Universidade de Santiago de Compostela
Departamento de Electrónica e Computación

TESIS DOCTORAL

BLANQUEADO ADAPTATIVO DE ESCALAS ESPACIO-TEMPORALES COMO MECANISMO COMPUTACIONAL DE ATENCIÓN VISUAL DINÁMICA

Presentada por:
Víctor Leborán Álvarez

Dirigida por:
Xosé M. Pardo López
Xosé R. Fernández Vidal

Santiago de Compostela,
28 de septiembre de 2015

Dr. **Xosé M. Pardo López**,
Profesor Titular de Universidad del Área de Lenguajes y Sistemas Informáticos de la Universidad de Santiago de Compostela. Dept. de Electrónica

Dr. **Xosé R. Fernández Vidal**,
Profesor Titular de Universidad del Área de Física Aplicada de la Universidad de Santiago de Compostela. Dept. de Física Aplicada

HACEN CONSTAR:

Que la memoria titulada **Blanqueado adaptativo de escalas espacio-temporales como mecanismo computacional de atención visual dinámica** ha sido realizada por D. **Víctor Leborán Álvarez** bajo nuestra dirección en el Departamento de Electrónica e Computación de la Universidade de Santiago de Compostela, y constituye la Tesis que presenta para optar al grado de Doctor.

Santiago de Compostela, 28 de septiembre de 2015

Fdo: **Xosé M. Pardo López**
Codirector de la tesis

Fdo: **Xosé R. Fernández Vidal**
Codirector de la tesis

Fdo: **Víctor Leborán Álvarez**
Autor de la Tesis

A mis padres,
sin ellos no habría llegado,
ni siquiera comenzado.

Agradecimientos

Me gustaría expresar mis agradecimientos a todas aquellas personas que directa o indirectamente han contribuido a la realización de esta *Tesis*. En primer lugar a Diego Cabello por haberme ofrecido la oportunidad de incorporarme al Grupo de Visión Artificial hace *algunos* años. Desde entonces he participando en proyectos realmente atractivos relacionados con las últimas tecnologías de la Visión Artificial aplicada a campos tan variados como la oftalmología, la reconstrucción 3D en entornos médicos o la supervisión y control de sistemas mecánicos.

Muchas gracias a mis múltiples directores de proyectos durante estos años, Antonio Mosquera, Manolo Penedo, Xose Manuel Pardo, Xosé Ramón y a todos aquellos con los que hemos colaborado, Paco, Bashir, Gomez-Ulla, Susana, Yanai, Marta, Castor, Moski ... que siempre han puesto de su parte mucho más de lo que esperaba.

Por supuesto gracias a Carlos Acuña, por haberme permitido formar también parte de su grupo accediendo al fantástico mundo de la Neurociencia y por haberme ofrecido la posibilidad de aprender como conseguir que el trabajo salga adelante sin necesidad alguna de imponer su autoridad ni en un solo momento. Muchas gracias por todos esos momentos de discusión de los que tanta formación he recibido de modo gratuito.

Gracias también a Patricia, Roberto y Mela por dedicar parte de su tiempo a explicarme sus tan complejos experimentos con todas esas torres estereotáxicas entrecruzadas que tanto se parecen a verdaderas obras de arte, y de nuevo a Roberto por sus lecciones de guitarra y sus extensos conocimientos de caza y pesca.

No puedo ignorar de ningún modo a todos mis compañeros del laboratorio 26, que no han sido pocos, desde Víctor Brea, Fernando, Manolo Feros, Nando, Rachel, Julio y Pigaso hasta mis últimos y más jóvenes compañeros y socios, Arturo, Roi, Fer, Víctor, Juan, Diego, Jose ... pasando por los menos jóvenes Antón, David, Alex, Natalia, Luna, Dani, Javier y como no a Adrián por tener tan -ordenado- siempre nuestro laboratorio y por sus sabios consejos acerca de la optimización del tiempo de trabajo.

Gracias al ontólogo, Cris, Roberto, Dani, Bardal y muchos otros compañeros y profesores de GSI y GAC por acercarse de cuando en vez a exponer sus inquietudes y permitirme así descansar un poco la mente y mantenerme al día de las últimas novedades en varias de mis aficiones, fotografía, bolsa, tecnología... (durante la

semana y también los fines de semana). También gracias a mis ahora nuevos compañeros del CIQUS, J. Manuel, Camilo, Bea, Paul, Alex, Elías, Lucía, Eric, Tinh y Fran que me han ayudado a pasar del mundo macro al mundo nano y volver a mis orígenes como físico electrónico.

Gracias a Pablo, Llerena, J. Manuel, Rachel y Miguel por amenizar todas esas sobremesas en el comedor de matemáticas, tantos cafés, y esos resúmenes de actualidad claramente sesgados, que años después han seguido siendo sesgados por Esteban, Vanesa, Chema, Fabio, Montse y Samuel en la cocina del CITIUS.

A David García, Digi porque sus conocimientos tecnológicos aunque difíciles de extraer de su mente son inagotables, por permitirnos a todos disfrutar siempre de las últimas novedades antes de que el propio mercado supiese de su existencia.

Gracias a Rachel por el continuo flujo de información, tantas discusiones y tanto código compartido durante todos estos años, sus conocimientos extremos de Matlab y su siempre disponibilidad para crear nuevos y creativos scripts imposibles de descifrar. También gracias a Antón, y de nuevo a Xose y Pardo por extraer parte de su tan solicitado tiempo para resolver mis múltiples dudas durante la realización de este documento.

Gracias a mi familia, sobre todo a los más próximos por haber estado siempre ahí. Pila, Benigno, Felipe, Sol, Gema, Quico ... y a mi nueva familia, Mercedes, Enrique, Merce, Toño, Diana, Ludo, Xulia y Mateo y a mis ahijados Darío y Raquel porque ellos nos hacen ver de nuevo el mundo desde una perspectiva que ya casi no recordábamos. Y como no, también a los que ya no están que han sido los que han despertado en mí ese interés por las cosas sencillas que me rodean y me han enseñado a valorarlas.

Gracias a mis padres y a mi hermana Rita, ya que sin ellos nada sería posible. Sobre todo por su total dedicación y ofrecerme todo lo que han podido de su parte e incluso mucho más de lo que han podido.

Por último, y no por ello menos importante, gracias a María porque ella ha sido la otra parte de la tesis que aquí no se ve reflejada.

28 de septiembre de 2015

"Puedes llegar a cualquier parte ...
siempre que andes lo suficiente."

Lewis Carroll
Matemático y escritor británico

"Hai un rapaz na miña clase que sabe contar ata infinito
... o que pasa é que lle leva dous días."

Darío
Mi ahijado (5 años)

Índice general

Índice de figuras

Índice de tablas

Notación y abreviaturas

- `Atención:` Concepto general que tiene en cuenta todos aquello que influye en los mecanismos de selección de objetivos tanto los mecanismos *bottom-up* determinados por la propia escena, como los *top-down* guiados por información de alto nivel.

- `BD:` Base de Datos.

- `Células M/P:` Células ganglionares tipo M, Magno y P, Parvo.

- `Dirección de la mirada:` Resultado de la composición de los movimientos coordinados de los ojos y la cabeza. Da lugar a una posición fija de la escena que se está observando y se considera una vía de información del funcionamiento interno de los mecanismos de atención.

- `FFT 2D/3D:` Transformada rápida de Fourier bi/tridimensional.

- `Flecha`  `:` En este documento esta flecha se superpone a las figuras indicando la dirección del movimiento o la trayectoria realizada por algunos de los objetos presentes en un vídeo. En ningún caso esta flecha ha sido vista por los sujetos durante la experimentación, tan solo se ha añadido como elemento informativo.

- `HSV:` Espacio de color que emplea las componentes de matiz, saturación y valor.

- `ICA:` Análisis de componentes independientes.

- `IFFT 2D/3D:` Inversa de la Transformada rápida de Fourier bi/tridimensional.

- `LAB:` Espacio de color tridimensional que emplea la componente de luminancia (L) y las de oposición de color (A y B).

- `LMS:` Espacio de color relacionado con la sensibilidad de los conos de la retina a las diferentes longitudes de onda, L(larga), M(media) y S(corta).

- `Mapa de atención:` Imagen en niveles de gris obtenida a partir de la información de las fijaciones de los sujetos.

- `Mapa de características:` En los modelos descritos en este trabajo, consiste en un primer nivel de agrupamiento de los mapas de saliencia, que refleja la contribución de una determinada característica al mapa de saliencia global (color, orientación, movimiento).

- `Mapa de saliencia:` Se refiere en este trabajo a la imagen en niveles de gris obtenida a partir de un determinado modelo de saliencia, representa las regiones de la imagen con mayor probabilidad de recibir fijaciones por parte de los sujetos.

- `ORG:` Sufijo empleado en las figuras para indicar que la imagen se corresponde con un fotograma del vídeo original.

- `Proto-Objetos:` En el contexto de la atención, son las regiones resultantes de aplicar algún modelo de saliencia, que al ser agrupadas en unidades coherentes, pueden ser percibidos como objetos reales [LZX+09].

- `PCA:` Análisis de componentes principales.

- `Redes WTA:` Redes neuronales competitivas (*Winner take all*).

- `RGB:` Espacio de color que emplea los colores rojo, verde y azul como componentes de la imagen.

- `ROC:` *(Receiving Operator Characteristic)*, en el contexto de este trabajo se emplea para referirse a la medida del area bajo la curva (AUC), y no a los valores en si mismo de la curva ROC.

- `Relevancia:` Importancia de un objeto o región dentro de un contexto. Se emplea en general en el contexto de los modelos *top-down*.

- `Saliencia:` (Psicología) Capacidad de relacionar las funciones cerebrales de integración. Permite hacer una selección, entre los diferentes estímulos que recibimos, para centrar nuestra atención en la información que nos interesa, quedando los demás estímulos amortiguados o anulados.

- `Saliencia:` (Computación) Concepto asociado habitualmente al contexto de la computación *bottom-up*. Es la capacidad de un objeto o región de atraer la atención, respecto a la región que lo rodea, debido a las diferencias en la información local.
- `Spikes:` También denominados potenciales de acción, son los impulsos eléctricos que viajan a través de las membranas celulares para trasladar la información de un lugar a otro.
- `SVH`: Sistema visual humano.
- `Tareas de observación libre:` Tareas en las que al sujeto simplemente se le indica que observe las escenas sin objetivo específico.
- `USC:` Universidad de Santiago de Compostela. (En el contexto de este trabajo este acrónimo no debe ser confundido con los de las BD externas empleadas de la Universidad del Sur de Carolina y la Universidad del Sur de California).
- `2D/3D:` Bi/tridimensional.

Introducción

El sistema visual de los seres vivos y, en particular, el de los primates presenta un conjunto de características que lo hace insuperable hasta el momento en la realización de tareas visuales, tanto en términos de eficiencia y robustez, como de rendimiento en general. A pesar de ello el sistema nervioso dispone de una capacidad limitada y no es posible el procesamiento simultáneo de toda la información sensorial adquirida. Para reducir esta información, el sistema visual humano dispone de un conjunto de procesos denominado colectivamente *atención visual selectiva*, que permite filtrar la información más relevante, en función de la tarea realizada. La atención visual puede verse afectada por los propios estímulos, y también puede estar sesgada por la información de la tarea.

Con el fin de dotar a los sistemas artificiales de una habilidad semejante a la *atención visual selectiva* de los humanos, en la última década, se han dedicado múltiples esfuerzos al desarrollo de modelos computacionales de saliencia, que han sido destinados principalmente a simplificar el coste computacional de las aplicaciones de visión artificial. La obtención de un sistema artificial capaz de simular los mecanismos de atención presentes en los humanos, presenta un enorme potencial ya que las tareas de atención visual son aplicables a campos tan diversos como la detección automática de objetos y acciones [HV10, SM10, SM09b], el reescalado automático de imágenes y vídeos [WY14], la creación automática de resúmenes del contenido de escenas o incluso de vídeos completos [MLZL02], la supervisión de la calidad de contenidos multimedia [ZP90] o el apoyo en tareas de videovigilancia [MSEB15]. Con la proliferación de nuevos sistemas de interacción hombre-máquina surgen nuevos ámbitos de aplicación para las técnicas que emplean la atención visual. Algunos ejemplos son; la identificación de gestos en juegos interactivos [SM11], la saliencia multisensorial 3D [LFD15], la visualización adaptativa en dispositivos móviles [BFGP12] o la realidad aumentada [SCDM10].

