j)^T$, podemos redefinir la diferencia perceptual cromática $\triangle C_{ij}$ entre ellos, originalmente definida en (3.1), como:

$$(\triangle C_{ij})^2 = (\triangle RG_{ij})^2 + (\triangle YB_{ij})^2 + (\triangle I_{ij})^2, \tag{4.6}$$

donde $\triangle RG_{ij} = RG_i - RG_j$, $\triangle YB_{ij} = YB_i - YB_j$ y $\triangle I_{ij} = I_i - I_j$.

4.2. Mapeo computacional del subespacio acromático

El subespacio acromático derivado por Izmailov y Sokolov en su segundo experimento es un espacio euclidiano bidimensional en el que cada combinación de niveles de luminancia para el estímulo y el fondo se asocia con un vector (Y_1, Y_2). Con el fin de definir un mapeo entre el espacio de color *RGB* y el subespacio acromático, en primer lugar es necesario asignar cada nivel de gris, normalizado de 0 a 1, a un nivel de luminancia en cd/m^2.

La función de visualización estándar de grises DICOM (GSDF [3]) definida en (NEMA, 2009) describe la relación logarítmica entre el nivel de luminancia y el índice de diferencia mínima (JND [4]). Una JND es la variación mínima de luminancia que puede ser percibida por un observador

[3] Grayscale Standard Display Function

[4] Just-Noticeable Difference

humano. La magnitud de esa variación no es constante, sino proporcional al nivel de luminancia previo en cada variación, de acuerdo con la ampliamente conocida Ley de Weber. En consecuencia, el índice JND, JND_j, se define en (NEMA, 2009) como el valor de entrada al GSDF de tal manera que un incremento en una unidad de JND_j da como resultado una diferencia de luminancia de un JND. Por lo tanto, el GSDF proporciona un mapeo directo, basado en una base perceptual, entre los niveles de luminancia y el espacio uniforme definido por el índice JND.

cd/m^2	0.2	1	2	10	20	100	200
JND_j	22.73	71.49	104.06	216.84	283.42	476.37	572.13
$\overline{JND_j}$	0	0.089	0.148	0.353	0.474	0.826	1

Cuadro 4.2: Indices JND y JND normalizados correspondientes a los siete niveles de luminancia usados por Izmailov y Sokolov.

Los índices JND de los siete niveles de luminancia utilizados por Izmailov y Sokolov se pueden calcular a partir de la formulación del GSDF inverso definido en (NEMA, 2009). Estos valores se muestran en la tabla 4.2. El intervalo uniforme entre el índice JND mínimo, $JND_{min} = 22{,}73$, que corresponde a la luminancia mínima considerada de 0.2 cd/m^2, y el índice JND máximo, $JND_{max} = 572{,}13$, asociado con la luminancia máxima considerada de 200 cd/m^2, se normaliza entre cero y uno:

$$\overline{JND_j} = \frac{JND_j - JND_{min}}{JND_{max} - JND_{min}}, \tag{4.7}$$

donde $\overline{JND_j}$ es el índice JND normalizado correspondiente a JND_j. La tabla 4.2 también muestra los índices JND normalizados correspondientes a los siete niveles de luminancia utilizados por Izmailov y Sokolov. Como se ha descrito anteriormente, cada vector (Y_1, Y_2) es una función de dos niveles de luminancia: uno para el estímulo y otro para el fondo. Siendo α y β los índices JND normalizados asociados con el estímulo y el fondo, respectivamente. Las figuras 4.5 y 4.6 muestran los valores de Y_1 e Y_2 obtenidos por Izmailov y Sokolov para todas las combinaciones de los siete valores de α (es decir, 0, 0.089, 0.148, 0.353, 0.474, 0.826, 1) y los tres valores de β (es decir, 0.089, 0.353, 0.826) considerados en su segundo experimento.

La figura 4.5 muestra que la evolución de Y_1, para un mismo valor de β, se aproxima a una suma de dos Gaussianas. Esto se puede apreciar, por ejemplo, para $\beta = 0{,}353$ (10 cd/m_2). A su vez, en la figura 4.6, si fijamos β, los valores de Y_2 siguen una función similar a la sigmoide de α.

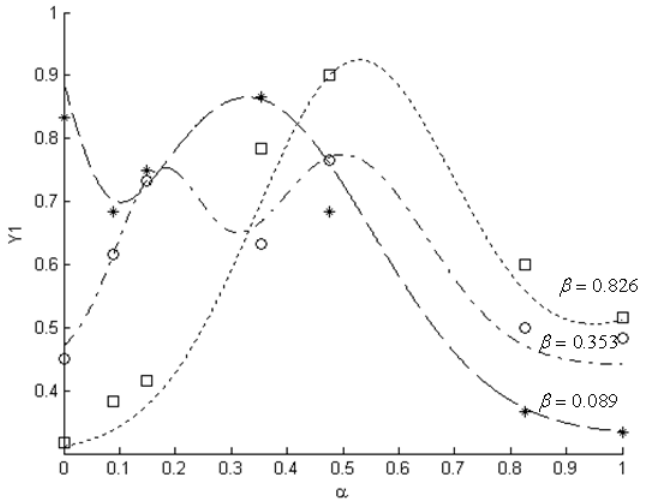

Figura 4.5: Valores de Y_1 para las diferentes combinaciones de estímulo (α) y fondo (β), así como su curva aproximada obtenida mediante regresión no lineal.

En particular, hemos formulado Y_1 como:

$$Y_1(\alpha,\beta) = \frac{s_0(\beta)}{\sigma_0(\beta)\sqrt{2\pi}} e^{-\frac{(\alpha-\mu_0(\beta))^2}{2\sigma_0^2(\beta)}} + \frac{s_1(\beta)}{\sigma_1(\beta)\sqrt{2\pi}} e^{-\frac{(\alpha-\mu_1(\beta))^2}{2\sigma_1^2(\beta)}} + d(\beta). \tag{4.8}$$

Se han estimado los siete coeficientes de la ecuación anterior $(s0, s1, \sigma 0, \sigma 1, \mu 0, \mu 1, d)$ para cada uno de los tres valores de β mediante ajuste no lineal[5] de los pares (Y_1, α), como se muestra en la figura 4.5. Posteriormente, se ajustó un polinomio de segundo grado [6] a los tres valores escalares obtenidos para cada coeficiente:

$$\begin{aligned}
s_0(\beta) &= 0{,}1409\beta^2 - 0{,}0964\beta + 0{,}0255 \\
s_1(\beta) &= 0{,}3109\beta^2 - 0{,}2397\beta + 0{,}0659 \\
\mu_0(\beta) &= -0{,}2295\beta^2 + 1{,}0105\beta - 0{,}1581 \\
\mu_1(\beta) &= 2{,}2817\beta^2 - 0{,}3646\beta + 0{,}3444 \\
\sigma_0(\beta) &= 0{,}2869\beta^2 - 0{,}1268\beta + 0{,}0890 \\
\sigma_1(\beta) &= 1{,}3638\beta^2 - 0{,}8679\beta + 0{,}2864 \\
d(\beta) &= -0{,}96700\beta^2 + 0{,}8441\beta + 0{,}2625
\end{aligned} \tag{4.9}$$

[5]Para el ajuste de curvas no lineales se ha utilizado la función *nlinfit* de MATLAB

[6]Para el ajuste de curvas poligonales se ha empleado la función *polyfit* de MATLAB

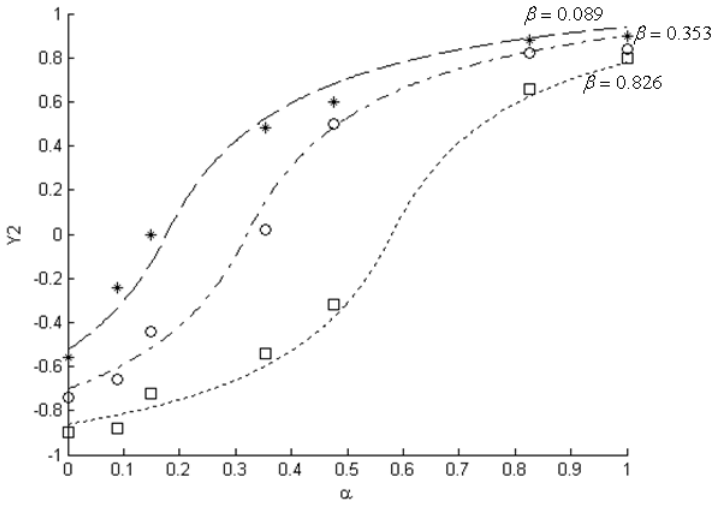

Figura 4.6: Valores de Y_2 para las diferentes combinaciones de estímulo (α) y fondo (β), así como su curva aproximada obtenida mediante regresión no lineal.

La figura 4.7 muestra la superficie bivaluada de la primera componente acromática (Y_1) en función de α y β según (4.8) y (4.9). A su vez, la figura 4.5 muestra las curvas obtenidas evaluando (4.8) para los tres valores β usados por Izmailov y Sokolov.

De manera similar, hemos obtenido la formulación de Y_2 después de un ajuste no lineal de los pares (Y_2, α) mostrados en la figura 4.6 para los tres valores de β:

$$Y_2(\alpha,\beta) = \frac{\gamma(\alpha - \eta(\beta))}{\delta + |\alpha - \eta(\beta)|}, \tag{4.10}$$

donde $\gamma = 1{,}2008$, $\delta = 0{,}2281$ y $\eta(\beta)$ es un polinomio de primer orden obtenido por ajuste polinomial de los valores de η obtenidos para los tres valores de β:

$$\eta(\beta) = 0{,}5495\beta + 0{,}1001. \tag{4.11}$$

La figura 4.8 muestra la superficie bivaluada que modela el segundo componente acromático Y_2 como una función de α y β de acuerdo a (4.10) y (4.11). A su vez, la figura 4.6 muestra las curvas correspondientes a las sigmoides obtenidas mediante la evaluación de (4.10) para los tres valores de β utilizados por Izmailov y Sokolov.

Tras haber formulado los componentes acromáticos Y_1 e Y_2 como funciones bivaluadas de α y β, siendo estos últimos los índices JND normalizados, es esencial definir el mapeo entre los vectores de color en el espacio RGB y estos índices JND normalizados. Para ello, es necesario

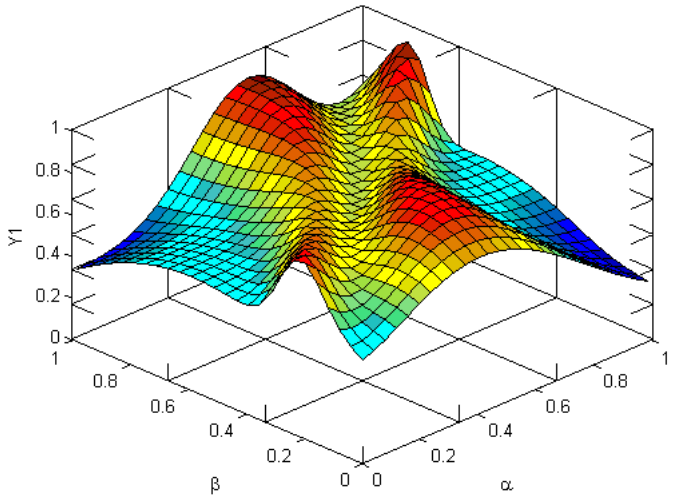

Figura 4.7: La superficie $Y_1(\alpha,\beta)$.

obtener una medida de la luminancia percibida asociada a cada vector *RGB* dado.