Existe una enorme variedad de posibles clasificaciones de los modelos de atención [BI13], pero en general, se suelen distinguir dos categorías básicas: modelos *bottom-up* y modelos *top-down*. La principal diferencia entre am-

bos reside en que los modelos *bottom-up*, usualmente más rápidos, están guiados por características de bajo nivel, mientras que los modelos *top-down* suelen estar guiados por información de más alto nivel, que involucra elementos como la memoria, los requerimientos de la tarea o las estrategias [LBF05, SM09b]. A menudo la saliencia *bottom-up* se suele asociar a la atención involuntaria (características espaciotemporales que, sin ningún esfuerzo y de modo involuntario, atraen la atención). Esta descripción *bottom-up* para el sistema visual humano es la que propone la hipótesis de saliencia visual. En contraposición, la hipótesis de control cognitivo da mayor peso a los factores *top-down*. Algunos autores concluyen que estos dos procesos (*top-down* y *bottom-up*) son independientes y actúan en ventanas temporales diferentes [vZDT04].

A lo largo de este trabajo se presenta un modelo computacional de atención visual selectiva cuya metodología se enmarca dentro de las técnicas *bottom-up*. Este modelo, al que hemos denominado AWSD, es capaz de detectar la saliencia tanto en imágenes estáticas como en vídeo. La idea básica sobre la que se sustenta el modelo AWSD es que la saliencia, tanto estática como dinámica, se produce en aquellos puntos donde la energía local espacio-tiempo posee la máxima desviación respecto a la distribución media de esta característica en un espacio multiescala. La energía local constituye un estadístico de alto orden que concentra gran cantidad de la información perceptualmente relevante. Para acceder a ella, el modelo utiliza el blanqueado como un mecanismo muy simple que condensa parte de las implicaciones de la hipótesis de Barlow [Bar61]. La reducción de la redundancia que consigue el modelo lleva a que el AWSD utilice las características óptimas para cada vídeo, al contrario que otros algoritmos que utilizan un espacio de características fijo y subóptimo para todos los vídeos.

Motivación

El sistema visual humano dispone de múltiples mecanismos para filtrar la gran cantidad de información que llega al cerebro y así poder procesarla. Múltiples modelos computacionales de atención visual comparten este objetivo, entonces ¿por qué no emplear la biología como fuente de inspiración? El problema que se plantea al crear un modelo de saliencia *bottom-up* es el siguiente: dada una imagen o vídeo, se pretenden detectar los objetos o acciones salientes en la escena, haciéndolo del modo más preciso y sin ningún conocimiento previo acerca del contenido.

Para resolver este problema se ha desarrollado un nuevo modelo que cubre parte de las carencias de los existentes (modelos rígidos y poco adaptativos).

La validación del modelo propuesto se ha realizado tomando como punto de partida la denominada *hipótesis de saliencia visual*, que afirma que las localizaciones de las fijaciones de los sujetos durante la realización de las tareas propuestas vienen determinadas principalmente por las propiedades *bottom-up*, o sea, la saliencia visual de la escena. Bajo esta hipótesis, el control de la atención es, en su mayor parte, una reacción ante las propiedades visuales de los estímulos que se presentan al observador.

A la hora de comparar el modelo propuesto con otros existentes, se ha observado que gran parte de los trabajos del ámbito valoran los resultados de modo global, con medidas que dan una información sintética del conjunto de datos de prueba. Por todo ello, se ha optado por validar los datos empleando una metodología robusta a través del análisis ROC temporal, que ofrece una información más completa que la descrita en publicaciones previas. De este modo se pueden extraer tanto medidas globales de los conjuntos de prueba, como medidas instantáneas de cualquiera de los vídeos del banco de pruebas.

Ante las carencias de las bases de datos de vídeos disponibles a través de la red, se ha creado una nueva base de datos de vídeos propia. Además, se han extraído muestras relevantes de varios de los vídeos, en las que se puede apreciar un mejor comportamiento de la metodología propuesta frente a experimentos clásicos, ante ciertos condicionantes.

Alcance

Las variadas aplicaciones de la visión por computador hacen que los resultados de este trabajo puedan ser empleados para conseguir una mejora de la eficiencia y un mayor potencial, haciendo posibles nuevos ámbitos de aplicación. Con todos estos usos en mente se han establecido inicialmente los siguientes objetivos:

- `Creación de un modelo biológicamente plausible.` El modelo propuesto dará una solución bioinspirada efectiva para el funcionamiento de un sistema de atención visual que será biológicamente plausible e intentará reflejar las capacidades del sistema visual humano. El modelo será lo más general posible, pudiendo emplearse en diferentes ámbitos sin necesidad de reparametrización.

- `Explicación de fenómenos.` El modelo propuesto deberá ser coherente con el comportamiento del SVH frente experimentos psicofísicos, tanto relacionados con características estáticas como dinámicas, tales como efectos *pop-out*, asimetrías visuales, etc.

- `Valoración objetiva.` Con el fin de comparar la capacidad del modelo propuesto frente a otros modelos del estado del arte, en este trabajo se valorarán las diferentes métricas existentes y se elegirán aquellas de uso más extendido que describan de modo más completo los resultados de dichas comparaciones.

Para cumplir estos objetivos, el modelo deberá ser capaz de predecir las fijaciones oculares de los sujetos en tareas de observación libre, en las que la influencia de los efectos *top-down* se hayan minimizado, sobre diferentes bases de datos.

Principales aportaciones

En el presente trabajo se podrían destacar las aportaciones que se indican a continuación y que serán desarrolladas en secciones posteriores:

- `Adaptación espacio-temporal.` Proponemos un método para adaptar las escalas espacio-temporales a la estadística de las secuencias de imágenes que permite cuantificar la saliencia de fenómenos dinámicos y estáticos por su contribución a la variabilidad de la energía local en un espacio de escalas blanqueado. Según nuestra información, hasta la fecha es la primera implementación de un mecanismo de adaptación visual de corto alcance en un modelo de saliencia dinámica. Además la máxima variabilidad de la energía local espacio-temporal permite determinar la relevancia de los movimientos sin necesidad de asumir modelos explícitos de fondo, coherencia temporal, flujo óptico o medidas de movimiento relativo.

- `Arquitectura unificada.` La buena respuesta de este algoritmo sobre secuencias de fotogramas es novedosa y muestra un buen comportamiento, dando explicación a múltiples fenómenos observados y siendo de aplicación en diferentes ámbitos. Proponemos una arquitectura que conjuga mecanismos plausibles (blanqueado, codificación sensorial eficiente, análisis multirresolución y mecanismos de adaptación visual) para lograr una *solución computacional* simple tanto para la saliencia estática como dinámica. En el caso de escenas estáticas, el propio modelo lleva a soluciones en las que la componente estática es la que impera en el modelo combinado.

- `Base de datos CITIUS.` Al revisar las diferentes bases de datos de vídeos, creadas empleando seguidores oculares y disponibles a través de Internet, se ha observado una serie de carencias: i) minimización de

los efectos *top-down* (no se han incluido escenas familiares, interiores, deportes de masas, etc.), ii) un número suficiente de fijaciones en cada fotograma de vídeo (necesario para la robustez de las medidas de evaluación) y iii) la incorporación de un conjunto de vídeos sintéticos que recopila efectos *pop-out* dinámicos. Con esta nueva base de datos y las previamente obtenidas se ha podido conseguir la información necesaria para valorar el comportamiento de los modelos de atención evaluados tanto de modo cualitativo como cuantitativo. Además, se ha dado acceso público a esta base de datos, permitiendo de este modo que sea una referencia para otros autores.

- `Test de fiabilidad para la ROC y NSS.` Uno de los problemas con los que se ha encontrado el avance de este trabajo ha sido la necesidad de cuantificar la calidad de los resultados obtenidos. En la evaluación experimental proponemos incorporar un test que nos permite estudiar la fiabilidad estadística de los modelos y/o medidas a lo largo de los vídeos. Adicionalmente, adaptamos el *shuffling* para lograr que la medida NSS [PIIK05] tenga en cuenta el sesgo central y los efectos de borde. Los resultados demuestran que el modelo creado es competitivo respecto al estado del arte.

Organización de la tesis

Esta tesis se ha organizado en cinco capítulos. En el capítulo 1 se presentarán unas nociones elementales acerca de los componentes básicos del sistema visual y la relación entre los movimientos oculares y la atención visual. Se mostrará cómo una parte de estos movimientos, conocidos como sacadas, están generados y controlados por los mecanismos de atención visual, siendo su principal objetivo centrar la fóvea, región de mayor resolución de la retina, sobre estas posiciones de la escena. Se hará una revisión de los principales modelos de atención visual tanto estáticos como dinámicos y de las principales aplicaciones de dichos modelos.

En el capítulo 2 se describirá el proceso de blanqueado adaptativo y su adaptación al modelo que se ha desarrollado en el Grupo de Visión Artificial del CITIUS y, por ello, nuestra propuesta. Se indicarán las características que emplea dicho modelo; color y características espaciales por una parte y características espacio-temporales por otra. Se presentarán los fundamentos psicofísicos de la adaptación contextual que refleja el modelo y se mostrará cómo se integran en un modelo único todas estas características que permiten el procesamiento tanto de escenas estáticas como dinámicas.

En el capítulo 3 se presentarán varias bases de datos de ejemplo de uso público y nuestra propuesta, la base de datos CITIUS. Se indicará el procedimiento seguido para su construcción y se resumirán las principales características. A continuación se realizará una breve descripción, desde el punto de vista computacional, del procedimiento seguido para la clasificación de los movimientos oculares y la creación de los mapas de atención visual humanos, empleando para ello la información de las fijaciones contenida en las bases de datos de prueba. En la sección 3.3 se indicarán las diferentes métricas evaluadas para realizar la valoración del modelo propuesto en el capítulo 2 y los motivos que han llevado a la elección de las métricas empleadas en este trabajo. Por último en la sección 3.4.1 se presentarán varios problemas, descritos en la bibliografía, que afectan a múltiples modelos y que pueden dar lugar a que las metodologías de valoración empleadas den resultados inesperados (sesgo central y efectos de borde). En la última sección de este capítulo se indican las soluciones que se han empleado para minimizar dichos efectos.

En el capítulo 4 se presentarán los resultados del funcionamiento del modelo AWSD propuesto en el capítulo 2. Se emplearán las métricas elegidas de la sección 3.3 para mostrar los resultados del modelo frente a experimentos clásicos. Se reflejarán las peculiaridades que lo hacen especialmente útil para muchas de las aplicaciones descritas en la sección 1.3. Se mostrará el comportamiento del modelo tanto sobre escenas naturales como en vídeos sintéticos para así apreciar las diferencias con los demás modelos empleados como comparación. Del mismo modo se mostrará cómo el modelo reproduce resultados psicofísicos conocidos relacionados con la forma, el color y el movimiento.

Por último, en el capítulo de conclusiones se indicarán las principales aportaciones y líneas de trabajo futuras que han surgido a partir de la realización del presente trabajo.

Capítulo 1

Atención visual. Modelos computacionales y aplicaciones

Las investigaciones en el campo de la fisiología, la neurociencia, la psicología y las ciencias de la computación, han sido muy útiles para la mejor comprensión del funcionamiento del sistema visual humano (SVH), y a su vez para el desarrollo de la visión por computador.

En este capítulo se abordarán cuestiones como: ¿qué es la atención?, ¿es uno o varios procesos?, ¿hay múltiples tipos de atención visual?, ¿existe una atención unitaria que explica todos los procesos selectivos?, ¿existe alguna posición neuronal donde se pueda localizar este mecanismo?, ¿en qué se diferencian atención, saliencia y relevancia? Aunque por supuesto, parte de estas cuestiones permanecen actualmente abiertas.

Gracias a múltiples simplificaciones, estudios basados en lesiones y a propiedades fisiológicas de las células que componen diferentes áreas del cerebro se puede dar respuesta a alguna de las mencionadas cuestiones y se pueden obtener modelos que reproducen en parte el complejo comportamiento del sistema visual de los primates. Estos modelos, conocidos como modelos de atención visual han ido evolucionando y cada vez funcionan mejor en tareas de aprendizaje y reconocimiento de objetos. El número de aplicaciones de dichos modelos crece a medida que se avanza en la comprensión del SVH y dicho proceso no parece que esté cerca de su fin.

Organización del capítulo

En primer lugar se describirán los conceptos elementales relacionados con la atención visual, sección 1.1, y los componentes básicos del SVH, sección 1.1.1. A continuación en la sección 1.1.2 se presentarán varios modelos psicofísicos de atención visual, mientras que en la sección 1.2 se mostrarán varios

modelos computacionales, que mediante diferentes aproximaciones intentan resolver múltiples problemas de la visión por computador. En la sección 1.2.3 se resumirán las principales estrategias de integración del movimiento empleadas por los modelos descritos. Por último, en la sección 1.3, se hará una revisión de las aplicaciones de todos estos modelos para la realización de múltiples tareas en diferentes ámbitos.