En el espacio de color CIELAB, la luminancia o luminosidad percibida L^* correspondiente a los tres canales *RGB* lineales *(R,G,B)* normalizados entre 0 y 1 es una función potencial de $\frac{1}{3}$ de la luminancia relativa Y (Sharma, 2002), siendo esta última (IEC, 1999):

$$Y = 0{,}2126R + 0{,}7152G + 0{,}0722B. \tag{4.12}$$

El componente de luminancia L^* tiene una relación lineal con el índice JND introducido anteriormente. En particular, dados dos colores con los mismos componentes cromáticos a^* y b^*, una diferencia de 2.3 unidades entre los componentes acromáticos L^* es aproximadamente igual a un JND según la ecuación CIE76 (Sharma, 2002). Por lo tanto, si el componente acromático L^* originalmente comprendido entre 0 y 100 se normaliza entre 0 y 1, y la ganancia de la cámara se establece de tal manera que la luminancia máxima considerada ($200cd/m^2$ en este caso) produce la máxima luminosidad ($L^* = 100$), el L^* normalizado puede ser mapeado directamente al espacio de índice JND normalizado definido en (4.7), permitiendo de esta manera la correlación buscada entre vectores de color *RGB* y el índice JND normalizado.

No obstante, dado que los sensores CMOS y CCD típicamente montados en cámaras de vídeo tienen una respuesta lineal a la luz debido a la fotosensibilidad casi lineal del silicio, la mayoría de las cámaras aplican una corrección no lineal (compresión gamma) para emular la respuesta lo-

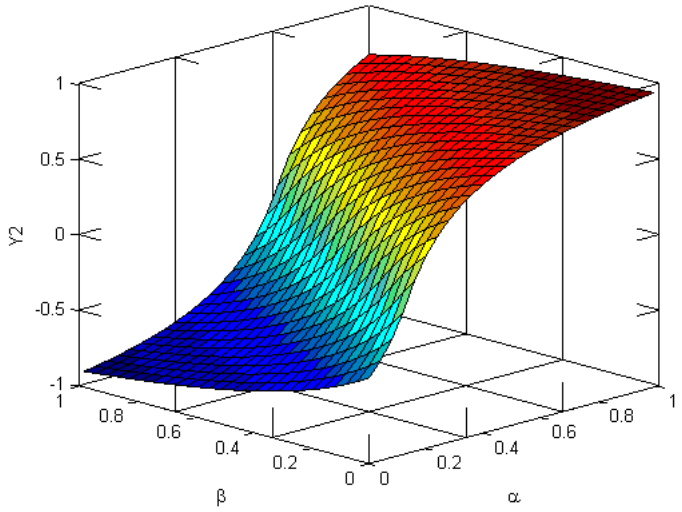

Figura 4.8: La superficie $Y_2(\alpha, \beta)$.

garítmica de la visión humana y, al mismo tiempo, aumentar su rango dinámico. La compresión gamma típicamente aplicada es una función potencial de $\frac{1}{2{,}2}$ sobre los tres canales *RGB* lineales generados por el sensor de la cámara, produciendo los canales *RGB* gamma-comprimidos (R', G', B') asociados a cada píxel de una imagen digital en color.

Para calcular L^* a partir de un vector *RGB* gamma-comprimido, en primer lugar es necesario linealizar el vector de color dado por medio de una compresión gamma inversa (es decir, expansión gamma), que implica la aplicación de una función potencial de 2.2 a los tres canales no lineales. Posteriormente, la luminancia Y se calcula a través de (4.12). Finalmente, L^* se obtiene aplicando la función potencial de $\frac{1}{3}$ definida en la especificación CIELAB (Sharma, 2002). Este proceso es intensivo desde el punto de vista computacional si se va a aplicar a todos los píxeles de una imagen digital en color.

Sin embargo, la norma ITU-R BT.601 (anteriormente conocida como CCIR 601), seguida por la mayoría de las cámaras de vídeo de definición estándar (SDTV[7]), define la luminancia o luma percibida, Y', correspondiente a un vector *RGB* gamma-comprimido (R', G', B') como una aglutinación ponderada de los tres componentes, con los siguientes pesos:

$$Y' = 0{,}299R' + 0{,}587G' + 0{,}114B'. \tag{4.13}$$

[7]Standard Definition Television

Dado que, como se ha mencionado anteriormente, la compresión gamma ya implica la aplicación de una función potencial de $\frac{1}{2{,}2}$ a los componentes lineales RGB, se considera que es una aproximación de la función potencial de $\frac{1}{3}$ aplicada para el cálculo de L^*. De esta manera, la luma puede considerarse, computacionalmente hablando, como una aproximación muy eficiente de la luminosidad L^* del CIELAB. En particular, suponiendo que los tres componentes gamma-comprimidos están normalizados entre 0 y 1, la luma se usa directamente en este trabajo para definir el índice JND normalizado (4.3) ($\overline{JND_j} = Y'$), y por tanto, los parámetros α y β necesarios para evaluar Y_1 e Y_2 en (4.8) y (4.10), respectivamente. Teniendo en cuenta que, dado que el nivel máximo de luminancia de la escena real representada en una imagen digital no se conoce en general, la asignación de la máxima luma al índice JND normalizado máximo representa en sí misma una normalización implícita de esa luminancia máxima desconocida al nivel máximo de $200 cd/m^2$ utilizado por Izmailov y Sokolov.

Así, hemos definido finalmente α como el luma de un píxel dado, mientras que β se obtiene como el luma promedio de los píxeles pertenecientes a su vecindario. Por razones computacionales, hemos considerado una vecindad 3x3 (es decir, los ocho píxeles adyacentes). Se requiere un trabajo adicional para analizar la influencia del tamaño y la forma del vecindario en el resultado final.

Parte II

APLICACIÓN DEL MODELO PERCEPTUAL

Capítulo 5

Modelo de atención visual basado en el modelo de Izmailov y Sokolov

Los modelos de atención visual suelen generar un mapa de saliencia a partir de una imagen digital dada, de modo que cada píxel de la imagen está asociado con un elemento del mapa, cuyo valor es proporcional a la atracción visual de ese píxel con respecto a algunas características visuales (color, brillo, etc.). El mapa de saliencia se obtiene mediante la integración de mapas de saliencia parciales (mapas de conspicuidad) generados para cada una de las características visuales tomadas en cuenta. Por ejemplo, el conocido modelo IKN (Itti et al., 1998) promedia los mapas de conspicuidad generados para tres características visuales: color, brillo y orientación (bordes). Así, tanto el color como el brillo se procesan independientemente y tienen un mismo peso en la saliencia final.

En este capítulo, proponemos un modelo de atención visual basado en la arquitectura de IKN pero que integra el color y el brillo a través de la adaptación computacional presentada en el capítulo 4 del modelo perceptual propuesto por Izmailov y Sokolov (Izmailov and Sokolov, 1991). Dicha integración esta fundamentada perceptualmente y produce resultados más consistentes que cuando ambas características visuales se procesan de forma independiente, como se mostrará en el siguiente capítulo en el ámbito de aplicación del seguimiento de objetos en secuencias de vídeo. Este modelo queda representado en la figura 5.1.

Un punto de imagen se considera saliente o destacable con respecto a una característica visual dada (por ejemplo, el brillo) si hay una diferencia significativa entre el valor de esta característica asociada a ese punto (centro) y los valores correspondientes a los puntos dentro de su vecindario (envolvente). Este llamado antagonismo centro-envolvente ocurre en las células fotorreceptoras

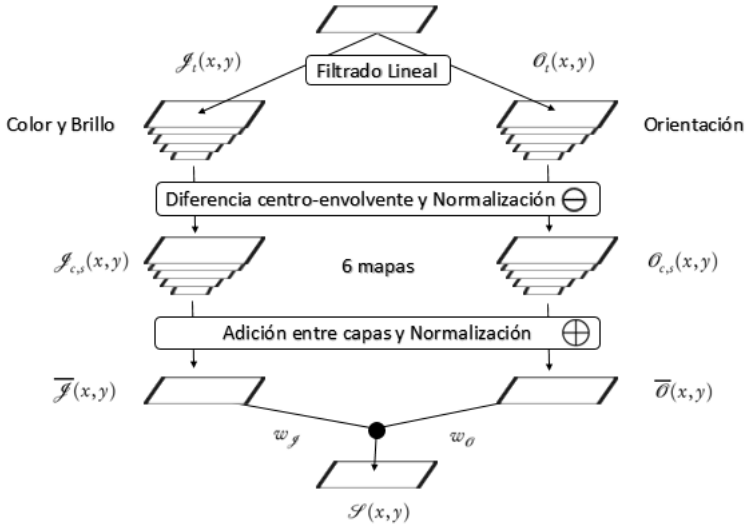

Figura 5.1: Representación gráfica del modelo propuesto.

de la retina de los primates y permite que la corteza visual, por ejemplo, detecte los bordes espaciales. En la práctica, las diferencias centro-envolvente se pueden calcular restando una versión de la imagen de bajo detalle, denominada de escala *gruesa*, de una versión detallada de la misma, denominada de escala *fina* (Itti et al., 1998).

Sea $\psi(x,y) = \psi_0(x,y)$ la imagen original a escala 0. La imagen en una escala t, $\psi_t(x,y)$, se obtiene aplicando un submuestreo al resultado de aplicar un filtro gaussiano a la imagen en escala $t-1$. Así, la imagen en escala t tiene un factor de reducción de $1:2^t$ con respecto a la imagen original. Por ejemplo, IKN aplica 9 escalas (es decir, escalas de 0 a 8). La imagen original y sus sucesivas aproximaciones de bajo detalle constituyen lo que se denomina una pirámide gaussiana.

Sea $\mathcal{F}_t(x,y)$ los valores de una característica visual correspondiente a los píxeles de $\psi_t(x,y)$. Un mapa de características, $\mathcal{F}_{c,s}(x,y)$, se define como el valor absoluto de la diferencia entre escalas, es decir, entre los valores de la característica visual en la escala *fina* c (centro), con la de nivel *grueso* s (envolvente), respectivamente: $\mathcal{F}_{c,s}(x,y) = |\mathcal{F}_c(x,y) \ominus \mathcal{F}_s(x,y)|$, donde el operador de diferencia entre escalas $\ominus$ interpola la escala de bajo detalle (*gruesa*) a la escala detallada (*fina*), para aplicar una sustracción punto a punto.

Los mapas de características se normalizan para poder combinar mapas correspondientes a diferentes características visuales. El operador de normalización $\mathcal{N}$ se define como: $\mathcal{N}(\mathcal{F}_{c,s}(x,y)) = (M-\overline{m})2\mathcal{F}_{c,s}(x,y)$, donde M es el valor del máximo global de $\mathcal{F}_{c,s}$ y $\overline{m}$ es el valor medio de sus máximos locales. De forma similar a IKN, determinamos varios mapas de ca-

racterísticas para cada característica visual, tal que $c_m \leq c \leq c_M$ y $s = c + \delta$, con $\delta_m \leq \delta \leq \delta_M$. En IKN, $c_m = 2$, $c_M = 4$, $\delta_m = 3$ y $\delta_M = 4$, lo que da seis mapas de características que abarcan siete escalas (de 2 a 8). Como se describirá en la sección 7.2, hemos configurado el modelo propuesto con $c_m = 4$, $c_M = 5$ $\delta_m = 1$ y $\delta_M = 3$, que también produce seis mapas que abarcan las cinco escalas superiores (de 4 a 8). Los mapas normalizados correspondientes a la misma característica visual se combinan en un solo mapa de conspicuidad, $\overline{\mathscr{F}}(x,y)$, a través del operador de adición entre escalas $\bigoplus$, que interpola los mapas dados a una misma escala (por ejemplo, escala 4 como se propone en IKN) y luego realiza una adición punto a punto:

$$\overline{\mathscr{F}}(x,y) = \bigoplus_{c=c_m}^{c_M} \bigoplus_{s=c+\delta_m}^{c+\delta_M} \mathscr{N}(\mathscr{F}_{c,s}(x,y)). \tag{5.1}$$

Por último, el mapa de saliencia final $S(x,y)$ se obtiene promediando los mapas de conspicuidad asociados con cada característica visual considerada. En particular, siete características visuales se toman en cuenta independientemente y se integran en IKN: una característica acromática, dos características cromáticas y cuatro características de orientación. La característica acromática es el componente de intensidad I definido en (4.1). Las características cromáticas son aproximadamente equivalentes a los componentes RG y BY definidos en (4.1). A su vez, las cuatro características de orientación son el resultado de los respectivos filtros de Gabor con orientaciones 0, 45, 90 y 135 grados, respectivamente.