1.1. Atención visual en Humanos

Los seres humanos se encuentran inmersos en un mundo lleno de información. La enorme cantidad de información adquirida a través del sistema visual hace imposible su procesamiento detallado y para compensar esta limitación existe en el cerebro un mecanismo biológico conocido como atención visual. Desde un punto de vista matemático se podría decir que la atención es un mecanismo que permite al SVH transformar problemas NP completos en problemas resolubles. Atendiendo a diversos factores se hablará de atención abierta o encubierta, *top-down* o *bottom-up*, interna o externa, dividida, sostenida, selectiva, espacial, temporal, dirigida a objetos, etc. Cada una de estas clasificaciones se centra en alguno de los factores que intervienen en el propio proceso de atención. Por ejemplo la atención abierta se produce cuando el sujeto dirige la mirada hacia algo que le resulta interesante, mientras que la atención encubierta es la que selecciona la parte de la escena que está siendo observada para su procesamiento, pudiendo convertirse a continuación en un nuevo centro de atención [BPE+01]. Un ejemplo muy ilustrativo de esta diferencia es la orientación de las orejas de los animales cuando perciben un sonido procedente de una determinada dirección. Esa parte correspondería a la atención abierta mientras que la selección de ese sonido de entre todos los provenientes de la misma dirección correspondería a la parte de atención encubierta [Kah73]. La lectura de este trabajo escuchando música simultáneamente sería un buen ejemplo de atención dividida. Mientras que pensar en las múltiples actividades que se podrían hacer en lugar de estar leyendo este trabajo supone un ejemplo de la atención interna, que nos permite elegir una de las posibles líneas de pensamiento en nuestro cerebro. Mantener la mirada sobre la pantalla del móvil a la espera de un mensaje de respuesta, o en la taza de café a la espera de que esté listo serían ejemplos de atención sostenida.

La pregunta que surge de modo natural ante todos estos tipos de atención es ¿qué dirige la atención?, y de ahí surge el concepto de saliencia, que proviene de la teoría de integración de señales de Treisman y Gelade [TG80]. En base a esta teoría, la combinación de un conjunto de características extraídas

de una escena produce los denominados mapas de referencia de localizaciones. Dichas representaciones, vinculadas a diferentes características, combinadas en un solo mapa topográfico, dan lugar al llamado mapa de saliencia (*saliency map*). De forma concisa, conforme a la descripción de Koch y Ulman [CS85], las localizaciones de los máximos en los mapas de saliencia serán las que guíen el foco de atención hacia diferentes posiciones de la imagen. Basándose en esta definición, el término saliencia tiene un origen *bottom-up*. Por el contrario cuando se habla de relevancia visual, se está atendiendo a la componente *top-down* de la tarea. En esa misma línea, Itti y Baldi [IB09] diferencian entre estímulos salientes, aquellos que son estadísticamente diferentes en un dominio espacial, y estímulos sorprendentes cuando son estadísticamente diferentes en un dominio temporal. En este trabajo se empleará el término saliencia refiriéndose indistintamente a ambas contribuciones.

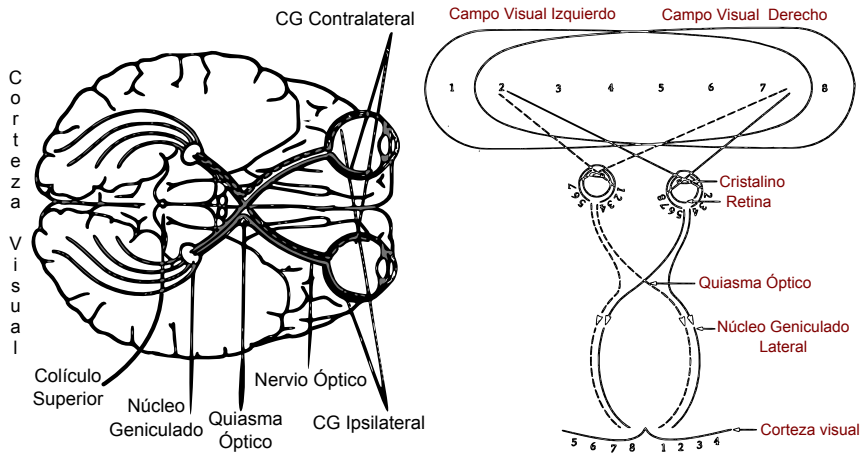
Tal como describen Borji e Itti [BI13], la atención puede ser dirigida por aquello que es más informativo [BT06, HZ08], lo que es más sorprendente [IB05, VEI12] o simplemente por el contenido que maximiza la recompensa en relación a la tarea que se está realizando (relevancia) [SB03].

Se ha indicado que hay múltiples tipos y aquello qué dirige la atención, además actualmente ya hay evidencias de que la atención no es un único proceso [CGTB11], pero ¿existe alguna posición neuronal donde se pueda localizar este mecanismo? Para responder a esta cuestión y facilitar la comprensión del funcionamiento de los mecanismos de atención referenciados a lo largo de este trabajo, se describirán brevemente los componentes más relevantes del SVH.

1.1.1. Organización del sistema de atención visual

Durante la observación de una escena, toda la información visual se recoge a través de la retina y se transmite a través del nervio óptico a diferentes centros del cerebro. Este flujo de procesamiento está sometido a múltiples interferencias que distorsionan la información recibida; aberraciones geométricas del cristalino, oclusiones debidas a los vasos sanguíneos, presencia del punto ciego en la región central de la retina o los movimientos sacádicos que alteran el contenido de la escena. Tal como se muestra en la figura 1.1a el flujo de la información visual en el cerebro da lugar a dos rutas principales, denominadas respectivamente ruta colicular y ruta retino-geniculada ya que dicho flujo se transmite respectivamente hacia el colículo superior y hacia el núcleo geniculado lateral, que forma parte del tálamo.

La ruta colicular, que se dirige del colículo superior hacia el núcleo pulvinar del tálamo, solamente transporta el 10 % de la información visual, pero presenta un papel muy importante en la atención visual y el control de



(a) (b)

Figura 1.1: (a) Componentes principales del sistema visual humano y (b) esquema de procesamiento visual, ruta retino-geniculada .

los movimientos oculares [WTS+10]. En cambio a través de la ruta retino-geniculada se transmite cerca del 90 % de la información visual hacia la corteza visual, que es la encargada del procesamiento espacial, temporal, cromático y de la disparidad [DWTS90]. En la figura 1.1b se muestra un esquema del procesamiento de la información visual a través de esta ruta. Los números 1 y 8 representan las partes del campo visual que proyectan tan sólo a la retina del ojo ipsilateral. Esa contribución es monocular debido a la presencia de la nariz que interfiere en el campo visual. El resto de los números proyectan a las dos retinas y a la corteza visual del lado contralateral.

A continuación se describirán los principales componentes del SVH que participan en la transferencia de información desde la retina hasta la corteza visual a través de las dos rutas anteriores.

La retina. Es la delgada lámina que se encuentra en la parte posterior del ojo y está compuesta por células fotosensibles de dos tipos, los conos y los bastones. La retina contiene aproximadamente $5 \cdot 10^6$ conos responsables de la visión fotópica y de color. Los bastones, cuyo número es aproximadamente 10^8, son básicos para la visión nocturna y habitualmente son descartados en estudios de agudeza visual.

Ni los conos ni los bastones están uniformemente distribuidos en la retina, existiendo una gran concentración de conos en la parte central de la retina, denominada fóvea. A lo largo de toda la periferia la concentración de conos desciende bruscamente. Debido a esta alta heterogeneidad de la distribución

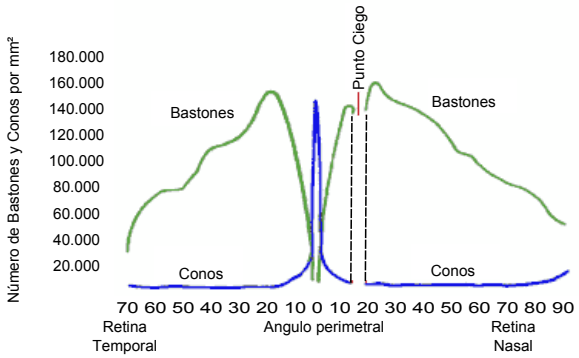


Figura 1.2: Distribución espacial de conos y bastones.

de los conos, la información de alta resolución tan solo está disponible en una pequeña parte del campo visual humano, aproximadamente 10º o sea el 2 % del campo visual [Wan95]. La figura 1.2 ilustra la distribución espacial de los fotorreceptores sobre la retina. Los fotorreceptores retinianos están conectados mediante células bipolares con las células ganglionares. Estas células pueden ser clasificadas en tres grandes categorías, las células ganglionares M (Magno), las P (Parvo) y un tercer grupo que engloba a las que no son ni M ni P. Las células ganglionares P están mayoritariamente conectadas con los conos de la fóvea y por ello están asociados con la visión central, mientras que las células ganglionares M están mayoritariamente conectadas con los bastones de la periferia y son acromáticas [LRD81].

Además de su distribución sobre la retina, las células M y P difieren entre ellas en su capacidad de resolución espacial y temporal. Por una parte los campos receptores de las células P son menores que los de las M y por ello presentan una mayor resolución espacial. Por otra parte las células M tienen una mayor resolución temporal ya que responden mejor a los cambios de patrones, mientras que las células P responden a estímulos estacionarios.

El tálamo. Está localizado en la cara anterior dorsal del cerebro medio, encima del hipotálamo y está compuesto por múltiples núcleos, entre ellos los núcleos geniculados laterales, núcleos pulvinares y núcleos reticulares. Contiene varios subsistemas que transfieren la información del campo visual proyectándola hacia la corteza visual a través de la ruta retino-geniculada.

Los núcleos geniculados laterales son la vía principal de transmisión de la información visual hacia la corteza cerebral, en ellos se encuentran tres tipos

de células, las parvocelulares, magnocelulares y coniocelulares. Cada uno de estos tipos de células alimentan un tipo diferente de flujo visual de la ruta retino-geniculada, respectivamente el flujo parvocelular, el magnocelular y el coniocelular. Las células parvocelulares y magnocelulares de los núcleos geniculados reciben aferencias de las células P y M de la retina, y del mismo modo las células ganglionares de la retina que no son ni M ni P alimentan el flujo coniocelular conectando con las células coniocelulares de los núcleos geniculados laterales [HR00].

El pulvinar es el mayor núcleo del tálamo ya que ocupa dos quintas partes de todo su volumen. Aunque no hay un acceso directo desde el nervio óptico, el pulvinar tiene conexiones recíprocas con todas las áreas corticales. Se ha asignado una gran relevancia a este núcleo en el control de la atención y de los movimientos oculares [LaB97, WTS+10].

El núcleo reticular es otro de los núcleos del tálamo que se cree que influye en los mecanismos de atención [Cri84]. Las células de este núcleo parecen modularse mediante señales de alerta, o sea señales generadas durante un estado de alerta para acelerar la selección de la información visual y con ello permitir una rápida reacción ante el peligro.

La corteza visual. Las aproximadamente 10^{10} células de la corteza visual están organizadas en un modo jerárquico. El área *V1*, también llamada corteza estriada está en la base de la jerarquía, puesto que representa la mayor entrada de la información visual procedente de los núcleos geniculados laterales a la corteza visual. Hay que destacar que el 50 % del área de *V1* está dedicada a la representación de la información adquirida desde la parte central de la retina, la fóvea. Esta región es una de las candidatas a contener una implementación neuronal del mapa de atención [GKG98, Li02].

Tal como se muestra en la figura 1.3, a partir de *V1* la información visual es transmitida a *V2* y desde ahí surgen dos vías diferentes, ya que el flujo parvocelular alimenta a *V4* mientras que el flujo magnocelular alimenta al área temporal central (emphMT). La región *V4* se asocia con la síntesis de formas complejas y a la combinación de color y forma [NSRM13, SKR13]. *V4* transmite la información visual procesada a través del flujo parvocelular a la corteza temporal inferior, que representa la última región visual para el reconocimiento de objetos, de ahí que ese flujo sea más conocido como el flujo -qué- [GM92].

El area *MT* representa una ruta importante para la información visual hacia la corteza parietal posterior (PP), que es responsable de la localización de los objetos, de ahí que a esta ruta se le conozca como el flujo -dónde- y tenga un papel muy importante en el control de la atención visual [GM92].

En estudios basados en lesiones se observa que los pacientes con lesiones en la región *MT* sufren aquinetopsia, una patología que consiste en la incapacidad de detectar el movimiento. Los pacientes perciben el movimiento como fotogramas independientes, por lo que no son capaces, por ejemplo, de calcular la velocidad de un vehículo al aproximarse.

El colículo superior y el campo frontal ocular. Estos centros cerebrales juegan un papel importante en el control del la atención visual y los movimientos oculares. El colículo superior está localizado en la superficie dorsal del tronco encefálico. Recibe las entradas visuales directamente desde la retina y desde la corteza visual, siendo las entradas de la retina principalmente células ganglionares tipo M. El colículo superior tiene proyecciones hacia numerosos núcleos del tálamo que a su vez tienen proyecciones a área corticales implicadas en los movimientos oculares sacádicos. Hay varias hipótesis que apuntan a que el colículo superior contribuye en gran parte a la construcción del mapa de atención global [DWTS90, FM06].