Alternativamente, nuestro modelo de atención visual se basa en dos características visuales: una característica cromática-acromática conjunta, $\mathscr{J}_t(x,y)$, basada en el modelo perceptual propuesto por Izmailov y Sokolov y una característica de orientación $\mathscr{O}_t(x,y)$, basada en el filtro de Sobel. En particular, definimos la característica cromático-acromática correspondiente a un píxel de la imagen $\psi(x,y)$ como un vector pentadimensional:

$$\mathscr{J}_t(x,y) = (X_{1,t}(x,y), X_{2,t}(x,y), X_{3,t}(x,y), Y_{1,t}(x,y), Y_{2,t}(x,y)), \tag{5.2}$$

donde $X_{i,t}(x,y)$ es el i-ésimo componente cromático del modelo perceptual de Izmailov y Sokolov evaluado en $\psi(x,y)$ según (4.5), mientras que $Y_{1,t}(x,y)$ e $Y_{2,t}(x,y)$ son los componentes acromáticos del modelo perceptual de Izmailov y Sokolov evaluados en $\psi(x,y)$ según (4.8) y (4.10), respectivamente. Teniendo esto en cuenta, definimos el mapa de características cromático-

acromático conjunto como:

$$\mathscr{J}_{c,s}(x,y) = |\mathscr{J}_c(x,y) \ominus \mathscr{J}_s(x,y)|, \tag{5.3}$$

donde la sustracción punto a punto inherente al operador de diferencia entre escalas $\ominus$ se define en este caso como la diferencia de percepción $\triangle S_{ij}$ entre los colores correspondientes a ambos puntos pentadimensionales, ${}^{i}\mathscr{J}_c(x,y)$ y ${}^{j}\mathscr{J}_c(x,y)$, de acuerdo con el modelo de Izmailov y Sokolov (3.3):

$${}^{i}\mathscr{J}_c(x,y) - {}^{j}\mathscr{J}_c(x,y) = \sqrt{(\triangle C_{ij,c})^2 + {}^{i}X_{3,c}(x,y)\,{}^{j}X_{3,c}(x,y)(\triangle W_{ij,c})^2}, \tag{5.4}$$

siendo $(\triangle C_{ij,c})^2$ definido como (4.6) y $(\triangle W_{ij,c})^2$ como (3.2). Los seis mapas de características se normalizan finalmente y se combinan con el operador de adición entre escalas (5.1), produciendo el mapa de conspicuidad cromático-acromático conjunto:

$$\overline{\mathscr{J}}(x,y) = \bigoplus_{c=c_m}^{c_M} \bigoplus_{s=c+\delta_m}^{c+\delta_M} \mathscr{N}(\mathscr{J}_{c,s}(x,y)). \tag{5.5}$$

A su vez, calculamos el mapa de conspicuidad de orientación mediante el operador de Sobel en lugar de los cuatro filtros de Gabor propuestos en IKN. Ya se ha demostrado que el detector de bordes de Sobel es una alternativa computacional eficiente a los cuatro filtros de Gabor aplicados en IKN (Won et al., 2008; Fernández-Carbajales et al., 2011). En particular, sea $I_t(x,y)$ la imagen de escala de grises correspondiente a $\psi_t(x,y)$, con la característica de brillo I calculada como (4.3). Definimos la característica de orientación de la imagen como:

$$\mathscr{O}_t(x,y) = \sqrt{(I_t(x,y) * S_x)^2 + (I_t(x,y) * S_y)^2}, \tag{5.6}$$

donde S_x y S_y son, respectivamente, los *kernels* 3x3 horizontales y verticales del operador de Sobel y $*$ denota el operador de convolución. El mapa de conspicuidad de orientación $\overline{\mathscr{O}}(x,y)$ se define como:

$$\begin{aligned} \overline{\mathscr{O}}(x,y) &= \bigoplus_{c=c_m}^{c_M} \bigoplus_{s=c+\delta_m}^{c+\delta_M} \mathscr{N}(\mathscr{O}_{c,s}(x,y)) \\ \mathscr{O}_{c,s}(x,y) &= |\mathscr{O}_c(x,y) \ominus \mathscr{O}_s(x,y)|. \end{aligned} \tag{5.7}$$

Definimos el mapa de saliencia final $\mathscr{S}(x,y)$ como el promedio ponderado de los dos mapas

de conspicuidad definidos anteriormente:

$$\mathcal{S}(x,y) = w_{\mathcal{J}}\overline{\mathcal{J}}(x,y) + w_{\mathcal{O}}\overline{\mathcal{O}}(x,y). \tag{5.8}$$

Ambos mapas de conspicuidad tienen el mismo peso, $w_{\mathcal{J}} = w_{\mathcal{O}} = 0{,}5$, siguiendo el ejemplo de IKN. De acuerdo con esta formulación, las saliencias se normalizan entre 0 y 1, siendo valores cercanos a 1 los que representan una gran atracción visual.

Capítulo 6

Algoritmo de seguimiento de objetos basado en el nuevo modelo de atención visual

El modelo de atención visual descrito en el capítulo 5 es ventajoso para la detección de rasgos visuales salientes o destacables, que pueden ayudar a mejorar otras aplicaciones de alto nivel de Visión por Computador. En particular, en esta sección se describe un sencillo algoritmo de seguimiento de objetos basado en la correspondencia directa de bloques de la imagen extraídos según estas características. El algoritmo de seguimiento propuesto asume una secuencia de vídeo adquirida con una cámara fija. Los objetos en movimiento en la escena se extraen mediante la aplicación del algoritmo de sustracción de fondo propuesto en (Cavallaro et al., 2005), que segmenta la imagen actual en píxeles de fondo y primer plano. Un algoritmo de etiquetado por componentes conexas aplicado a los píxeles de primer plano determina regiones aisladas, que se denominarán *blobs* en lo sucesivo. Los *blobs* con un área por debajo de un umbral predefinido se descartan al ser considerados ruido producido por la etapa de segmentación previa.

Dado un conjunto de *blobs* separados extraídos de la imagen actual, el objetivo de un algoritmo de seguimiento de objetos en secuencias de vídeo es asociar cada nuevo *blob* con otro *blob* candidato extraído de imágenes anteriores, siempre y cuando ambos *blobs* sean visualmente similares. En particular, el algoritmo propuesto determina la similitud visual entre cada *blob* de la imagen actual y todos los *blobs* extraídos de las últimas imágenes M (M se ha fijado en 25 en el desarrollo de este trabajo). También se mantiene una lista histórica de *blobs* activos, la cual contiene los *blobs*

que han sido exitosamente asociados, así como los *blobs* que no lo han sido. Estos últimos tienen un contador de vida asociado, inicialmente fijado a cero, que se incrementa después de procesar cada nueva imagen. Si este contador alcanza un valor de M, el *blob* no asociado que tiene este contador se elimina de la lista. Si un nuevo *blob* puede asociarse con un *blob* anterior, el *blob* anterior de la lista se sustituye por el nuevo. De lo contrario, el nuevo *blob* se agrega a la lista. En ambos casos, el contador de vida del nuevo *blob* se restablece a 0.

Además de la información geométrica, como el área, la anchura, la altura y las coordenadas del centroide, el algoritmo almacena, para cada *blob*, las partes tanto de la imagen original como del mapa de saliencia que intersecan con la máscara binaria del *blob*. Con el fin de reducir la saliencia de las características visuales pertenecientes a los bordes de oclusión (los contornos de imagen debidos a la frontera entre el objeto en primer plano y el fondo de la escena) que no caracterizan así el interior del objeto rastreado y por lo tanto el objeto en sí, el mapa de saliencia del *blob* se modula multiplicándolo por la transformada de distancia morfológica del *blob*.

Se determina la similitud visual entre un nuevo *blob* B_{new} y uno anterior B_{old} de la siguiente manera. Se determinan las coordenadas (x,y) correspondientes al valor máximo del mapa de saliencia modulada asociado con B_{new}. Esta es la ubicación de la característica más saliente, visualmente hablando, en el interior del *blob*. Un bloque de imagen rectangular, de tamaño $w \times h$ centrado en esa ubicación, se define en la imagen del *blob*, siendo w y h una fracción de la anchura y altura del *blob*, respectivamente (en este trabajo, el tamaño del bloque es de un quinto del tamaño del *blob*). Este bloque de imagen extraído de B_{new} se busca en toda la imagen asociada con B_{old} a través de la coincidencia de bloque, usando la suma de diferencias absolutas (SAD). Una vez que el primer bloque ha sido procesado y su menor medida de diferencia con B_{old} almacenada, se extrae un nuevo bloque centrado en la segunda característica visual más saliente o destacada en B_{new}.

Con el fin de evitar una concentración de bloques alrededor de la máxima saliencia, todos los valores de saliencia en un vecindario de $2w \times 2h$ centrado en la máxima saliencia encontrada anteriormente son fijados a cero. Esto corresponde al concepto de inhibición local o "inhibición de retorno"que se considera típicamente en la atención visual, por ejemplo en (Itti et al., 1998). Una vez que se encuentra el segundo bloque de imagen, se obtiene nuevamente su mínima diferencia con B_{old}. El proceso se itera hasta que un máximo de N bloques de imagen extraídos de B_{new} (N ha sido experimentalmente fijada a 5 en este trabajo), o el mapa de saliencia modulada se anula completamente debido a la inhibición local. Para mayor completitud, se aplica el mismo proceso

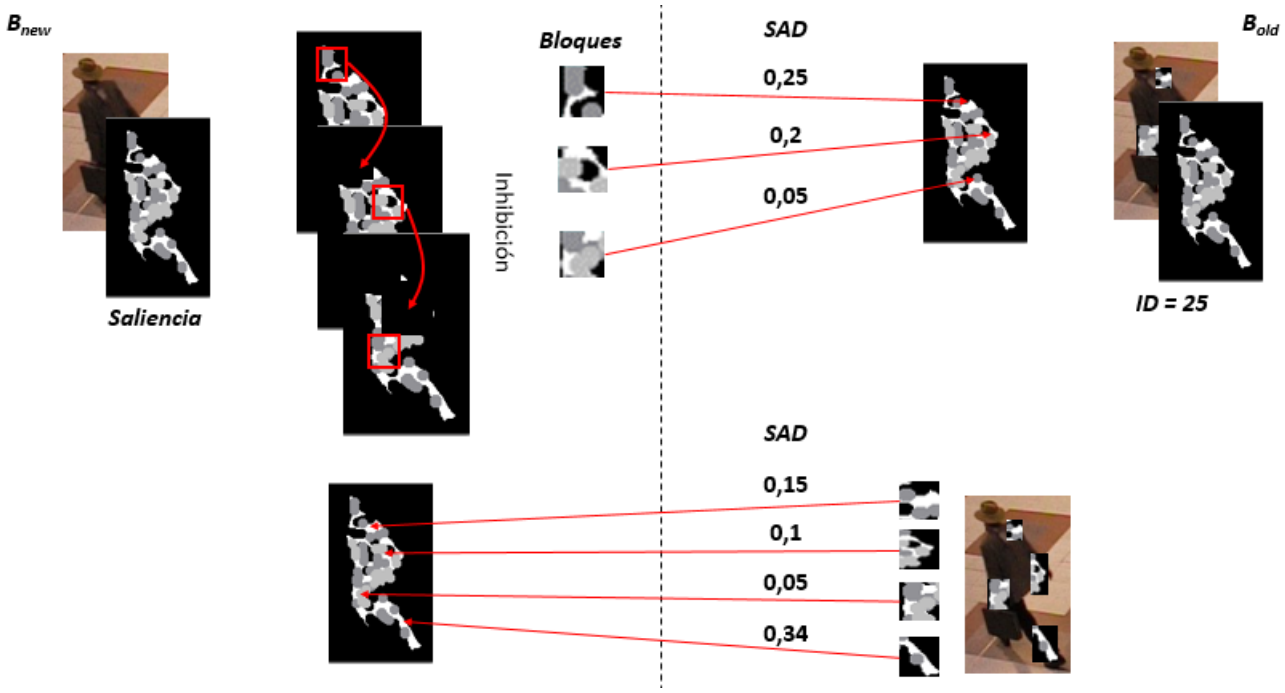

Figura 6.1: Ejemplo de extracción y asignación de bloques.

de extracción de bloques a B_{old} y se calculan las diferencias mínimas de esos bloques de imágenes con B_{new}.

Finalmente, obtenemos la similitud visual entre los dos *blobs* como la inversa del promedio de las diferencias estimadas como se ha descrito anteriormente. Las similitudes visuales obtenidas de esta manera y los pares de *blobs* asociados se clasifican entonces en orden descendente de similitud, excluyendo aquellas parejas cuya similitud es inferior a un umbral mínimo. En este punto, el algoritmo elige el emparejamiento con la máxima similitud. El nuevo *blob* asociado se empareja con su correspondiente *blob* antiguo. A continuación, se considera el siguiente emparejamiento, y un nuevo *blob* se empareja con un *blob* antiguo, excepto si el *blob* antiguo ya ha sido emparejado anteriormente. Este procedimiento codicioso se aplica hasta que no haya más emparejamientos disponibles. Los nuevos *blobs* que no han podido ser emparejados después de todo el proceso se consideran como nuevos objetos y se añaden a la lista histórica como se ha descrito anteriormente.

Capítulo 7

Resultados experimentales

En este capítulo vamos a presentar los resultados experimentales de rendimiento del algoritmo de seguimiento propuesto en el capítulo anterior basado en el modelo de atención visual propuesto en el capítulo 5 y lo comparamos con modelos alternativos de diferencias cromáticas, así como de modelos de atención visual. Hemos segmentado cuatro secuencias de vídeo (denotadas como A, B, C y D) del conjunto de datos PETS (CVPR06, 2006) en fragmentos de secuencias de imágenes con múltiples objetos en movimiento con el fin de evitar períodos donde no hay presencia de objetos en movimiento, así como secuencias de seguimiento muy simples (con un único objeto). Cada fragmento es una secuencia de vídeo de prueba independiente. Los fragmentos iniciales y las secuencias de vídeo obtenidas, después de aplicar la técnica propuesta, se muestran en un sitio web complementario[1].

Se ha anotado cada imagen de cada secuencia de prueba para definir: 1) el fondo de detección, que consiste en una máscara binaria que indica los píxeles que constituyen los objetos en movimiento dentro de cada imagen (es decir, los píxeles de primer plano), 2) un identificador numérico, 3) las coordenadas de la esquina inferior izquierda del cuadro delimitador y 4) las dimensiones de éste para cada objeto dentro de la imagen. El identificador numérico es único para cada objeto de la secuencia de prueba.

El algoritmo de seguimiento de objetos consiste en una etapa de detección inicial que segmenta cada nueva imagen en objetos separados (*blobs*) y una etapa posterior de seguimiento que determina el identificador numérico de cada objeto dentro de cada imagen.

Hemos evaluado el rendimiento del seguimiento en todos los experimentos mediante las me-

[1]Las secuencias de vídeo en alta resolución, datos detallados de los experimentos y las secuencias de video resultantes de aplicar la técnica propuesta se encuentran en https://sites.google.com/site/fgmtracking/

didas de ATA (medidas CLEAR) y MOTA (medidas VACE) descritas en (Kasturi et al., 2009). Ambas medidas se normalizan entre cero y uno, siendo uno el valor indicativo del mejor rendimiento posible. Además, hemos incluido las medidas MMR, FNR y FPR propuestas en (Dicle et al., 2013) para evaluar mejor la robustez de los algoritmos respecto a la calidad de la detección.

En las siguientes tres secciones se van a describir los tres grupos de experimentos realizados. El primer grupo evalúa el algoritmo de seguimiento propuesto anteriormente considerando diferentes modelos de diferencias cromáticas. El segundo grupo evalúa el algoritmo de seguimiento propuesto considerando diferentes modelos de atención visual. En ambos casos, el modelo de atención visual propuesto, que se basa en la adaptación computacional del modelo perceptual teórico de Izmailov y Sokolov, es superior a las otras alternativas probadas en este ámbito de aplicación.

Finalmente, con el fin de ilustrar los beneficios de incorporar el modelo de atención visual propuesto en algoritmos de seguimiento alternativos, el tercer grupo de experimentos aplica el modelo de atención visual propuesto en dos algoritmos de seguimiento de objetos del estado del arte: (Bouguet, 2000) y (Kwon and Lee, 2011), así como su comparación con las versiones originales de los mismos y el algoritmo de seguimiento propuesto.

7.1. Rendimiento del seguimiento para diferentes modelos de diferencia cromática

Con el fin de validar la diferencia cromática perceptual definida en (3.3), basada en la adaptación computacional propuesta del modelo teórico de Izmailov y Sokolov (4.3), se ha evaluado el rendimiento del algoritmo de seguimiento introducido en el capítulo 6 usando diferentes modelos de diferencias cromáticas: (*a*) La diferencia cromática propuesta (3.3) usando el mapeo propuesto al espacio perceptual original de Izmailov y Sokolov (4.3), indicado por **IS**. (*b*) La diferencia cromática propuesta (3.3) utilizando el mapeo al espacio perceptual original basado en (Itti et al., 1998), como se define en (4.1), indicado por **IS2**. (*c*) La diferencia cromática **CIE76**. (*d*) La diferencia cromática **CIE94**. (*e*) La diferencia cromática **CIE2000**. Para las ecuaciones CIE, $\triangle C$ en (3.3) es sustituido por la $\triangle E$ de CIE según se define en (Sharma, 2002).

La figura 7.1 muestra las medidas relacionadas con el seguimiento (es decir, 1-ATA y 1-MOTA) para cada una de las cuatro secuencias PETS consideradas y los diferentes modelos de diferencia

cromática. Cada punto dado representa el rendimiento promedio de los fragmentos extraídos de la secuencia correspondiente.

Estos resultados muestran que la adaptación computacional del modelo teórico de Izmailov y Sokolov (4.3) produce resultados de seguimiento significativamente mejores que los otros modelos de diferencias cromáticas.

7.2. Rendimiento del seguimiento para diferentes modelos de atención visual

Como se indicó anteriormente, también se han evaluado el rendimiento del algoritmo de seguimiento de objetos propuesto en el capítulo 6 con los siguientes modelos de atención visual: (*a*) La adaptación computacional del modelo de Izmailov y Sokolov descrito en el capítulo 5, denotado como **IS**. (*b*) El Modelo de Itti, Koch y Niebur descrito en (Itti et al., 1998), denotado como **IKN**. (*c*) La variación de IKN propuesta por Won, Lee y Son en (Won et al., 2008), denotada como **WLS**. (*d*) La variación de IKN y WLS previamente propuesta por los autores (Fernández-Carbajales et al., 2011), denotada como **HYBRID**. (*e*) El modelo propuesto por Hou y Zhang (Hou and Zhang, 2007), denotado como **HZ**. (*f*) El modelo propuesto por Maruta, Sato e Isshi (Maruta et al., 2010), denotado como **MSI**. (*g*) El modelo propuesto por Judd, Ehinger, Durand y Torralba (Judd et al., 2009), basado en una combinación de diferentes niveles de rasgos de imagen, denotado como **JEDT**. (*h*) El método propuesto por Avraham y Lindenbaum en (Avraham and Lindenbaum, 2010), basado en un modelo estocástico validado, que estima la probabilidad de que una parte de la imagen sea de interés, denotado como **ESALIENCY**. (*i*) Un modelo simple en el que los bloques de imagen se extraen usando como medida de saliencia la función de caracterización de bordes del detector de esquinas Harris, denotado como **HARRIS**. (*j*) Un modelo simple que utiliza como medida de saliencia la magnitud del gradiente estimada con el operador de Sobel, denotado como **SOBEL**. (*k*) Un algoritmo muy simple que extrae aleatoriamente bloques de la imagen, denotado como **ALEAT**.

Se han probado dos variaciones de los métodos basados en mapas de saliencia (en todas excepto con ALEAT) dependiendo de si los bloques de imagen se extraen según los máximos locales del mapa de saliencia (denotado como **MAX**) o mediante los máximos locales resultantes de la suma de saliencias sobre una ventana local del tamaño del bloque (denotado como **SUM**). Este

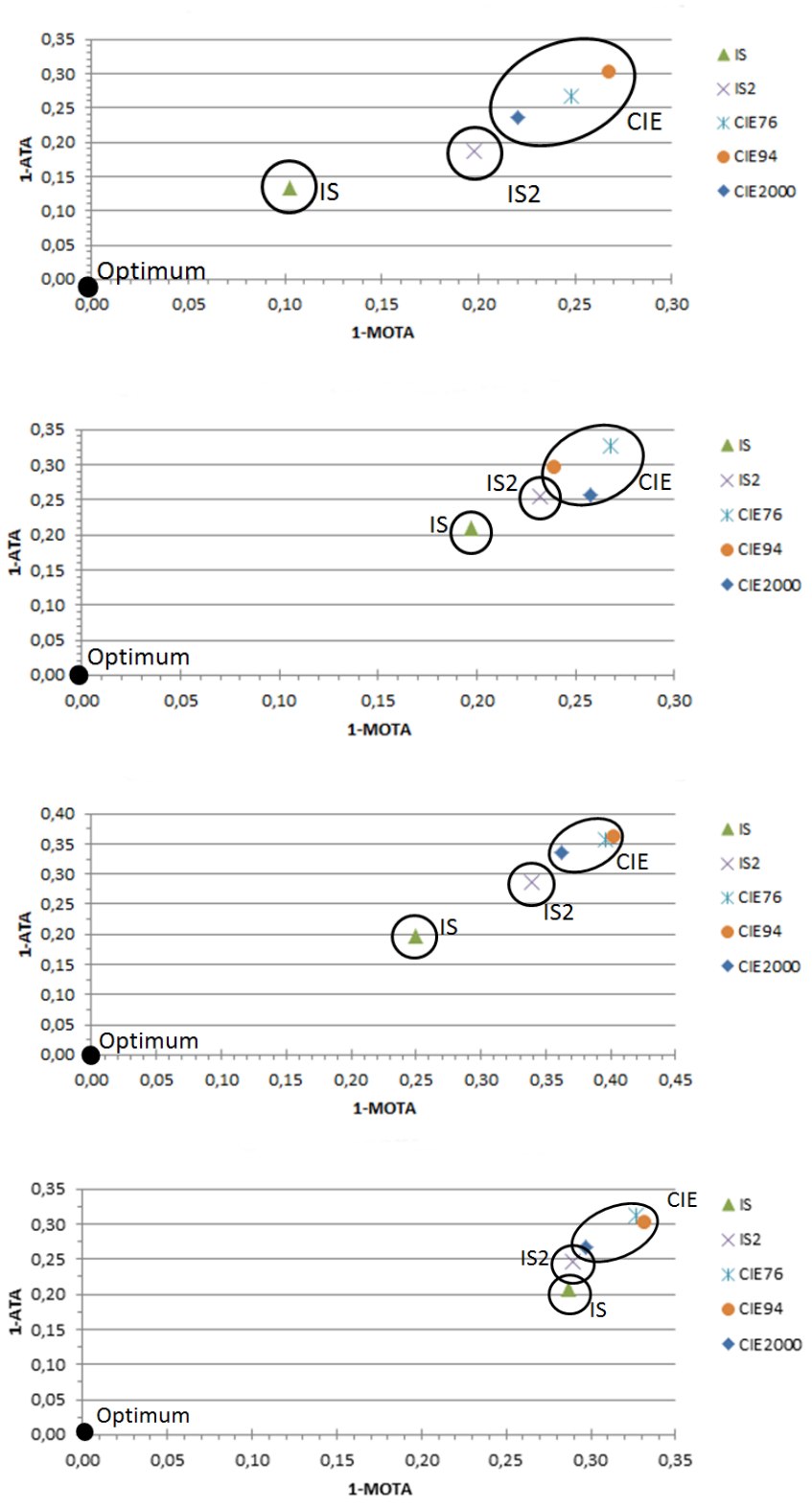

Figura 7.1: Rendimiento del seguimiento para los diferente modelos de diferencias cromáticas desde la secuencia A (superior) hasta la D (inferior).