Por otra parte el campo ocular frontal que está localizado en el lóbulo frontal del cerebro, está implicado en la generación de los movimientos oculares. Recibe aferencias desde diferentes áreas de la corteza visual y está directamente conectado con los centros oculomotores. Algunos trabajos hacen sospechar que el campo ocular frontal está relacionado con los mecanismos de control de la atención [TS99].

Áreas *MST* y *7a*. Las Areas *MST* y *7a* localizadas en el lóbulo parietal, mostradas en la figura 1.3 y descritas en [FVE91, KSJ+00], se han relacionado con la identificación de patrones de movimiento. Por una parte, el área *MST* muestra ser selectiva ante patrones complejos de movimiento, tales como expansiones, (aumento del nivel de *zoom*), contracciones (disminución del *zoom*), y rotaciones. Por otra parte el área *7a* muestra selectividad ante varios patrones de movimiento: la traslación y el movimiento espiral como sucede con *MST*, la rotación de la escena completa y los movimientos radiales (expansión y contracción) pero con mayores campos receptores [MBMG04]. Las características de las regiones *MST* y *7a* sugieren la existencia de procesamientos relevantes en relación al modelo presentado en este trabajo, y en conjunto sobre los mecanismos de atención visual.

Interconexión entre fisiología y modelos de atención

El sistema visual está organizado de modo que tan solo se percibe a una gran resolución una pequeña porción del campo visual, siendo esta parte de la retina la fóvea. La mayor capacidad de procesamiento de la corteza visual se

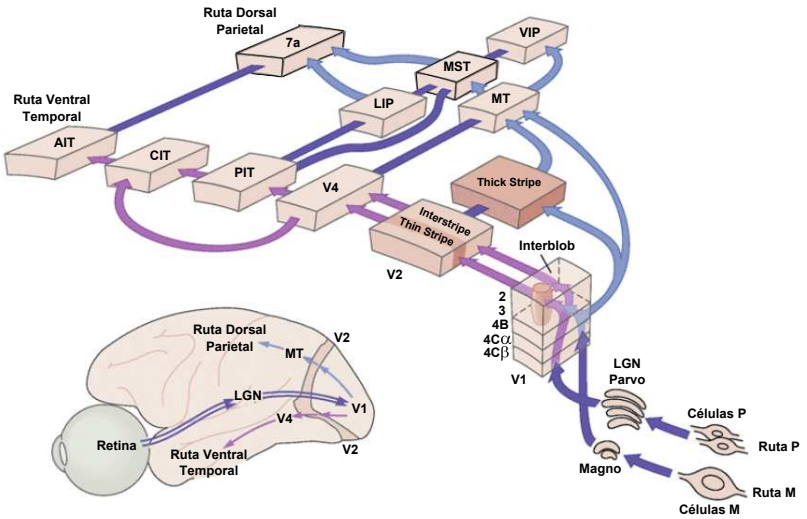


Figura 1.3: Interconexiones entre las rutas principales [KSJ+00]. Se ha simplificado el gráfico por lo que múltiples conexiones no aparecen.

dedica a la información visual procedente de la fóvea. Por ello para analizar todo el campo visual es necesario un desplazamiento de la fóvea que permita cubrir las regiones más importantes de la escena.

A lo largo de las rutas visuales del SVH de los primates se han localizado múltiples mecanismos de adaptación contextual, desde los fotorreceptores y las células ganglionares de la retina hasta las células de la corteza estriada y extraestriada [RR09, CWS+07]. Todos estos mecanismos parecen proporcionar una decorrelación entre las respuestas neuronales que permite mejorar la eficiencia de la representación [Koh07].

Por otra parte, la participación de diversas áreas del cerebro en la identificación de contrastes de movimiento, ha sido observada en áreas como *V1*, *MT*, *MST*, *7a* [TPL+02]. Estas áreas presentan una selectividad a diferentes velocidades y orientaciones permitiendo identificar contrastes de magnitud y fase del movimiento.

La existencia de una implementación neuronal del mapa de atención permanece abierta desde que, en 1985, Koch y Ulman [CS85] sugirieron como posible ubicación a los núcleos geniculados laterales del tálamo. Otros múltiples candidatos han sido propuestos: el pulvinar, el colículo superior, *V1*, *V4*, el cortex parietal. Incluso es posible que exista un único mapa que se obtenga a partir de propiedades de la imagen extraídas directamente de la región *V1* [Hen03], o quizás son múltiples las regiones que contribuyen a la construcción de múltiples mapas, conocidos como mapas de prioridad, que combinan tanto fuentes *top-down* como fuentes *bottom-up* [FM06][Nie07].

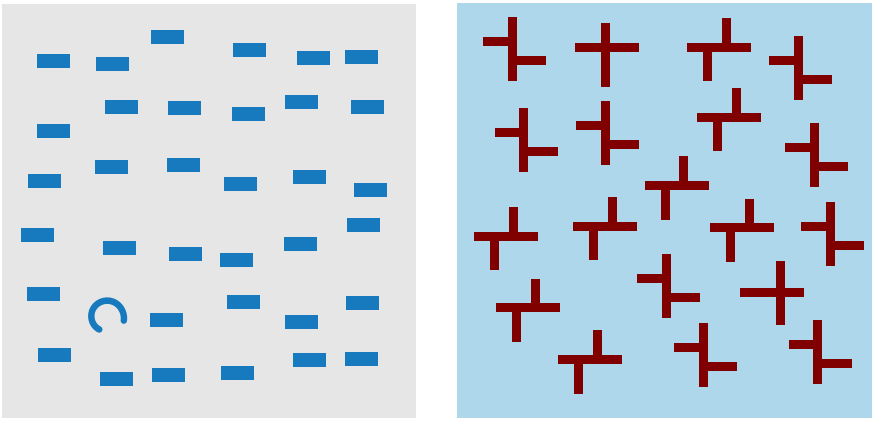
El SVH es realmente complejo y a medida que se avanza de áreas meramente sensoriales (*V1* para estímulos visuales o *A1* para los auditivos) a áreas más integradoras (corteza parietal posterior, temporal o premotora), las respuestas son más contextuales y complejas integrando información multisensorial. Incluso es posible que las mismas áreas realicen tareas diferentes en función de su implicación en un determinado contexto. Este comportamiento que ha sido observado con fuentes de información como la auditiva o la somestésica, también ha sido observado en áreas como la corteza premotora ventral [PVLA09], en la que las mismas neuronas recuperan información de la memoria, integran y comparan información basándose para todo ello en información visual.

Gracias a múltiples simplificaciones, estudios basados en lesiones y a propiedades fisiológicas de las células que componen cada una de estas áreas del cerebro, se pueden obtener modelos que reproducen en parte el complejo comportamiento del sistema visual de los primates.

1.1.2. Modelos psicofísicos de atención visual dinámica

Son múltiples los modelos que han intentado dar una explicación plausible a una gran variedad de observaciones psicofísicas acerca de la atención visual en los humanos [SS77, TG88, Ros99, HH05a], concepto que pertenece a la literatura clásica de la psicología [Jam50]. Una de las primeras observaciones, que posteriormente fue explicada mediante la teoría de Treisman, consistió en un experimento de búsqueda visual en el que una serie de objetos conocidos estaban entremezclados con varios distractores. En este experimento se analizaba el tiempo de reacción y se comprobó que en ciertos casos, como el de la figura 1.4a el tiempo de reacción era constante independientemente del número de distractores (efecto *pop-out*) mientras que en otros casos, como el que se muestra en la figura 1.4b, el tiempo de reacción aumentaba de forma lineal con el número de distractores. Estímulos claramente diferentes a los que lo rodean, en tan sólo una determinada característica, atraen nuestra atención de modo inmediato, sin necesidad de examinar detalladamente la escena e independientemente del número de distractores (*pop-out*). Por el contrario cuando los distractores son claramente heterogéneos, o cuando el objeto buscado difiere en múltiples características de los objetos que lo rodean los sujetos necesitan examinar los objetos de la escena de uno en uno.

De las publicaciones iniciales de Treisman et al. se puede extraer un conjunto de características identificadas por el SVH: la longitud de las líneas, la orientación, el contraste, el color, la curvatura y el cierre. Siendo algunas de estas propiedades asimétricas (por ejemplo, una línea inclinada frente a muchas verticales genera saliencia, mientras que lo contrario no). Otras



(a) (b)

Figura 1.4: Experimentos de *pop-out*, objeto buscado (a) con una característica diferente, (b) mezclado con distractores heterogéneos.

características fueron posteriormente añadidas a esta lista; el número de elementos, las intersecciones, los terminadores, la intensidad, el parpadeo, la dirección del movimiento, la profundidad debida a la visión estereoscópica, la dirección de la luz, etc. y posiblemente esta lista continúe creciendo a medida que aumentan los conocimientos acerca del SVH.

A raíz de todos estos experimentos se planteó la existencia de dos fases durante el proceso de atención. Gran parte de los modelos existentes comparten esta idea de que los mecanismos de atención consisten en dos etapas independientes, una etapa de atención temprana (*preattentive*), que opera de modo paralelo sobre todo el campo visual y una etapa posterior de atención (*attentive*), que es la responsable de procesar unas pocas regiones de la escena de modo secuencial [FBR05]. Estas dos etapas demuestran ser altamente eficientes ya que los recursos de mayor consumo, en los que se emplean operaciones complejas y robustas, se invierten tan solo en el análisis de regiones candidatas previamente seleccionadas durante un procesamiento paralelo menos costoso y más impreciso. En este aspecto algún trabajo pone límites a esa capacidad de procesamiento paralelo durante la etapa inicial, demostrando que la capacidad de identificación en paralelo puede verse afectada durante la realización de tareas duales de búsqueda de caracteres. Esto indica que los objetivos de la tarea pueden modular la capacidad de procesamiento de esa etapa inicial [JCN97].

En este sentido se pueden encontrar dos corrientes que proponen teorías contrapuestas; conforme a las teorías de *selección temprana*, debido a la limitación de los recursos disponibles, los mecanismos de atención son los que

permiten seleccionar los estímulos que posteriormente serán supervisados. Esta selección sucede después de la identificación de las características físicas de los estímulos y después de la separación de los objetos del campo visual del fondo de la escena. Por el contrario, las teorías de *selección tardía* afirman que todos los objetos pueden ser procesados de modo automático en paralelo sin necesidad de ningún mecanismo de selección y es al final, tras la interpretación semántica, cuando el mecanismo de selección actúa permitiendo obtener la información necesaria de cada uno de los objetos presentes en la escena [Laa99].

En general, la mayor parte de los modelos sugieren que las interacciones competitivas son la base de los efectos observados en atención visual en humanos. Incluso hay un cierto acuerdo en el hecho de que los desequilibrios durante el proceso competitivo provocados por lesiones pueden explicar ciertos trastornos neurológicos [HH05a]. Entre los modelos psicofísicos más relevantes, usados como fundamento para múltiples implementaciones de modelos dinámicos, cabe destacar los siguientes:

Modelo spotlight Posner et al. [PSD80]. Este modelo de 1980 se fundamenta en que la información de una región del campo visual se selecciona del mismo modo que cuando se enfoca dicha escena con un haz de luz, de ahí su nombre. Esta selección es realizada durante un proceso de atención encubierta que ocurre cuando los ojos están en una posición estacionaria. Es un proceso diferente al que se produce cuando se fija la atención en un punto [CB99]. Este foco se dirige hacia una determinada región del campo visual de tal modo que la calidad de la percepción es amplificada en dicha región. En esta teoría se presupone que no existe procesamiento pre-atencional en la región que está fuera del foco de atención.

Modelo texton de Julesz et al. [JGS+73]. Fue propuesto en un trabajo cuyo objetivo era identificar patrones presentes en las texturas, que puedan ser localizados [Tre85]. Para ello se definen básicamente tres tipos de estructuras: terminadores, segmentos entrecruzados y *blobs* alargados con sus propiedades de color, orientación y grosor. El modelo *texton* se basa en las capacidades del sistema visual humano para identificar de modo inmediato diferencias de características en elementos de textura (*textons*) durante la fase de atención temprana. Esta hipótesis afirma que solamente una diferencia en los elementos de textura o en su densidad puede ser identificada de modo preatentivo y, del mismo modo que Treisman, afirma que estos procesos ocurren de modo paralelo, mientras que la atención focalizada ocurre de modo secuencial.