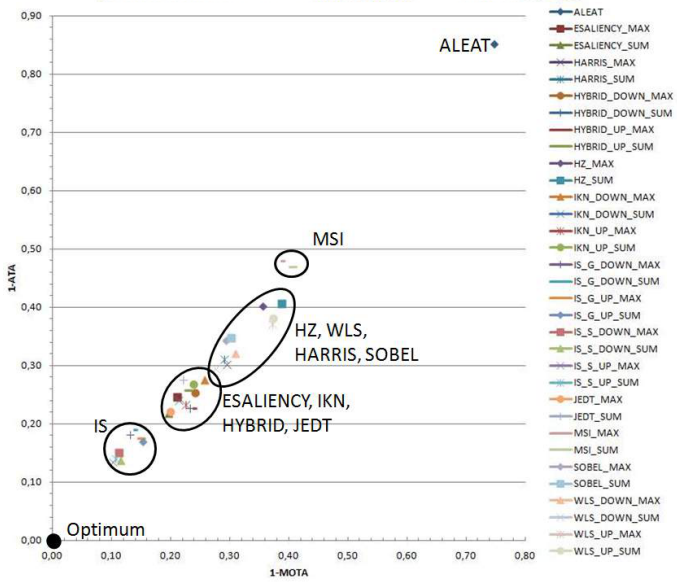

Figura 7.2: Rendimiento del seguimiento de los diferentes modelos de atención visual (secuencia de prueba A)

último tiene como objetivo filtrar el ruido de la saliencia. A su vez, hemos probado dos variaciones adicionales de los métodos de escala múltiple IS, IKN, WLS e HYBRID dependiendo de si los mapas de características integrados pertenecen al rango bajo de escalas (de 2 a 6), denotado como **DOWN**, o al rango alto (de 4 a 8), denotado como **UP**. En particular, de acuerdo con (5.1), DOWN corresponde a $c_m = 2, C_M = 3, \delta_m = 1 y \delta_M = 3$, mientras que UP corresponde a $c_m = 4, C_M = 5, \delta_m = 1 y \delta_M = 3$. Por último, hemos considerado dos variaciones del modelo IS propuesto en función de si el operador Sobel se utiliza para extraer los mapas de orientación como se describe en el capítulo 5 (denotado como **S**) o usando los filtros de Gabor según la aproximación aplica en IKN (denotado como **G**).

Dado que todos los algoritmos de seguimiento evaluados tienen la misma fase de detección de *blobs*, diferenciándose exclusivamente en el modelo de atención visual aplicado, sólo las medidas ATA y MOTA relacionadas con la evaluación del rendimiento de estos algoritmos tienen variaciones. En particular, de la figura 7.2 a la 7.5 solo se muestra el resultado de ambas medidas (es decir, 1-ATA y 1-MOTA) correspondientes a las cuatro secuencias de PETS consideradas para los diferentes modelos de atención visual probados. Cada punto dado representa el rendimiento promedio de los fragmentos extraídos de la secuencia correspondiente.

Estos resultados muestran que las diferentes variaciones del modelo de atención visual propuesto producen resultados de seguimiento significativamente mejores que los otros modelos

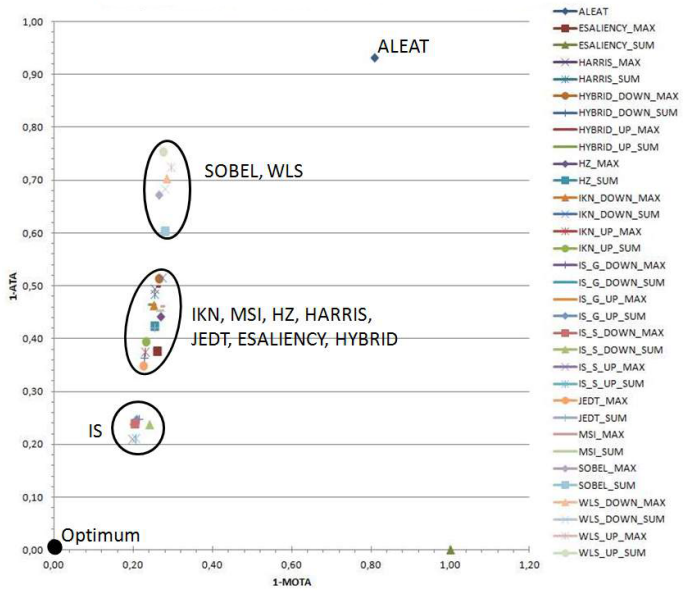

Figura 7.3: Rendimiento del seguimiento de los diferentes modelos de atención visual (secuencia de prueba B)

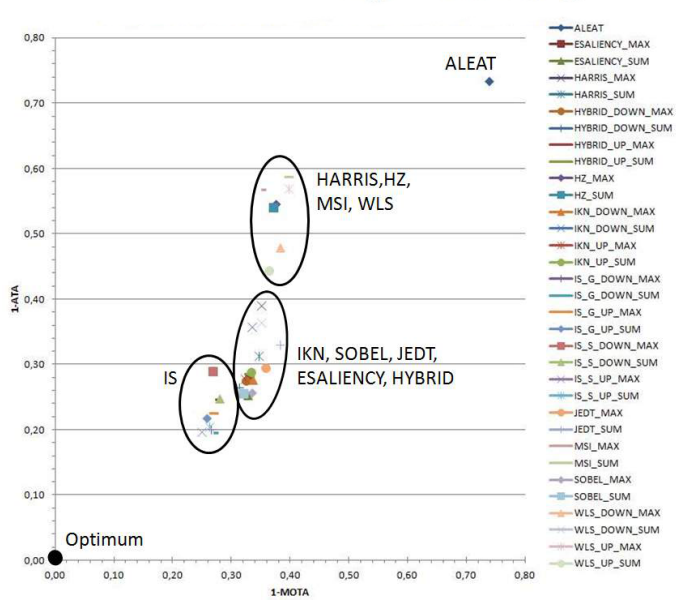

Figura 7.4: Rendimiento del seguimiento de los diferentes modelos de atención visual (secuencia de prueba C)

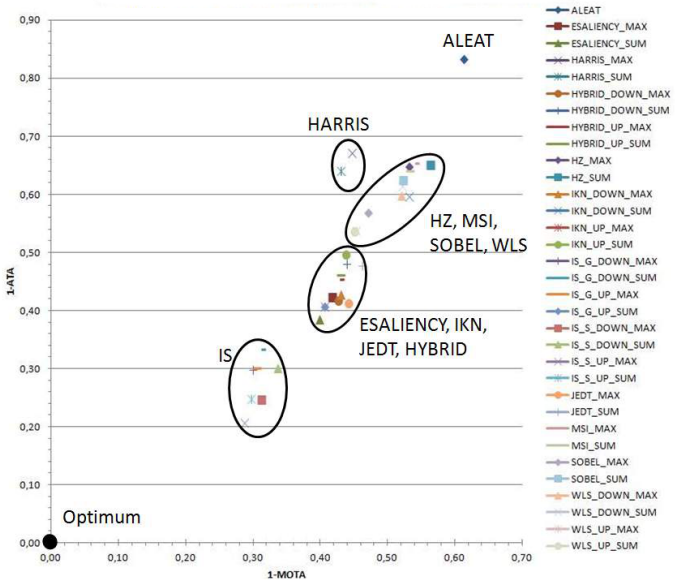

Figura 7.5: Rendimiento del seguimiento de los diferentes modelos de atención visual (secuencia de prueba D)

probados. Entre estos últimos, IKN, ESALIENCY y JEDT son los modelos con el rendimiento más cercano. Curiosamente, los modelos simples y extremadamente eficientes, basados tanto en el detector de Harris como en el operador de Sobel, proporcionan un rendimiento en el seguimiento muy competitivo, en algunos casos incluso comparable al de modelos de atención visual más avanzados y mucho más complejos.

Con respecto a las variaciones del modelo de atención visual propuesto, estos resultados indican que la versión Sobel tiene un mejor rendimiento de seguimiento que la versión basada en los filtros de Gabor, además de una eficiencia computacional mucho más alta. Con respecto a SUM y MAX, ambas alternativas han mostrado un desempeño similar. No obstante, SUM se prefiere debido a sus propiedades de filtrar el ruido y por poderse calcular eficientemente a través de imágenes integrales. Por último, trabajar en el rango superior de escalas (UP) generalmente produce un mejor rendimiento que cuando se considera el rango inferior (DOWN), además de su mayor rendimiento relacionado.

Además de la evaluación anterior, también hemos analizado la robustez al ruido del algoritmo de seguimiento propuesto con los modelos de atención visual alternativos, siguiendo el mismo enfoque propuesto en (Dicle et al., 2013). En particular, hemos llevado a cabo los mismos dos conjuntos de experimentos con la introducción de falsos positivos y falsos negativos aleatorios. En el primer conjunto, hemos aumentado el número de falsos positivos mediante la inyección de falsas

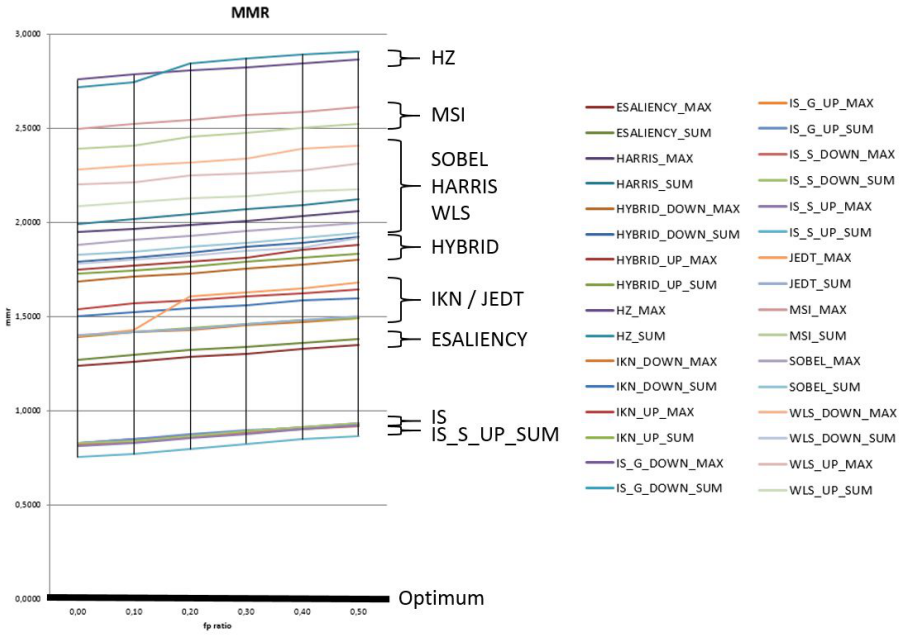

Figura 7.6: Resultados de MMR con incremento de falsos positivos.

detecciones uniformemente distribuidas, mientras que en el segundo conjunto, se ha realizado una inyección de falsos negativos mediante la eliminación uniforme de detecciones verdaderas. Se han evaluado las mismas medidas propuestas en (Dicle et al., 2013): (*a*) Ratio de Falsos Negativos (FNR[2]), (*b*) Ratio de Falsos Positivos (FPR[3]) y (*c*) Ratio de Emparejamientos Fallidos (MMR[4]). Las figuras 7.6 a 7.8 muestran los resultados promedio correspondientes a los cuatro conjuntos de prueba con el aumento de falsos positivos, mientras que las figuras 7.9 y 7.10 recogen los resultados con el aumento de falsos negativos. Las gráficas correspondientes a FPR, obtenidas para el aumento de falsos negativos, no se muestran, ya que FPR en este caso siempre es nulo, como se muestra en (Dicle et al., 2013). Estos resultados son consistentes con los anteriores, mostrando la ventaja del modelo de atención visual propuesto con respecto a los otros enfoques probados.

Teniendo en cuenta estas conclusiones, finalmente se han configurado el algoritmo de seguimiento de objetos propuesto en el capítulo 6, denominado FGM, de acuerdo con el modelo de atención visual IS_S_UP_SUM. Se ha implementado FGM en C ++ sobre un Intel Core 2 Duo a 1,6 GHz con 2 GB de RAM. El tiempo medio de ejecución por imagen, para todo el conjunto

[2]False Negative Ratio

[3]False Positive Ratio

[4]MissMatch Ratio

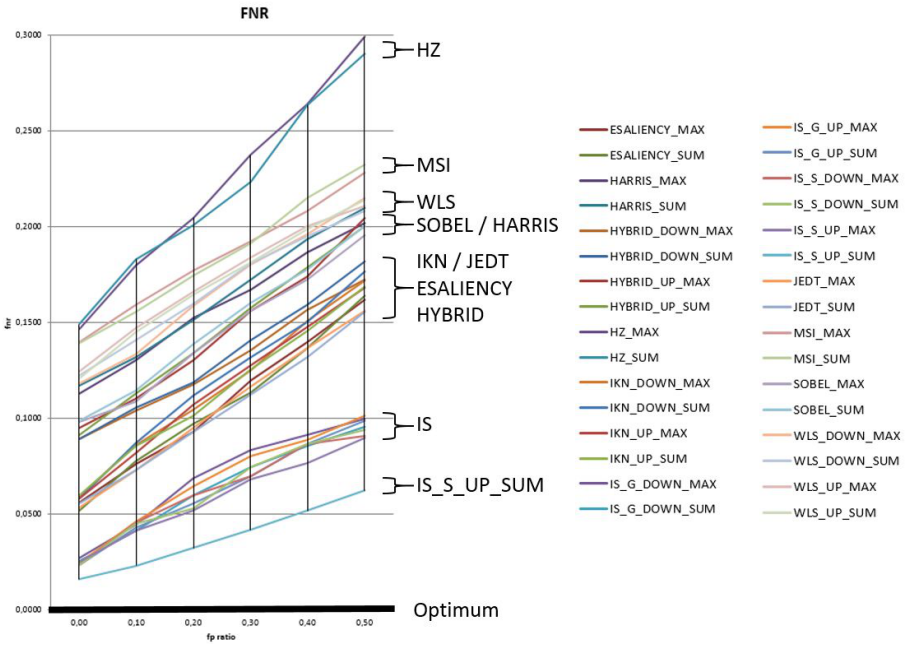

Figura 7.7: Resultados de FNR con incremento de falsos positivos.

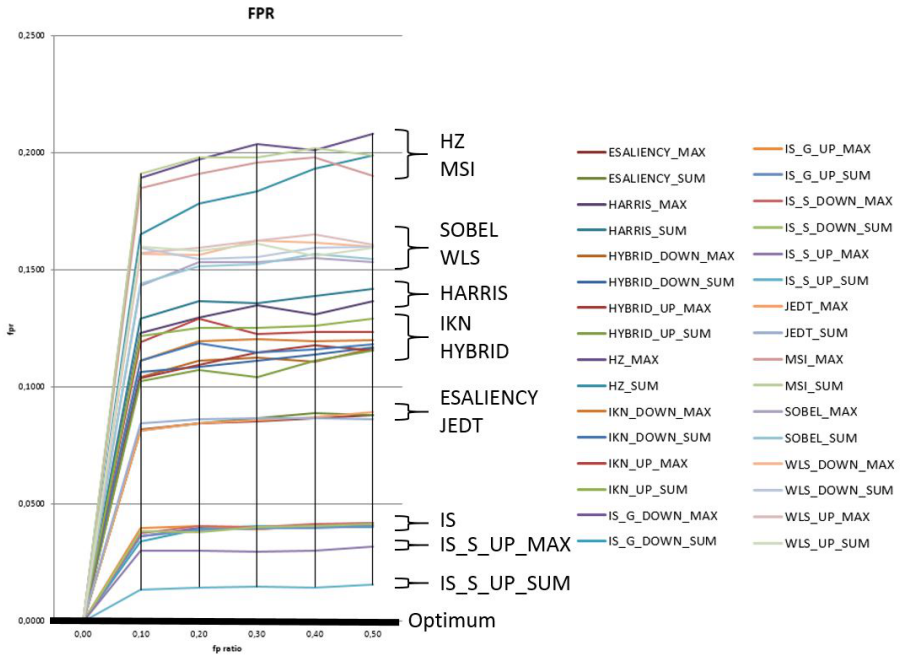

Figura 7.8: Resultados de FPR con incremento de falsos positivos.

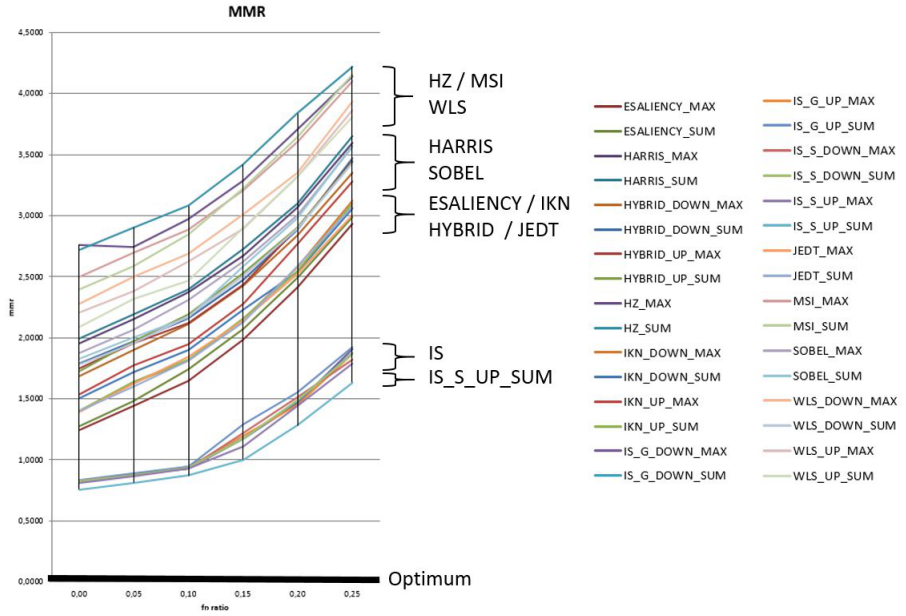

Figura 7.9: Resultados de MMR con incremento de falsos negativos.

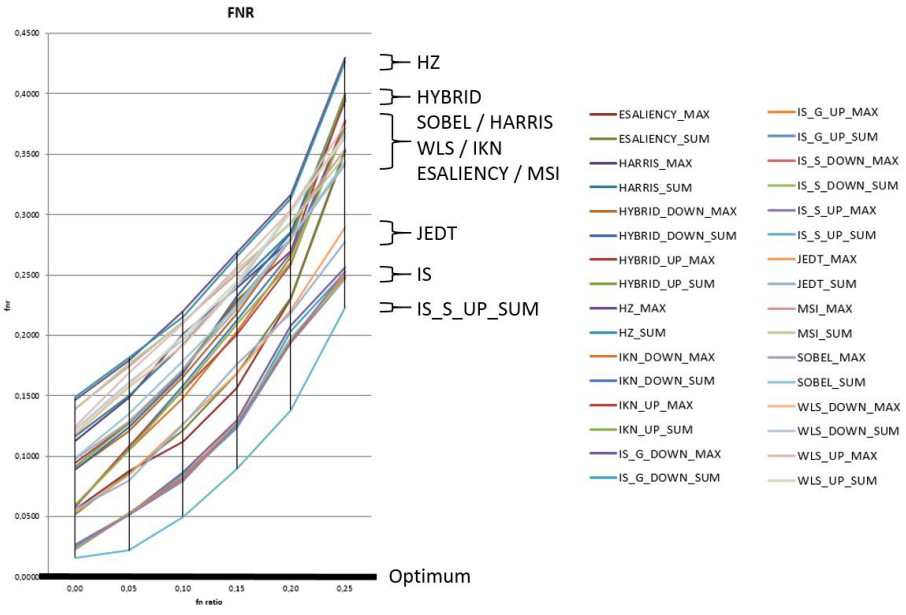

Figura 7.10: Resultados de FNR con incremento de falsos negativos.

de pruebas, considerando imágenes en color de una resolución de 720x576, con un único objeto seguido por imagen, es de 150 ms. Hay que tener en cuenta que las etapas de procesamiento directamente relacionadas con el seguimiento (caracterización, asociación, ...) son las únicas que se han incluido en este estudio de rendimiento computacional.

7.3. Rendimiento del seguimiento para diferentes algoritmos de seguimiento de objetos en secuencias de vídeo

El algoritmo simple de seguimiento mediante asociación de bloques propuesto en el capítulo 6 fue diseñado como un banco de pruebas para comparar el rendimiento de los diferentes modelos de diferencia cromática analizados, así como de los modelos de atención visual en un ámbito de aplicación práctica. Nos quedamos en ese nivel de sencillez con el objetivo de que el buen o mal rendimiento del algoritmo no fuera atribuible a la bondad intrínseca del algoritmo de seguimiento en sí.

Sin embargo, el modelo de atención visual propuesto en este trabajo puede ser incorporado en otros algoritmos de seguimiento del estado del arte actual con el fin de mejorar su rendimiento. En particular, el algoritmo de seguimiento de objetos en secuencias de vídeo desarrollado se basa en un detector de saliencia que aplica el modelo de percepción de color propuesto. Los *blobs* se asocian visualmente basándose en bloques de imagen centrados en los píxeles con mayor saliencia. Siguiendo la misma estrategia, el modelo de atención visual propuesto puede ser incorporado en algoritmos de seguimiento alternativos basados en puntos de interés, tales como SIFT, mediante el filtrado de las características de acuerdo a su saliencia. A su vez, los algoritmos de seguimiento basados en otros enfoques también pueden ser ajustados convenientemente para acomodar los mapas de saliencia precalculados.