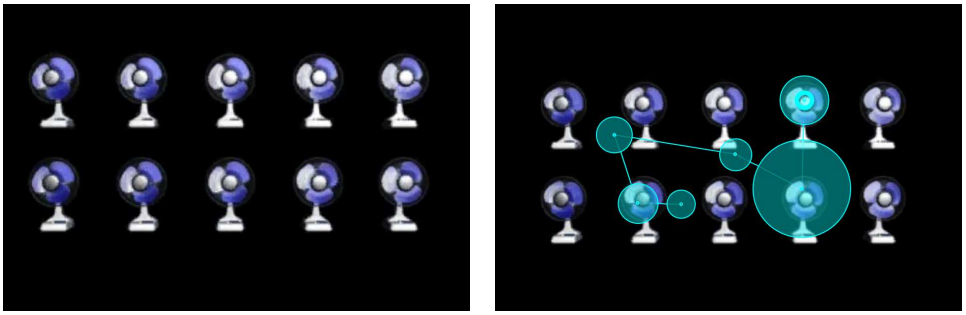
Modelo zoom-lenz de Eriksen y Yeh [EY85]. Mediante este modelo se describe la influencia de la distancia entre el objeto buscado y los distractores estableciendo una analogía con una lente tipo *zoom*, en la que los distractores, fuera del centro, se encuentran desenfocados en la zona de penumbra pero siguen interviniendo en cierto modo en el proceso de búsqueda. Es en cierto modo una evolución del modelo de *spotlight* en el que además se contemplan las diferentes escalas mediante la modificación del tamaño del foco. Es interesante destacar que en este modelo aunque la información a priori pueda facilitar la detección de un determinado objeto no inhibe el procesamiento de los demás objetos.

Modelo del mapa booleano de Huang et al. [HP07, HTP07]. El modelo ofrece una visión del comportamiento del SVH durante la realización de tareas de búsqueda guiada, dividéndolas en dos etapas; la de selección y la de acceso. La etapa de selección es la que determina que objetos son relevantes frente a los que no lo son, durante esta etapa se escogen ciertos objetos de la escena mientras que durante la etapa de acceso se determina que propiedades de los objetos elegidos serán las que se procesen.

Además de los modelos aquí descritos, en la sección 1.2 se describirán varios modelos computacionales bioinspirados, algunos de ellos relacionados con los modelos psicofísicos de esta sección. Estos modelos también ofrecen una explicación a múltiples fenómenos psicofísicos observados en el comportamiento del sistema de atención visual de los humanos.

1.1.3. Relación entre movimientos oculares y atención visual

Al observar una escena, los sujetos realizan un desplazamiento de la fóvea, región de alta resolución, hacia diferentes posiciones de la escena que atraen su atención. Precisamente los movimientos sacádicos son los responsables de este desplazamiento. Una vez que se han producido estas sacadas, el objetivo puede ser procesado con mucho mayor detalle durante cierto tiempo, conocido como tiempo de fijación. De este modo los sujetos al explorar una escena realizan secuencias de sacadas sucesivas deteniéndose en ciertos destinos, creando así una ruta de exploración, tal como la que se muestra en la figura 1.5b durante un experimento de observación libre. Las líneas que unen los centros representan las sacadas mientras que los círculos de diferentes tamaños representan la duración de la fijación en cada punto de fijación. Sin embargo durante estos períodos de fijación, los ojos no permanecen estáti-



(a) (b)

Figura 1.5: (a) Imagen extraída de un experimento de la USC de observación libre, y (a) la trayectoria de exploración de un sujeto superpuesta a la imagen original

cos, ya que hay microsacadas que parecen jugar un papel importante en el mantenimiento de las respuestas de las células de la vía visual. Si los ojos permaneciesen estáticos, la percepción de la imagen se desvanecería, al igual que sucede con el contacto de nuestra piel con la ropa que deja de percibirse pocos intantes despues de ponerla.

La relación existente entre los movimientos oculares y la atención visual se remonta a los primeros trabajos de A. Yarbus en los años 60 [TWK+10, Yar67] y posteriormente fue demostrada mediante estudios fisiológicos [MF01], conductuales [RRDU87, MRW97] y técnicas de análisis de imagen cerebral [BPE+01]. En cierto modo los movimientos oculares ofrecen una ventana hacia la comprensión del funcionamiento de los mecanismos de atención.

Los mecanismos de atención permiten examinar múltiples objetivos y elegir el más relevante, realizando en ese caso la correspondiente sacada hacia dicho objetivo para poder así revisarlo con mayor nivel de detalle. Se ha comprobado que la atención se desplaza hasta cuatro veces antes de que se produzca el movimiento de sacada [Mil93]. Esta relación entre el desplazamiento de la atención y los movimientos oculares se ha formulado mediante el modelo de EMMA (*Eye Movemente and Movement of Attention*), modelo que considera el mecanismo de atención visual como un módulo capaz de predecir los movimientos oculares [Sal00]. También hay trabajos posteriores que afirman que aunque ambos elementos, atención y sacadas están relacionados, no están estrictamente vinculados, pudiendo ser diferentes los módulos que guían el control de los movimientos y la atención [RRP+03]. Puesto que el número de sacadas que realiza un sujeto, durante tareas de exploración, es del orden de 15.000 por hora, es razonable pensar que existan mecanismos de bajo coste computacional que guían estos movimientos oculares.

Además, el hecho de que los objetos interesantes de una escena en general sean visualmente salientes permite a los modelos de atención basados en información de saliencia predecir las fijaciones de los sujetos [EI08]. Incluso las primeras fijaciones sobre una escena, cuando se realizan tareas guiadas por objetivos, pueden ir en la dirección del objeto buscado, aun cuando no está presente en la escena. Esto es debido a que los conocimientos previos pueden guiar los mecanismos de control hacia posiciones candidatas adecuadas. Estos sistemas de control además de emplear la información visual disponible tienen en cuenta otros factores, como la memoria a corto plazo que puede contener información previamente procesada de la escena actual, información semántica, espacial o visual de largo plazo de escenas semejantes previamente observadas o los objetivos y planes del propio sujeto. De hecho en tareas guiadas se observa que, cuando las escenas son más significativas para los sujetos, esta fuerte relación entre las posiciones de las fijaciones y el mapa de saliencia disminuye [OTCH03].

En este trabajo se han elegido tareas de observación libre en las que los sujetos no están focalizados en ningún objetivo en particular, intentando así minimizar la influencia de los efectos *top-down* y por ello favoreciendo una buena correlación entre los movimientos oculares, las fijaciones de los sujetos y los mapas de saliencia.

En cuanto a la duración de las fijaciones se ha visto que está influenciada por múltiples factores como la luminosidad de las escenas, el contraste o el color, siendo por ejemplo mayor la duración de las fijaciones para fotografías en color que para dibujos lineales en blanco y negro. Tener en cuenta la duración de las fijaciones, ponderando la contribución de las posiciones de fijación, puede cambiar de modo significativo las distribuciones de las regiones de atención [Hen03]. Además, en el caso de tareas guiadas si éstas requieren el empleo de la memoria, la duración también aumenta al igual que sucede si las regiones en las que se producen las fijaciones presentan un menor contraste. Esto sugiere que dicha duración está condicionada por la información extraída de las características de la propia fijación.

1.2. Modelos computacionales de atención visual

Existen extensas revisiones que describen los diferentes modelos computacionales de atención visual y sus aplicaciones a la visión artificial [FRC10, Toe11, BSI12]. Algunos de estos modelos se han inspirado en la biología, otros son meramente computacionales y en otros casos ambas estrategias se

fusionan en el mismo modelo. En general, estos modelos pueden ser clasificados dependiendo de la hipótesis básica que asumen [Tso11]. En la vida real, la información *top-down* de alto nivel (el objetivo de la tarea, las expectativas o las emociones) interfiere con la información *bottom-up* relacionada con las características de la propia imagen dando lugar a los generalmente llamados modelos de atención. Cuando se habla de modelos de saliencia, se suele referir a aquellos modelos que parten de la información intrínseca de las imágenes o vídeos, o sea, a los modelos *bottom-up*.

En este trabajo nos centraremos en los modelos espacio-temporales, i.e. aplicados sobre vídeo, basados en mapas de saliencia. Este subgrupo tiene su origen en la teoría de integración de características de Treisman et al. [TG80], y el vínculo común es la presencia de mapas asociados a ciertas características de los estímulos (color, forma, textura, movimiento, etc). Es interesante indicar que estos modelos en general no diferencian entre la atención abierta y encubierta. La primera se produce cuando el sujeto centra la información presente en la escena mediante los movimietos oculares, mientras que la segunda es la que tiene lugar en el cerebro durante el proceso de selección de regiones relevantes sin que se produzca necesariamente una fijación sobre ellas. El empleo del mapa de saliencia refleja la contribución de ambos tipos de atención sin distinción.

Una de las primeras arquitecturas computacionales de atención visual *bottom-up* propuesta por Kock y Ullman [CS85] tiene su origen en la teoría descrita por Treisman [TG80]. Esta teoría se basa en la proposición de un procesamiento paralelo inicial simple de características integrales, capaces de capturar la atención, frente a un procesamiento serie secuencial, necesario para detectar agrupaciones de características. Esta descripción lleva a la explicación de ciertos efectos de *pop-out* observados durante la realización de experimentos de búsqueda semejantes a los descritos en la sección 1.1.2.

Múltiples modelos basados en la teoría de Treisman presentan un paso intermedio entre los mapas de características y los mapas de atención conocidos como los mapas de visibilidad. Estos mapas de visibilidad resaltan las partes de la escena que difieren en gran medida de su entorno en alguna de sus características .

Una característica común a gran parte de los modelos de atención es el alto coste computacional asociado. Al igual que en los modelos psicofísicos se tiende a considerar la optimización del coste energético, en los modelos computacionales se tiene muy en cuenta este coste computacional. En general las tareas visuales suelen tener un alto coste computacional asociado, debido a la gran cantidad de información que se puede adquirir mediante la imagen. Además dicho coste aumenta a medida que los modelos son más independientes de factores como la luminosidad, las escalas o las poses.

Puesto que una imagen estática puede ser vista también como un vídeo en el que el contenido no cambia a lo largo del tiempo, se pueden emplear de modo directo modelos dinámicos sobre imágenes estáticas. De modo análogo, para aplicar un modelo estático sobre un vídeo se puede utilizar el modelo estático fotograma a fotograma o por el contrario elegir un modelo dinámico en el que específicamente se procesan los datos en el dominio espacio-temporal. A continuación se describirán varios de los modelos de atención estáticos y dinámicos más relevantes.

1.2.1. Modelos estáticos

En la figura 1.6 se muestra un esquema temporal con varios de los modelos estáticos más significativos, también denominados modelos espaciales. Inicialmente fueron propuestos como modelos de atención para imágenes estáticas, aunque algunos de ellos han sido directamente aplicados sobre escenas dinámicas.

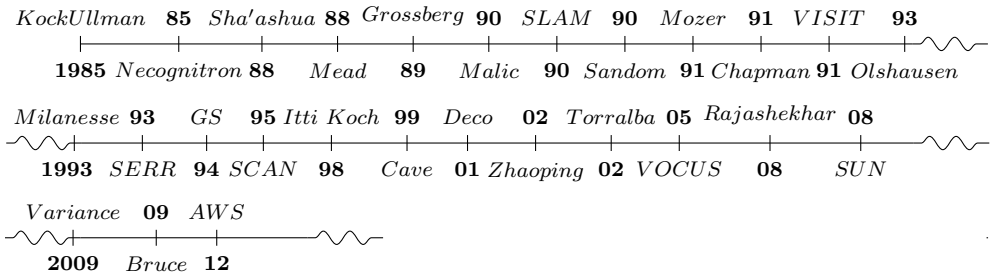


Figura 1.6: Evolución temporal de los principales modelos estáticos relacionados con atención visual.

Parte de los modelos que se describen a continuación integran múltiples características de la imagen y algunos además integran información 2D con información 3D de distancia. Modelos como los de Koch y Ulman [CS85], Itti y Koch [IKN98, IK01] o el modelo AWS [GDLFVP12] se pueden agrupar en lo que se denominan modelos basados en la hipótesis del mapa de saliencia, que tiene su origen a su vez en la teoría de integración de señales de Treisman [TG80]. Este tipo de modelos suelen contener varios componentes esenciales; una serie de características asociadas a diferentes propiedades de los estímulos, mecanismos del tipo centro y alrededores o diferencias de gaussianas que permitan determinar las características más relevantes frente al contenido que las rodea, algún mecanismo de normalización que permita combinar los mapas asociados a características diferentes y por último un mapa de saliencia en el que se codifica mediante el nivel de gris de la imagen

la probabilidad de que se produzcan fijaciones en cierta posición espacial. En algunos casos existe una etapa a mayores que consiste en la conversión del mapa de saliencia en un listado temporal de posiciones de fijación, para lo que se pueden emplear redes WTA (*winner-take-all*) combinadas con mecanismos de inhibición de retorno que evitan revisar posiciones que ya han sido previamente fijadas.

Por otra parte están los modelos que emplean una aproximación estadística al problema de la atención. Se ha observado que por ejemplo, las frecuencias espaciales o las densidades de aristas son mayores en las posiciones de fijación [MRW95], o que el contraste local presenta un mayor valor en dichas posiciones [DE03]. Mediante el análisis de estas y otras propiedades locales de las imágenes en las regiones próximas a las posiciones de fijación se pueden crear modelos que predigan las posiciones candidatas a recibir fijaciones. También hay modelos que emplean la información global en los que la saliencia de cada pixel se mide como la distancia en el espacio LAB al valor promedio de toda la imagen [AHES09].