Como prueba de concepto de las estrategias mencionadas, hemos aplicado el modelo de atención visual propuesto a dos algoritmos de seguimiento de objetos en secuencias de vídeo disponibles públicamente: la implementación piramidal del algoritmo de seguimiento de Lucas y Kanade integrado en OpenCV (Bouguet, 2000), denotado como KLT, y el algoritmo de seguimiento de muestras visuales propuesto en (Kwon and Lee, 2011), denotado como VTS. En KLT, hemos incorporado el modelo de atención visual propuesto mediante la modificación de las características a ser seguidas. En lugar de usar los puntos originales de Shi y Tomasi, calculamos la saliencia de

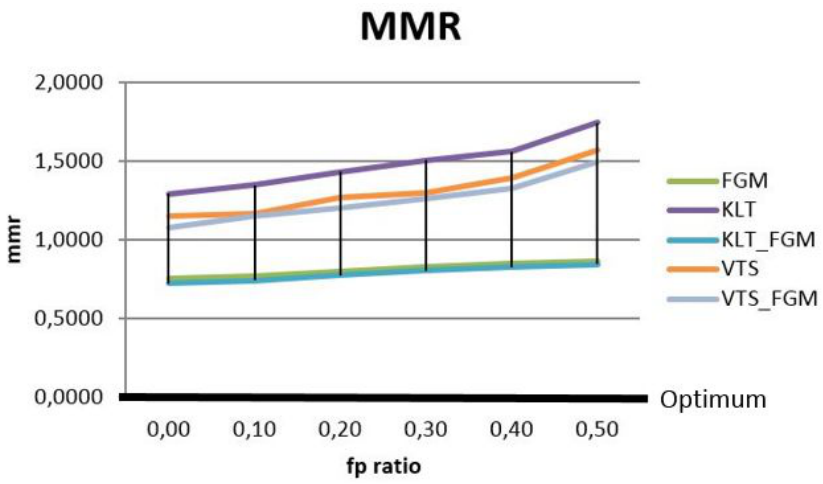

Figura 7.11: Resultados de MMR con incremento de falsos positivos.

cada uno de los elementos a seguir y seleccionamos los puntos con mayor saliencia como las características de partida para inicializar el algoritmo KLT. Este algoritmo de seguimiento mejorado se ha denotado como KLT_FGM. Alternativamente, hemos incorporado el modelo de atención visual en VTS mediante la mejora de la selección de la zona inicial a ser seguida por el algoritmo, ya que VTS requiere la selección manual del área objetivo. Utilizamos este área inicial como una semilla. A continuación, se realiza un proceso iterativo, en el cual, el área se desplaza sobre la imagen mediante la aplicación de un algoritmo de ascenso de gradiente sobre el mapa de saliencia de la imagen. El algoritmo de seguimiento mejorado de esta forma se ha denominado VTS_FGM. Los experimentos han evaluado el rendimiento de los algoritmos de seguimiento originales (FGM, KLT, VTS) y de sus versiones (KLT_FGM, VTS_FGM) alternativas.

Las figuras 7.11 a 7.13 muestran los resultados promedios correspondientes a los cuatro conjuntos de ensayos obtenidos para el aumento de falsos positivos, mientras que las figuras 7.14 y 7.15 muestran los resultados para el aumento de falsos negativos. Estos resultados demuestran que el rendimiento de los algoritmos de seguimiento originales mejora significativamente cuando el modelo de atención visual propuesto se integra en ellos, tal y como se ha descrito anteriormente, siendo esta probablemente la mayor contribución práctica de la presente tesis. Además, estos resultados también muestran que el algoritmo de seguimiento propuesto, a pesar de su sencillez, produce unos resultados significativamente comparables con el estado del arte actual.

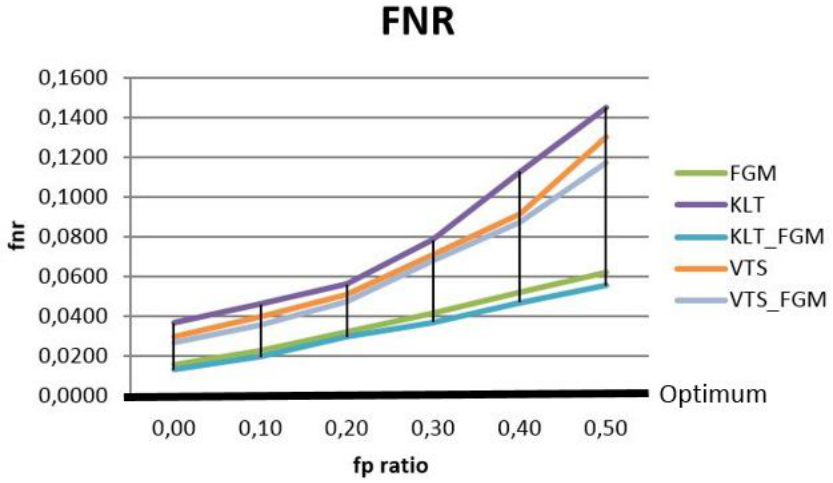

Figura 7.12: Resultados de FNR con incremento de falsos positivos.

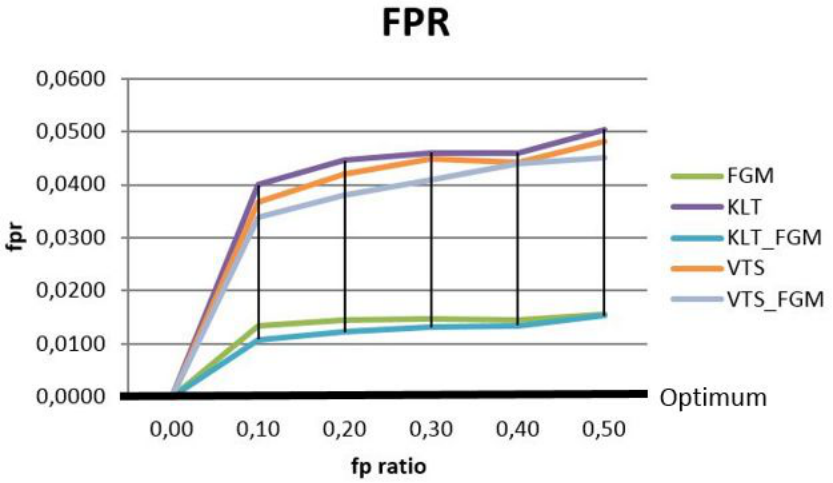

Figura 7.13: Resultados de FPR con incremento de falsos positivos.

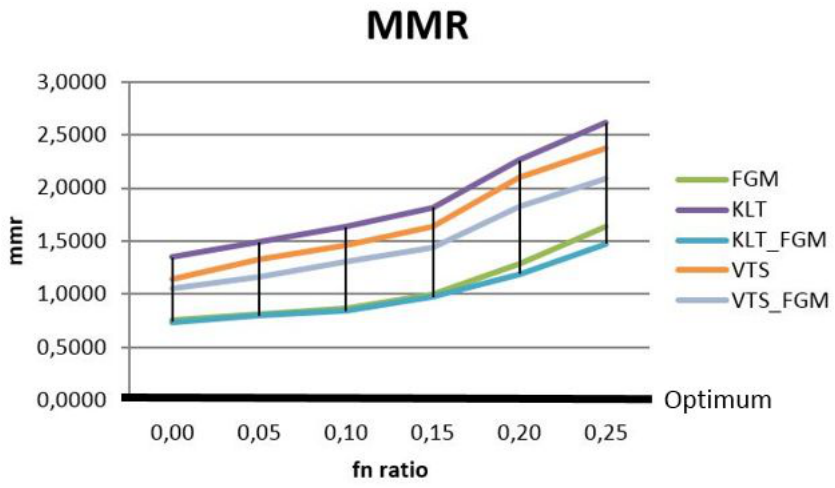

Figura 7.14: Resultados de MMR con incremento de falsos negativos.

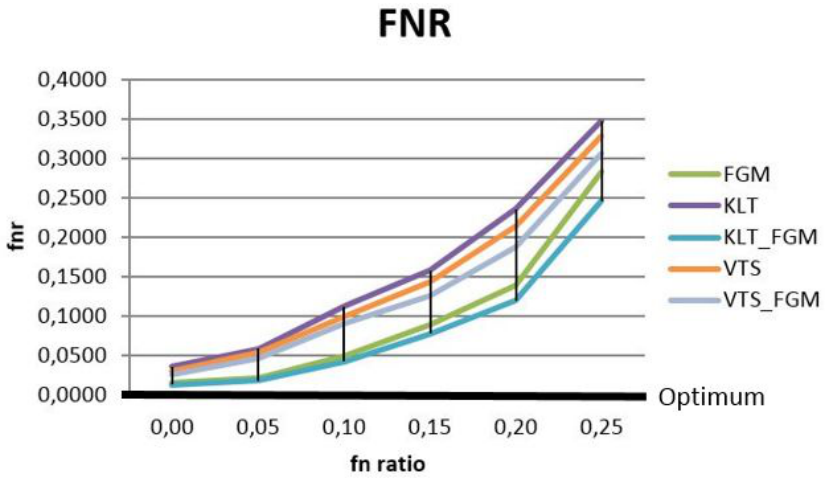

Figura 7.15: Resultados de FNR con incremento de falsos negativos.

Parte III

CONCLUSIONES

Capítulo 8

Conclusiones

Este capítulo presenta los comentarios finales de esta tesis. Está organizado de la siguiente manera. La sección 8.1 resume las contribuciones hechas a lo largo del desarrollo de esta tesis. La sección 8.2 propone futuras líneas de investigación que surgen de este trabajo. Por último, la sección 8.3 enumera las publicaciones que se han derivado de esta tesis.

8.1. Contribuciones

En esta tesis se han realizado las siguientes contribuciones a los campos del Procesamiento de Imágenes y la Visión por Computador:

1. **Modelo computacional del modelo psicofísico perceptual de Izmailov y Sokolov**: Aunque el modelo psicofísico perceptual conjunto de color y brillo de Izmailov y Sokolov es realmente innovador y representa una aproximación bioinspirada mucho más cercana a la percepción humana del color y el brillo, no había sido utilizado hasta el momento a nivel práctico. El principal problema era el estar definido únicamente como un modelo teórico y no contar con una representación adaptada al ámbito computacional. Este problema ha sido abordado en el capítulo 4, proponiendo un modelo computacional eficaz y capaz de calcularse en tiempo real. Esto es especialmente útil, ya que permite la explotación de las ventajas de este modelo perceptual, principalmente cuando se maneja conjuntamente el brillo y el color, en múltiples aplicaciones, como la ejemplificada en el modelo de atención visual propuesto (capítulo 4).

2. **Significado perceptual de las variables psicofísicas de Izmailov y Sokolov**: El mode-

lo teórico presentado por Izmailov y Sokolov (capítulo 3) consiste en un vector de cinco características difícilmente comprensibles para su utilización dentro de cualquier modelo computacional al no ser extrapolables a espacios de color o variables extraíbles de los mismos. El análisis de estas variables y su descripción en el espacio de color RGB, así como la luma, puede fomentar el uso de este modelo incluso fuera del modelo computacional descrito.

3. **Actualización del modelo de atención visual de IKN**: La arquitectura presentada por IKN (Itti et al., 1998) representa una de las piedras angulares en los modelos de atención visual, siendo un referente claro para los modelos actuales. La actualización presentada en el capítulo 5, mediante una integración perceptual del color y el brillo a través de la adaptación del modelo computacional presentado en el capítulo 4, produce unos resultados más consistentes que el modelo original de IKN. Esto contribuye indirectamente a muchas aplicaciones en el campo de la Visión por Computador, como se ejemplifica en el capítulo 6 con el seguimiento de objetos.

4. **Nuevo algoritmo de seguimiento en secuencias de vídeo**: Con el fin de evaluar objetivamente los beneficios del modelo de atención propuesto basado en el modelo perceptual de Izmailov y Sokolov, se ha propuesto en el capítulo 6 un método simple de seguimiento de objetos en secuencias de vídeo basado en bloques que son extraídos según las saliencias visuales más características de la imagen. Los resultados experimentales (capítulo 7) muestran que, aún siendo una aproximación muy simple, su rendimiento es significativamente elevado en situaciones complejas (objetos visualmente muy similares) con agrupaciones y auto-oclusiones.

5. **Mejora de los algoritmos de seguimiento de objetos usando modelos de atención visual**: Los resultados experimentales del capítulo 7 muestran que los algoritmos de seguimiento de objetos en secuencias de vídeo pueden ser significativamente mejorados incorporando el modelo de atención visual propuesto (capítulo 6) en ellos de una manera directa, siendo esta una de las mayores contribuciones prácticas de esta tesis.