A continuación se describen brevemente los modelos incluidos en la figura 1.6, que son una muestra representativa de los modelos estáticos desde los años 80.

Modelo de Koch y Ulman [CS85]. Es un modelo bioinspirado, fundamentado en la teoría de Treisman, define dos etapas de procesamiento; una de atención previa y otra de atención secuencial. Es un modelo puramente *bottom-up*, guiado por los datos, por lo que los saltos en la atención se producen únicamente debidos a la información contenida en las imágenes. Se basa en las siguientes premisas: los mecanismos de atención visual operan sobre imágenes con múltiples características, la información de saliencia de una región se puede representar mediante un campo escalar que se denomina mapa de saliencia, esta información está relacionada con el contexto que la rodea y las redes WTA y la inhibición de retorno son mecanismos adecuados para guiar la atención.

Modelo Necognitron de K. Fukushima [Fuk86]. Este modelo basado en el uso de redes neuronales es uno de los primeros que incluye el concepto de atención. Aunque no es un modelo puramente bioinspirado muestra como los mecanismos de atención pueden ser empleados para resolver problemas de ingeniería. Este modelo se centra en el reconocimiento de múltiples objetos superpuestos en una imagen permitiendo ligeras distorsiones de forma y tamaño. Permite recuperar patrones completos a partir de muestras imperfectas de los mismos siendo una aplicación típica el reconocimiento de caracteres.

Este modelo está basado en aprendizaje competitivo iterativo auto organizado. Presenta un mecanismo de desplazamiento de la atención mediante el cual la atención va variando de posición a medida que se van reconociendo los caracteres, generando una señal de inhibición sobre las ventanas ya procesadas.

Modelo de Sha'ashua y Ulman [SU88]. También conocido como modelo de saliencia estructural, aunque no está relacionado con ningún modelo de visión temprana, si está relacionado con la búsqueda visual. El objetivo de este modelo es la extracción de estructuras salientes de una imagen mediante el agrupamiento de información local parcial, como por ejemplo segmentos de un contorno fragmentado. Para resolver este problema se emplean técnicas de programación dinámica que permiten obtener resultados en tiempos razonables, devolviendo las curvas que representan las estructuras globales más salientes de la imagen a medida que aumenta el número de iteraciones.

Modelo de Mead y Mahowald [MM88]. Este modelo presenta una aproximación global al funcionamiento de la retina mediante una analogía entre los circuitos eléctricos y las estrategias de procesamiento de señal de la retina. Describe los procesamientos realizados por las primeras capas de células de la retina, resaltando el procesamiento paralelo que en ellas se realiza de modo natural. Aunque este modelo se centra en la capacidad de filtrado espacial de la retina, ha sido optimizado para emplear con secuencias temporales por Beaudot et al. [BPH93] y extendido por Herault et al. [HD07].

Modelo de Grossberg [GMT89]. Este modelo bioinspirado fue desarrollado para la segmentación de imágenes y extracción de texturas. Varios de sus trabajos se dirigen hacia modelos de visión pre-atencional, especialmente los formados en las áreas *V1*, *V2* y *V4*. El modelo diferencia dos módulos separados que actúan de modo cooperativo. Un sistema de extracción de los contornos de los objetos, y un sistema de extracción de características que extrae las regiones uniformes en el interior de los contornos. Ambos módulos permiten generar una representación de color, forma y profundidad invariante ante cambios de luminosidad.

Modelo de Malic y Perona [MP90]. Este modelo, plausible desde el punto de vista biológico, aborda la discriminación de las diferentes categorías de texturas propuestas por B. Julesz [JGS+73, Tre85]. Consta de tres etapas en las cuales, primero se calculan los mapas de características mediante la convolución de la imagen con un banco de filtros orientados. En una segunda

etapa se realiza la supresión de las señales débiles mediante una inhibición no lineal y en la última etapa se combinan todas las respuestas resultantes para detectar los límites de las regiones de textura.

Modelo SLAM de Phaff et al. [PdHH90]. Es uno de los modelos de atención selectiva (SLAM), desarrollado para modelar comportamientos descubiertos en tareas de filtrado. Las características inicialmente empleadas por este modelo fueron el color, la forma y la posición, y en un modelo posterior se añadió la palabra, o sea el texto que describe la característica. Este modelo fue construido mediante un esquema multicapa basado en una estructura activatoria e inhibitoria. A pesar de que el modelo no tiene en cuenta la codificación espacial de las características, resuelve satisfactoriamente problemas previamente comunicados por J.R. Stroop [Str35], en los que la in/congruencia de las características de los estímulos presentados a los sujetos afectaban a la rapidez de las respuestas. Uno de los ejemplos más conocidos de estas tareas consiste en decir el color de una palabra que está escrita en un color diferente.

Modelo de Sandon [San90]. Este modelo piramidal comienza con la construcción de mapas de características partiendo de operadores de procesamiento de imagen tales como detectores de bordes. A diferencia de otros modelos no se basa en características como la intensidad o el color. A continuación estos mapas son procesados por operadores de interés que refuerzan el contraste y posteriormente son integrados mediante redes WTA. Es un modelo multiescala y está basado en una estructura piramidal.

Modelo de Mozer [Moz91, MS96]. Este modelo emplea técnicas de relajación iterativa para realizar la selección de la región de atención. La red de atención empleada, denominada AM, forma parte de un sistema conexionista complejo empleado para reconocer objetos 2D en una escena. El modelo está compuesto de un detector de características, un mecanismo atencional, una red de reconocimiento y una memoria visual de corto plazo que almacena la descripción del objeto. Del mismo modo que el modelo Necognitron, el mecanismo atencional, la red AM, se emplea para seleccionar una región de tamaño y forma variable en la que se procederá a reconocer los objetos. Una de las aplicaciones de este modelo es el reconocimiento de caracteres.

Modelo de Chapman [Cha91]. Modelo desarrollado para servir como componente esencial de un agente autónomo responsable de vincular la percepción y la acción. En este modelo la atención es la responsable de la selec-

ción de los objetivos hacia los que se desplaza el sistema. Es un modelo fundamentado en la teoría de Treisman y una de las primeras implementaciones del modelo original de Koch y Ulman. El modelo trata de reproducir los comportamientos observados en tareas de búsqueda descritos por A. Treisman y S. Gormican [TG88]. Implementa un modelo que tras obtener los mapas de características, los transforma en mapas de activación mediante una umbralización y a partir de ellos realiza una etapa de procesamiento paralelo y otra serie simulando el comportamiento de los mecanismos de atención.

Modelo VISIT de S. Ahmad [AO91, Ahm92]. Este modelo, semejante al modelo de Chapman, es un modelo conexionista de atención que integra tanto mecanismos *bottom-up* como *top-down* para elegir los objetos de interés de una determinada escena. Parte de una descomposición de la imagen en un conjunto de mapas de características sobre los que detecta la actividad mediante operaciones lógicas. Su modelo implementa dos rutas separadas, una que determina el foco de atención e inhibe el resto de la imagen y otra ruta que determina la prioridad de las localizaciones en función de su relevancia. Este modelo simula el proceso de búsqueda serie y paralelo y su implementación se basa en el empleo de redes neuronales artificiales.

Modelo de Olshausen et al. [OAVE93]. Este es un modelo bioinspirado que refleja el comportamiento de las neuronas de la región *V1*. El objetivo de este modelo es la obtención de una representación invariante ante cambios de posición y escala de los objetos presentes en una escena. El mapa de saliencia se emplea como referencia inicial para determinar las regiones en las que pueden estar presentes objetos relevantes. La gran ventaja de esta perspectiva centrada en los objetos reside en el hecho de que se minimiza la cantidad de información necesaria para recuperar un objeto previamente observado desde una perspectiva diferente.

Modelo de Milanesse [Mil93]. Este modelo *bottom-up* emplea como características las componentes de oposición de color, la intensidad y la orientación. Estos mapas iniciales son filtrados mediante operadores centro y alrededores para obtener así los mapas de visibilidad. Dichos mapas muestran las regiones más sobresalientes para esa característica. Por último esos mapas son integrados en un solo mapa de saliencia tras un proceso de relajación no lineal. En una posterior revisión [MWG+94] se añade un mapa de visibilidad relacionado con información *top-down* y un sistema de control del movimiento.

Modelo SERR de Humphreys y Muller [HM93]. Es un modelo de búsqueda por rechazo recursivo (SERR), fue desarrollado para simular los efectos de las agrupaciones observados durante la realización de tareas de búsqueda. Emplea una arquitectura jerárquica que simula el proceso de búsqueda de estímulos construidos con líneas horizontales y verticales que se asemejan a letras. Se observa que cuando los distractores no son homogéneos la búsqueda se vuelve ineficiente, comportamiento que reproduce el modelo. Otro de los resultados que explica es el enorme aumento de la capacidad de identificación cuando están presentes más de un objetivo (estímulo buscado).

Modelo GS de Wolfe [Wol94]. Modelo desarrollado para tareas de búsqueda guiada (GS). Se basa en la creación de un mapa de activación que será el que posteriormente determine el orden secuencial del proceso de búsqueda. Este modelo intenta simular los datos conductuales de los sujetos. Al igual que otros modelos basados en la teoría de Treisman, define dos etapas de procesamiento, una paralela de atención previa (*preattentive*) en la que se identifican características visuales básicas como el color o las orientaciones de las líneas y otra etapa serie de tipo atencional en la que se procesan agrupaciones de características. En el modelo GS también se tiene en cuenta la información tipo *top-down*. El proceso secuencial posterior a la creación del mapa de activación consiste en recorrer por orden las regiones de mayor saliencia hasta que se alcanza el objetivo o se alcanza un determinado umbral. Existen varias implementaciones posteriores, GS1-GS4 en las que se amplían las funcionalidades del modelo descritas en [ZL13].

Modelo SCAN de Postma [PvdHH97]. Este modelo se fundamenta en una red neuronal escalable cuyo objetivo es la identificación de patrones de modo invariante ante desplazamientos. Basándose en las propiedades del SVH, invarianza ante transformaciones y escalabilidad, crea una red basada en elementos básicos, red de puertas. En dicha red bidimensional de conectividad local cada puerta es responsable de supervisar el estado de un elemento, por ejemplo un píxel. Crea conectividades entre elementos próximos y así permite elegir distintas regiones sobre una imagen, implementando en cierto modo el concepto de foco de atención. La combinación de este modelo con una red de clasificación permite identificar patrones conocidos o almacenar nuevos patrones que posteriormente podrán ser identificados.

Modelo de Itti y Koch [IKN98, IK01]. Este modelo es una implementación del modelo de Koch y Ulman [CS85]. La estructura de procesamiento paralelo mostrada en la figura 1.7 ha sido una referencia continuada por mu-

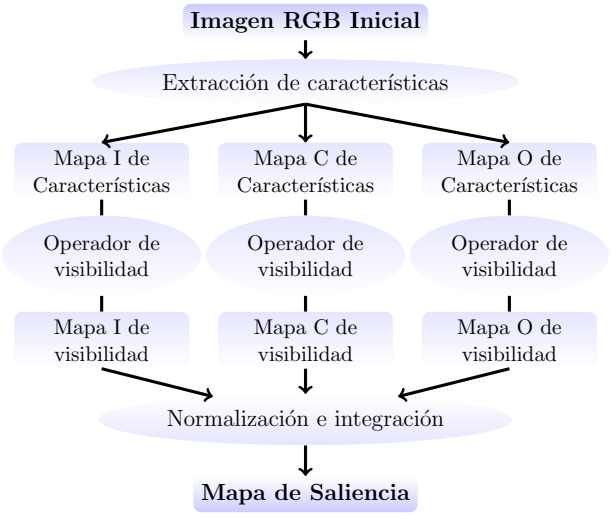


Figura 1.7: Esquema del modelo de Itti y Koch. En este esquema se procesan siete características de modo paralelo; la intensidad (I), dos componentes cromáticas (C) y cuatro orientaciones (O).

chos otros modelos. En este esquema, se construyen paralelamente una serie de mapas fijándose para ello en determinadas características. En el primer modelo propuesto, dichas características eran la intensidad, dos componentes cromáticas y cuatro orientaciones, implementadas mediante filtros de Gabor. En una segunda fase de este procesamiento paralelo se resaltan las características diferenciales, o sea se aplica un operador que resalte aquellas partes que sobresalen respecto a su entorno dando lugar a lo que se conoce como mapas de visibilidad V_c, y por último se integran todos estos mapas mediante la ecuación 1.1, obteniendo así el mapa de saliencia S_M.

$$S_M = \sum_{k=1}^{l} N(V_c) \tag{1.1}$$

Siendo k el número de características evaluadas, y N el operador de normalización. El modelo propone una etapa posterior en la que mediante redes WTA e inhibición de retorno se consigue determinar el orden en el que se realizarán las fijaciones a partir de la información contenida en los mapas de saliencia.

Modelo de Cave [Cav99]. Es un modelo conexionista semejante al GS, se denomina modelo de puertas de características ya que es un modelo piramidal en el que las características van pasando de un nivel a otro a través de

ciertas *puertas* que permiten o no el paso de las características al siguiente nivel. Igual que con el modelo GS, el mecanismo de control de estas puertas está guiado tanto por mecanismos *top-down* como *bottom-up*. Fue desarrollado con el objetivo de modelar y dar explicación a estudios de atención visual como la atención dividida o la selección espacial guiada por características.

Modelo de Deco et al. [DZ01, DPZ02]. Es un modelo conexionista semejante al modelo GS, conocido como modelo de búsqueda dinámica. Modelo fundamentado en la neurociencia, que intenta dar explicación a los fenómenos de visión preatentiva (de procesamiento paralelo) y atentiva (con procesamiento secuencial). Implementa un modelo que emula la integración realizada por el cerebro sobre las dos vías principales de la corteza cerebral, la del *qué* y la del *cómo* tal como se describe en la sección 1.1.1. Asocia los tiempos de reacción, obtenidos durante la realización de tareas de búsqueda, al número de iteraciones necesarias para alcanzar un mapa de atención ganador mediante interacciones competitivas entre varios mapas de características.

Modelo de Zhaoping [Li02]. Este modelo se basa en la existencia un único mapa de saliencia en la corteza visual primaria, independientemente de las características de partida, al contrario que sucede con el modelo de Koch y Ulman [CS85] en el que son varios los mapas de características que se combinan de modo competitivo. Propone una metodología que enlaza la variación de la dificultad durante la realización de tareas de búsqueda, o sea los comportamientos psicofísicos, con la fisiología y la anatomía de la región *V1*.

Modelo de Torralba [JEDT09, Tor03, TOCH06]. En este modelo A. Torralba incluye un procesamiento paralelo inicial que sirve de guía a la atención modulando el mapa de saliencia en función de la información contextual. Para realizar esta identificación del contexto emplea información estadística de las características locales de la imagen. Esta información permite identificar la escena a través de las relaciones espaciales aún sin haber realizado para ello la identificación de los objetos presentes en la escena.

Modelo VOCUS de Frintrop et al. [FBR05]. El modelo fue diseñado para realizar búsquedas guiadas de objetos. El modelo VOCUS añade un componente *top-down*, que pondera las localizaciones previamente visitadas, sobre una arquitectura *bottom-up*, basada en el modelo original de Itti y Koch [IKN98]. Los pesos asociados dependen tanto de los elementos buscados, como de las propiedades del fondo, actuando mediante una estructura

excitatoria-inhibitoria. Mediante esta metodología en promedio se consigue localizar los objetos a lo largo de las tres primeras localizaciones visitadas.

Modelo GAFFE de Rajashekhar [RvdLBC08]. El modelo GAFFE (*Gaze-Attentive Fixation Finding Engine*) identifica de modo automático regiones visualmente interesantes basándose para ello en las estadísticas de las posiciones de fijación. Se observa que las características locales de las imágenes, como la luminancia y el contraste toman valores superiores en las posiciones donde se producen las fijaciones [RZ99a, DE03]. El modelo implementa además el concepto de fóvea mediante el suavizado progresivo de la imagen a medida que aumenta la distancia a la posición de fijación previa. De este modo los parches extraídos de los que se obtienen las estadísticas estarán difuminados en función de la distancia a la que se encontrasen en el instante de la fijación previa.

Modelo SUN de Zhang et al. [ZTM+08]. Este modelo defiende que las estadísticas obtenidas de las características visuales de las escenas naturales juegan un importante papel en la creación del mapa de saliencia. Esta información representa la experiencia visual previa del sujeto, por ello es un modelo en el que hay una cierta componente *top-down*. Dichas estadísticas serán las que se comparen con la nueva información adquirida.

Modelo VARIANCE de Itti y Baldi [IB05]. Este modelo fue empleado como una referencia sencilla que cuantificase la capacidad explicativa de las características locales sobre imágenes estáticas [IB09]. Dicho modelo calcula la varianza local de la luminosidad sobre parches de imagen de 16x16 pixels y tiene su origen en la descripción de Reinagel y Zador [RZ99b].

Modelo de Bruce y Tsosos [BT09]. Este modelo bioinspirado se basa en la estructura de la región *V1*. Emplea medidas de información estadística para predecir las posiciones de las fijaciones de los sujetos. Para ello aplica un proceso de maximización de la información donde la saliencia se modela como una medida de autoinformación [BT06]. Mediante este modelo se consiguen reproducir múltiples fenómenos psicofísicos y se obtienen unos buenos resultados de predicción de fijaciones oculares.

Modelo AWS [GDLFVP12]. Modelo basado en que el factor clave que permite determinar la saliencia visual es la adaptación contextual de corto alcance que surge del proceso de blanqueado adaptativo. Para obtener la saliencia realiza un proceso de blanqueado de las imágenes de partida,

una descomposición multiescala y multiorientación de las componentes de color a través de los filtros log Gabor. Luego mediante la decorrelación de las respuestas multiescala a través de la extracción de una medida local de la variabilidad y con un posterior promediado se obtiene una medida única y eficiente de la saliencia. Para obtener la decorrelación se aplica el análisis de componente principales (PCA) sobre un conjunto multiescala de características de bajo nivel. Este modelo también reproduce efectos psicofísicos como la asimetría en el caso de presencia ausencia de características o la ley de Weber aplicada a la saliencia de un segmento como función de la diferencia con los de su entorno.

Múltiples modelos de los aquí descritos integran la información asociada a diferentes características en un único mapa de atención final del modelo. En general esta integración consiste en dos etapas; una primera etapa en la que los diferentes mapas asociados a características muy diferentes (luminosidad, color, orientación, etc.) se reescalan a un mismo rango para que puedan ser comparables y una segunda etapa que simula un proceso competitivo del que surge el mapa de características final. Esta misma estructura de reescalado/normalización se puede aplicar a otros niveles, por ejemplo para la combinación de los mapas de visibilidad, o los mapas estáticos y dinámicos para el caso de secuencias de vídeo.

1.2.2. Modelos dinámicos

La identificación de la saliencia en dominios espacio temporales, ha sido de gran ayuda en aplicaciones como la clasificación de vídeos [MGP07] o la identificación de gestos o tareas [SM11]. Algunos modelos de saliencia son una aplicación directa de modelos estáticos a escenas en movimiento, mientras que otros han sido específicamente desarrollados para tratar con el movimiento o se han basado en extensiones de detectores de características 2D especialmente adaptados al dominio espacio temporal. Mediante estos nuevos elementos y otros creados específicamente para cada caso se han construido nuevos modelos dinámicos, sintetizados en la figura 1.8, que se describirán brevemente en esta sección. La predicción de la información de movimiento,

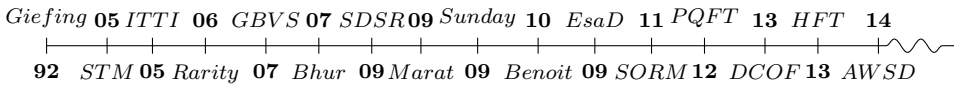


Figura 1.8: Diagrama temporal de los principales modelos dinámicos.

por ejemplo mediante los filtros de Kalman o las transformadas de Fourier,

combinada con la información de atención ofrece una vía para la consecución de resultados que cada vez se asemejen más a los obtenidos por los humanos. Independientemente del rápido crecimiento de la capacidad de cómputo de los ordenadores actuales, la dificultad inherente al procesamiento de imágenes de vídeo requiere el empleo de mecanismos atencionales para reducir el volumen de datos adquiridos a un subconjunto más manejable.

En cuanto a la clasificación de los modelos dinámicos, de modo análogo a los estáticos existe una clasificación global que diferencia aquellos que priorizan las características de las escenas, modelos *bottom-up*, frente a aquellos que tienen en cuenta elementos cognitivos, conocimientos previos, expectativas, el objetivo de la tarea o simplemente la recompensa. Se ha observado que en tareas en las que el objetivo es claro, p. ej. en tareas de búsqueda, la información *top-down* prevalece sobre los elementos *bottom-up*. En múltiples casos, a los modelos *bottom-up* se le añaden elementos de carácter *top-down* que permiten mejorar la capacidad predictiva del modelo, por ejemplo el modelo SUNDAY [ZTW09] o el modelo de Wolfe [Wol94].

Respecto al tipo de información empleada, hay modelos que emplean información de regiones, y por ello son más adecuados para la segmentación de objetos, como el modelo de Achanta et al. [AHES09], mientras que otros hacen uso de información local como el de Seo y Milanfar [SM09b]. Otros combinan ambos tipos de información, local y global mediante el uso de grafos, como el GBVS de Harel y col [HKP07]. En algunos modelos, la información es analizada en el dominio espacial, mientras que en otros se emplea el dominio de la frecuencia [GZ10, LLA+13]. Este es el caso del modelo multirresolución PQFT de Guo y Zhang, desarrollado para la compresión de imágenes y vídeos [GZ10]. Modelos como el PQFT o el HFT de Li y Levine [LLA+13] también se basan en el empleo de la transformada de Fourier.

Un factor común a todos ellos es que al analizar los resultados del procesamiento de modelos estáticos previos, usando escenas en movimiento, se ha observado que los cambios temporales guían la saliencia tanto como la propia saliencia espacial [TPL+02][MSHH11]. De ahí la gran importancia de incluir el movimiento, tal como muestran por ejemplo el modelo GBVSmotion de Harel y col [HKP07] o el de A. Bhur [BWMH07b].

A continuación se incluye una breve descripción de los modelos de la figura 1.8, que son una muestra representativa de los modelos dinámicos de saliencia.

Modelo de Giefing et al. [GJM92]. Refleja la estructura fisiológica de la fóvea y el procesamiento paralelo de la región central frente a la región periférica. Fue diseñado para controlar un sistema de visión activa e imple-

menta el propio concepto de fijaciones y sacadas en el mecanismo de control. De tal modo que, tras las sacadas, revisa la escena buscando puntos objetivo que recorre secuencialmente. Realiza el análisis de movimiento durante las fijaciones e implementa una cierta memoria, que evita revisar localizaciones recientemente vistas (inhibición de retorno). Del mismo modo implementa la capacidad de olvidar mediante mecanismos de difusión y relajación, permitiendo de ese modo revisar de nuevo localizaciones anteriormente vistas cuando trascurre cierto tiempo. El modelo parte de la transformada log-polar de la imagen inicial simulando con ello el comportamiento de muestreado no uniforme de la retina. A continuación se obtienen los mapas de características empleando la correlación cruzada con varios patrones de referencia tales como líneas o cruces, semejantes a los observados en la región *V1* de la corteza. La principal aportación de este modelo frente al original de Sandom [San90] consiste en el empleo de las derivadas temporales para la identificación de objetos en movimiento.

Modelo STM de Tsotsos et al. [TCW+95, TPL+02, TLMT+05]. Se centra en la identificación de patrones de movimiento, como los observados en las áreas de la corteza *V1*, *MT*, *MST* y *7a*, mediante un modelo neuronal jerárquico realimentado. Se fundamenta en la selección espacial mediante la eliminación de localizaciones irrelevantes y en la selección de características mediante la inhibición de aquellas que también son irrelevantes. Las restricciones de la tarea actuarán sobre los pesos de estas inhibiciones en función del tipo de tarea, tomando los valores por defecto cuando no se defina ninguna prioridad. Refleja las conectividades jerárquicas de las áreas visuales consideradas e intenta que cada una de las capas del modelo extraiga las características correspondientes a la región que representa (velocidad local en *V1*, movimientos de expansión en *MST*, etc.).

Modelos CIO, CIOFM y SURPRISE de Itti et al. Las diferentes implementaciones del modelo de Koch y Ulman [CS85], tienen en común el hecho de que la saliencia se obtiene partiendo de los mapas individuales de color (C), luminancia (I) y orientación (O), a diferentes escalas espaciales. A continuación mediante operadores de centro y alrededores se analizan las diferencias entre las respuestas de los filtros para varias escalas y se obtienen los mapas de características. Tras normalizarlos y combinarlos se obtiene el mapa de saliencia global. Esta formulación, consistente con la teoría de integración de características [Tre85], da lugar a un modelo que predice con gran acierto las fijaciones humanas [IKN98][Itt06].

La versión dinámica del modelo, CIOFM, añade un canal de parpadeo

(F, *flicker*, diferencia de fotogramas) y un canal de movimiento (M) a las características básicas procesadas, confirmando que el movimiento es un buen predictor de la saliencia y que la combinación del movimiento con las demás características mejora aún más los resultados del modelo [Itt05].