8.2. Trabajo Futuro

Esta tesis deja abiertas varias líneas de investigación para su futuro desarrollo. Una vez demostrado que el modelo perceptual propuesto puede extenderse a un modelo de atención visual y su posterior aplicación al campo del seguimiento de objetos, la línea de investigación futura más importante consiste en proponer nuevas extensiones del modelo perceptual conjunto de color y brillo para otras aplicaciones. Además, se pueden derivar de este trabajo las siguientes líneas de investigación:

1. Estudio de variaciones en el cálculo de β, dentro del modelo perceptual adaptado de Izmailov y Sokolov.

2. Optimización de la técnica propuesta para que sea adecuada para el procesamiento de vídeo en tiempo real.

3. Estudio cuidadoso de las escalas y, por lo tanto, de los mapas de caracterización que se integran actualmente para formar los mapas de conspicuidad cromático-acromático y de orientación en el modelo de atención visual propuesto.

4. Estudio de simplificaciones o aproximaciones de la formulación analítica que constituyen el modelo computacional propuesto.

5. Desarrollo de un modelo paralelizable que pueda explotar sistemas de procesamiento paralelo (Big Data) o las unidades de procesamiento gráfico (GPU).

6. Mejora de otros algoritmos de seguimiento del estado del arte incorporando el modelo de atención visual propuesto.

7. Desarrollo de otras aplicaciones de alto nivel sobre el modelo de atención visual propuesto en el campo de la Robótica y la Visión por Computador, como puede ser la detección y reconocimiento de objetos, segmentación semántica, etc.

8.3. Publicaciones

Las siguientes publicaciones han sido derivadas de esta tesis:

1. El cuerpo principal de esta tesis fue publicado en Diciembre de 2016 en el volumen 60 de la revista internacional "*Pattern Recognition*"(Fernández-Carbajales et al., 2016), con un factor de impacto de 4,582 (en 2016) o 4,991 si contabilizamos los últimos 5 años.

2. En 2011, parte de las comparativas obtenidas de los modelos de atención visual fueron presentadas en el congreso internacional "*Advanced Video and Signal-Based Surveillance (AVSS)*", (Fernández-Carbajales et al., 2011). (AREA 2011 B).

APÉNDICES

Apéndice A

Listado de abreviaturas y símbolos

Los símbolos y abreviaturas utilizados a lo largo de este trabajo se recogen en la siguiente tabla:

Abreviatura/símbolo	Significado
ATA	Average Tracking Accuracy
CMOS	Complementary Metal–Oxide–Semiconductor
CCD	Charge-Coupled Device
DICOM	Digital Imaging and COmmunication in Medicine
FNR	False Negative Ratio
FPR	False Positive Ratio
GSDF	Grayscale Standard Display Function
ITU-R	Unión Internacional de Telecomunicaciones - Radiocomunicaciones
JND	Just-Notificable Difference
MMR	MisMatch Ratio
MOTA	Multiple Object Tracking Accuracy
NGL	Nucleo Geniculado Lateral
RSME	Root-Mean-Square Error
SAD	Sum of Absolute Differences
SDTV	Standard Definition TeleVision

Bibliografía

Avraham, T., Lindenbaum, M., 2010. Esaliency (extended saliency): meaningful attention using stochastic image modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 693–708.

Begum, M., Karray, F., 2011. Visual attention for robotic cognition: A survey. IEEE Transactions on Autonomous Mental Development 3 (1), 92–105.

Borji, A., Itti, L., 2013. State-of-the-art in visual attention modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (1), 185–207.
URL http://dx.doi.org/10.1109/TPAMI.2012.89

Bouguet, J.-Y., 2000. Pyramidal implementation of the lucas kanade feature tracker. Intel Corporation, Microprocessor Research Labs.

Brown, L. M., Datta, A., Pankanti, S., 2012. Exploiting color strength to improve color correction. In: 2012 IEEE International Symposium on Multimedia, ISM 2012, Irvine, CA, USA, December 10-12, 2012. pp. 179–182.
URL http://dx.doi.org/10.1109/ISM.2012.43

Burdick, H. E., 1997. Digital Imaging: Theory and Applications. McGraw-Hill.

Burton, D., 1996. Approximate RGB values for visible wavelengths, http://www.physics.sfasu.edu/astro/color/spectra.html.
URL http://www.physics.sfasu.edu/astro/color/spectra.html

Cannons, K. J., Wildes, R. P., 2014. The applicability of spatiotemporal oriented energy features to region tracking. IEEE Trans. Pattern Anal. Mach. Intell. 36 (4), 784–796.

Cavallaro, A., Steiger, O., Ebrahimi, T., 2005. Semantic videos analysis for adaptive content delivery and automatic description. IEEE Transactions on Circuits and Systems of Video Technology 10 (15), 1200–1209.

Choi, W., Pantofaru, C., Savarese, S., 2013. A general framework for tracking multiple people from a moving camera. Pattern Analysis and Machine Intelligence (PAMI).

Comaniciu, D., Meer, P., 2002. Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. 24 (5), 603–619.

CVPR06, 2006. Performance evaluation of tracking and surveillance (PETS 2006), http://www.cvg.rdg.ac.uk/pets2006/data.html.
URL http://www.cvg.rdg.ac.uk/PETS2006/data.html

Dicle, C., Camps, O. I., Sznaier, M., 2013. The way they move: Tracking multiple targets with similar appearance. In: IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013. IEEE, pp. 2304–2311.
URL http://dx.doi.org/10.1109/ICCV.2013.286

Engel, S., Zhang, X., Wandell, B., 1997. Colour tuning in human visual cortex measured with functional magnetic resonance imaging. Nature 388 (6637), 68–71.

Felzenszwalb, P., Mcallester, D., Ramanan, D., 2008. A discriminatively trained, multiscale, deformable part model. Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on.

Fernández-Carbajales, V., García, M. A., Martínez, J. M., 2011. Improving the efficiency and accuracy of visual attention. In: 2011 8th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS). pp. 349–354.

Fernández-Carbajales, V., García, M. n., Martínez, J. M., 2016. Visual attention based on a joint perceptual space of color and brightness for improved video tracking. Pattern Recognition 60, 571–584.

Frintrop, S., 2006. Vocus: a visual attention system for object detection and goal-directed search. In: IN LECTURE NOTES IN ARTIFICIAL INTELLIGENCE (LNAI. Springer.

Geusebroek, J. M., van den Boomgaard, R., Smeulders, A. W. M., Geerts, H., 2001. Color invariance. IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (12), 1338–1350. URL https://ivi.fnwi.uva.nl/isis/publications/2001/GeusebroekTPAMI2001

Gijsenij, A., Gevers, T., van de Weijer, J., 2011. Computational color constancy: Survey and experiments. IEEE Transactions on Image Processing 20 (9), 2475–2489. URL https://ivi.fnwi.uva.nl/isis/publications/2011/GijsenijTIP2011

Gordon, N. J., Salmond, D. J., Smith, A. F. M., 1993. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. Radar and Signal Processing, IEE Proceedings F 140 (2), 107–113.

Guild, J., 1931. The colorimetric properties of the spectrum. Philosophical Transactions of the Royal Society of London, 149–187.

He, R., Yang, B., Sang, N., Yu, Y., Bai, G., Li, J., 2015. Integral region-based covariance tracking with occlusion detection. Multimedia Tools and Applications 74 (6), 2157–2178.

Hou, X., Zhang, L., 2007. Saliency detection: A spectral residual approach. In: IEEE International Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07. pp. 1–8.

IEC, 1999. Multimedia systems and equipment - Colour measurement and management - Part 2-1: Colour management - Default RGB colour space - sRGB. International Electrotechnical Commission.

Itti, L., Koch, C., Niebur, E., 1998. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (11), 1254–1259.

Izmailov, C., Sokolov, E., 1991. Spherical model of color and brightness discrimination. Psychological Science 2 (4), 249–259.

Judd, T., Ehinger, K., Durand, F., Torralba, A., 2009. Learning to predict where humans look. In: IEEE International Conference on Computer Vision, 2009. pp. 2106–2113.

Kalman, R. E., 1960. A new approach to linear filtering and prediction problems. ASME Journal of Basic Engineering.

Karaulova, I. A., Hall, P. M., Marshall, A. D., 2002. Tracking people in three dimensions using a hierarchical model of dynamics. Image Vision Comput. 20 (9-10), 691–700.

Kasturi, R., Goldgof, D., Soundararajan, P., Manohar, V., Garofolo, J., Bowers, R., Boonstra, M., Korzhova, V., Zhang, J., 2009. Framework for performance evaluation of face, text and vehicle detection and tracking in video: Data, metrics and protocol. IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2), 319–336.

Khan, S. M., Shah, M., 2009. Tracking multiple occluding people by localizing on multiple scene planes. IEEE Trans. Pattern Anal. Mach. Intell. 31 (3), 505–519.

Koch, C., Ullman, S., 1985. Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. Human Neurobiology 4, 219–227.

Kumar, P., Dick, A., 2013. Adaptive earth movers distance-based bayesian multi-target tracking. IET Computer Vision.

Kwon, J., Lee, K. M., 2011. Tracking by sampling trackers. In: IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011. pp. 1195–1202.

Lucas, B. D., Kanade, T., 1981. An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2. IJCAI'81. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 674–679.

Maruta, H., Isshi, M., Sato, M., 2010. Salient region extraction based on local extrema on natural images. In: IEEE International Conference on Image Procesing, 2010. pp. 1113–1116.

NEMA, 2009. Digital Imaging and Communications in Medicine (DICOM): Part 14: Grayscale Standard Display Function. National Electrical Manufacturers Association.

Ning, J., Zhang, L., Zhang, D., Yu, W., 2013. Joint registration and active contour segmentation for object tracking. IEEE Trans. Circuits Syst. Video Techn. 23 (9), 1589–1597.

S., W., L., H., J.R., L., 2004. 2d rigid-body target modelling for tracking and identification with gmti/hrr measurements. IEEE Control Theory and Applications 151 (4), 429–438.

Shan, C., Tan, T., Wei, Y., 2007. Real-time hand tracking using a mean shift embedded particle filter. Pattern Recogn. 40 (7), 1958–1970.

Sharma, G., 2002. Digital Color Imaging Handbook. CRC Press, Inc., Boca Raton, FL, USA.

Smith, A. R., 1978. Color gamut transform pairs. SIGGRAPH Comput. Graph. 12 (3), 12–19.

Tomasi, C., Kanade, T., 1991. Detection and tracking of point features. Tech. rep., International Journal of Computer Vision.

Viola, P., Jones, M. J., 2004. Robust real-time face detection. Int. J. Comput. Vision 57 (2), 137–154.

Wen, S., Cai, Z., Hu, X., 2015. Constrained extended kalman filter for target tracking in directional sensor networks. Int. J. Distrib. Sen. Netw. 2015, 3:3–3:3.

Won, W., Lee, M., Son, J., 2008. Implementation of road traffic signs detection based on saliency map model. In: IEEE International Conference on Computer and Information Technology Workshops, 2008. pp. 374–378.

Wright, D. W., 1928. A re-determination of the trichromatic coefficients of the spectral colours. Transactions of the Optical Society 30, 141–164.

Yang, H., Lou, J., Sun, H., Hu, W., Tan, T., 2001. Efficient and robust vehicle localization. IEEE Internationa Conference on Image Processing.

Yarbus, A. L., 1967. Eye movements and vision.

Yilmaz, A., Javed, O., Shah, M., 2006. Object tracking: A survey. ACM Comput. Surv. 38 (4), 1–45.

Yilmaz, A., Li, X., Shah, M., 2004. Object contour tracking using level sets. In: Asian Conference on Computer Vision, ACCV 2004, Jaju Islands, Korea.

Young, T., 1802. Bakerian lecture: On the theory of light and colours. Philosophical Transactions of the Royal Society of London, 12–48.

Zhou, Q. M., Aggarwal, J. K., 2006. Object tracking in an outdoor environment using fusion of features and cameras. Image and Vision Computing 24, 1244–1255.