El modelo SURPRISE relaciona la saliencia con la desviación frente al modelo esperado, dicho modelo se construye a partir de un conjunto de modelos internos del mundo visual. Es un modelo Bayesiano fundamentado en la teoría de la información [IB09], que posteriormente fue también adaptado a escenas dinámicas [IB05].

Modelo Rarity-G/L de Mancas et al. [MMTGM06, RMGD12]. Los modelos de rareza GR (global) y GL (Local) se apoyan en el hecho de que las regiones importantes son en múltiples casos inusuales, o sea emplean el concepto de *rareza* para obtener la información de saliencia. El modelo GR, obtiene la saliencia como una medida de auto información de la media y la varianza local de la intensidad de la imagen, bajo la hipótesis de que las características minoritarias de la imagen son las que atraen la atención. El modelo LR emplea el contraste local como medida de saliencia, basándose en la hipótesis de que las regiones con un alto contraste atraen la atención. Esta última medida depende de la escala a la que se observe la escena por lo que para compensar este efecto el modelo realiza un análisis multiescala sobre imágenes filtradas mediante filtros Gaussianos.

Modelo GBVS de Harel et al. [HKP07]. El modelo de saliencia visual basada en grafos calcula el mapa de saliencia a partir de la función de distancia multiescala de los mapas de disimilaridad de características. Crea los mapas de características a diferentes escalas espaciales y del mismo modo que el modelo de Itti y col [IKN98] emplea el color, la intensidad y la orientación. Con esos mapas el modelo construye un grafo conectado sobre todas las localizaciones de los mapas de características y le asigna pesos a las uniones entre los nodos inversamente proporcionales al valor de la semejanza entre los valores de las características y su distancia espacial. A continuación se realiza un proceso de normalización de los pesos de los elementos del contorno y se define una relación de equivalencia entre los pesos y las probabilidades de transición. La distribución final de equilibrio representa el mapa de activación y tras combinar los mapas asociados a las diferentes características se obtiene el mapa de saliencia del modelo.

Modelo de Hou y Zhang [HZ07]. El modelo de detección de saliencia mediante la aproximación de espectro residual es un modelo que emplea la

información del espectro logarítmico de una imagen para la construcción del mapa de saliencia en el dominio espacial. Es independiente de conocimientos a priori como las características o las categorías de las imágenes. Se fundamenta en la hipótesis de codificación eficiente de Barlow [Bar61], o sea la eliminación de la información redundante de las entradas sensoriales. Mediante este modelo se explora la idea de que las singularidades estadísticas del espectro son las responsables de la existencia de regiones relevantes en las imágenes, y partiendo de esa idea se puede extraer el mapa de saliencia.

Modelo de Bhur [BWMH07b]. Se basa en los descubrimientos acerca del procesamiento del movimiento en el cerebro. Para la estimación del movimiento se identifican las correspondencias entre regiones de diferentes fotogramas (técnicas BMA), obteniendo así un campo de velocidades que se empleará como componente básico de la parte dinámica del modelo. En la comparación del movimiento contempla el contraste de magnitud y fase, o sea el contraste entre la magnitud de la velocidad local y las velocidades de las regiones que lo rodean, y el contraste entre las direcciones de las velocidades. Compara diferentes técnicas BMA y a la vez proporciona una comparativa de los resultados obtenidos con diferentes estrategias de integración de los modelos estático y dinámico.

Modelo SDSR de Seo y Milanfar [SM09a, SM09b]. En el modelo SDSR (detección mediante autosemejanza) el mapa de saliencia del modelo se obtiene midiendo la similaridad entre la matriz de características de un píxel respecto a los píxeles vecinos, bajo la hipótesis de que el contraste local atrae la atención. El modelo crea una estructura local asociada a cada píxel que contiene los descriptores locales y a continuación mediante la matriz de similaridad mide la probabilidad a priori de cada píxel en su entorno. Este modelo, aplicable a imágenes de cualquier tamaño, es un modelo no paramétrico.

Modelo de Marat [Sop09, MHPG+09]. Combina el procesamiento estático obtenido a través del análisis de las orientaciones y frecuencias espaciales, con el procesamiento dinámico, obtenido del análisis del módulo de la velocidad a través del cálculo del flujo óptico. Para esta parte dinámica se realiza una compensación previa del movimiento de la cámara, en fotogramas consecutivos, para poder diferenciar los movimientos de los objetos respecto al fondo.

Modelo SUNDAY de Lingyun Zhang [ZTW09]. El modelo estadístico es una extensión a secuencias dinámicas del algoritmo denominado SUN [ZTM+08, KTZC09]. Calcula la saliencia *bottom-up* como una medida de autoinformación de las características locales de la imagen. Se fundamenta en el hecho de que el sistema visual estima la probabilidad a priori de que un determinado objeto se encuentre en cada una de las localizaciones de la escena, a partir de las características observadas. Además el modelo presenta una componente *top-down*, a través del empleo de estadísticas obtenidas de imágenes naturales. Información que emplea para comparar con las estadísticas obtenidas de la propia imagen.

Modelo de Benoit et al. [BCDH10]. Es uno de los modelos bioinspirados en los que se integra el movimiento. En este modelo se describen los primeros componentes de entrada del SVH; la retina y *V1*, la primera de las áreas de la corteza visual. Mediante módulos simples pero altamente eficientes permite realzar los contornos de los objetos, extraer contornos en movimiento, analizar las orientaciones presentes en la escena y detectar eventos relacionados con el movimiento en función del contexto.

Modelo ESA-D de Esa Rathu et al. [RH09]. Emplean esta técnica de obtención del mapa de saliencia inicalmente como método para la eliminación del fondo y posteriormente para la segmentación de objetos mediante el modelo CRF (*conditional random field*)[RKSH10]. El mapa de saliencia utiliza como punto de partida el contraste local de características como la iluminación el color o el movimiento de este modo mediante la comparación de histogramas locales identifica regiones con un alto contraste entre una región y su entorno. Afirma su especial adecuación para escenas con un alto contraste de movimiento o escenas con cámara en movimiento.

Modelo SORM de Kim et al. [KJK11]. Fue desarrollado para aplicaciones de tiempo real, por lo que se prioriza la rapidez de los cálculos. Emplea una aproximación espaciotemporal en la que la saliencia se calcula a partir de las signaturas de las orientaciones de las aristas y de color. Para ello, se calculan las diferencias entre una región y las circundantes para la componente espacial y las diferencias de los gradientes temporales para la componente temporal. Muestra muy buenos resultados en aplicaciones de reencuadre y extracción de movimiento.

Modelo PQFT de de Guo y Zhang [GZ10]. Este modelo multirresolución fue desarrollado para la compresión de imágenes y vídeos. Se fundamenta

en el hecho de que las regiones salientes de una escena están más relacionadas con la fase del espectro de la transformada de Fourier que con la amplitud de dicho espectro. Las características que integra el modelo son la intensidad, el color y el movimiento. Una de las características relevantes de este modelo además de su eficiencia en el proceso de compresión de vídeo es la capacidad de procesamiento a tiempo real. Una aplicación para reencuadre de vídeos basada en la misma información es descrita por Fang et al. [FLL+12].

Modelo DCOF de Zhong et al.[ZLR+13]. Modelo específicamente diseñado para el análisis de secuencias temporales. Emplean la información estática basándose en la implementación del modelo GBVS que fusiona con la información de movimiento obtenida mediante una variante del cálculo del flujo óptico. Esta modificación del flujo óptico además de tener en cuenta dos fotogramas adyacentes, analiza la consistencia de las regiones en movimiento a lo largo del tiempo.

Modelo HFT de Li et al. [LLA+13]. Se fundamenta en el hecho de que la información global de una imagen permite extraer las regiones *no-salientes* y por exclusión determinar las regiones salientes. Para ello, el modelo emplea transformadas de Fourier hipercomplejas (HFT) desde un punto de vista global integrando la intensidad, las diferencias de color y el movimiento. Estas transformadas permiten trabajar sobre mapas de que contienen múltiples características. Bajo ciertos condicionantes, este modelo puede ser visto como una extensión del modelo PQFT.

Modelo de Pengfei [WFC+13]. Se puede ver como una variante del modelo de ITTI al que se le añade un canal de información asociado al movimiento en el espacio tridimensional, obtenido gracias a las nuevas imágenes de rango que ofrecen dispositivos como el Kinect™ de Microsoft, entre otros. En este modelo se construye un espacio de velocidades y se da mayor peso a aquellos objetos que se mueven en la dirección del observador, atendiendo a razones evolutivas, dichos objetos se comportan como depredadores que se aproximan hacia el sujeto atrayendo su atención.

Gran parte de los modelos aquí descritos han sido evaluados por su capacidad de reproducir resultados de experimentos psicofísicos, predecir las posiciones de las fijaciones oculares de los humanos y han sido comparados al menos con la implementación del modelo de Kock y Ulman realizada por Itti y Koch [IKN98, IK01].

1.2.3. Estrategias de integración del movimiento

En esta sección describiremos brevemente las diferentes estrategias de integración empleadas por los modelos descritos en la sección 1.2.2 para la obtención del mapa de saliencia combinado único, S_{MC}, de cada fotograma de un vídeo. En general, la estrategia más habitual consiste en la combinación de los mapas de características, realizando una etapa de normalización y después una de competición, pero también se pueden emplear estrategias que dan prioridad al movimiento. Los diferentes modelos revisados han planteado estrategias como las que se describen a continuación.

Estrategia competitiva. Mediante esta estrategia el mapa de saliencia S^A_{MC} se obtiene empleando la ecuación 1.2 a partir de los mapas estáticos S^S_M y dinámicos S^D_M.

$$S^A_{MC}(t) = N(S^S_M) + N(S^D_M) \tag{1.2}$$

siendo N el operador de normalización que permite obtener a partir de los mapas de visibilidad aquellos con menor número de máximos.

Una variante de esta estrategia emplea directamente cada uno de los mapas estáticos asociados a las diferentes características S^{Si}_M (intensidad, color, orientación, etc.) y el mapa asociado al movimiento S^D_M.

$$S^A_{MC}(t) = N(S^{S1}_M) + ... + N(S^{Si}_M) + N(S^D_M) \tag{1.3}$$

De ese modo cualquiera de las características empleadas y el movimiento están al mismo nivel.

Estrategia condicionada por el movimiento. Mediante esta estrategia la componente de movimiento modula la actuación de la componente estática.

$$S^B_{MC}(t) = \begin{cases} S^S_M(t) & si\ S^D_M(t) < \varepsilon \\ 0 & en\ otro\ caso \end{cases} \tag{1.4}$$

siendo ε el umbral que determina el valor para el cual se considera la contribución estática. De este modo los objetos salientes del modelo estático tan sólo compiten cuando hay movimiento. Esta estrategia ha demostrado ser muy útil en aplicaciones de videovigilancia.

Estrategia ponderada Es una combinación de las estrategias anteriores en la que se pondera la contribución de ambas con un determinado factor que se determinará en función de la aplicación para que tenga mayor peso una u otra estrategia.

$$S_{MC}(t) = \alpha \cdot S^A_{MC}(t) + (1 - \alpha) \cdot S^B_{MC}(t) \tag{1.5}$$

Figura 1.9: Esquema del modelo de Nabil Ouerhani. Este esquema realiza por una parte la extracción estática de características y por otra la extracción dinámica.

Estrategia de movimiento prioritario. Es una variante de la estrategia anterior en la que el movimiento condiciona la actuación de la componente estática o la dinámica. Se emplea un umbral que determinará la actuación de una u otra componente.

$$S_{MC}(t) = \begin{cases} S_M^S(t) & si\ S_M^D(t) < \varepsilon_1 \\ S_M^D(t) & si\ S_M^D(t) > \varepsilon_2 \\ S_{MC}^A(t) & en\ otro\ caso \end{cases} \tag{1.6}$$

En caso contrario se obtiene el mapa final del mismo modo que en la estrategia competitiva ya sea mediante la ecuación 1.2 o 1.3.

Para la construcción del mapa de saliencia, en los modelos dinámicos se emplean varios fotogramas consecutivos. En cambio, los modelos estáticos tan sólo emplean la información de cada uno de los fotogramas. En la figura 1.9 se muestra un diagrama con la metodología combinada empleada por el modelo de Nabil Ouerhani [Oue04], en ella se realiza un procesamiento paralelo de las características estáticas, semejante al mostrado en el esquema de la figura 1.7, y por otra parte se realiza un análisis del movimiento empleando el flujo óptico, a partir de ambos mapas se obtiene el vídeo con la información de saliencia para cada fotograma con la combinación de los mapas estáticos y dinámicos.

Existen varias revisiones detalladas con más información acerca de los modelos estáticos [HHK12], dinámicos y estrategias de integración [BWMH07a]

aquí descritos. Cabe destacar la revisión de Dietmar Heinke y Glyn W. Humphreys del año 2005 [HH05a], la revisión de A. Toet del 2011 [Toe11] y la revisión de Ali Borji y Laurent Itti [BI13] en la que se agrupan las diferentes metodologías en función de trece parámetros, que permiten determinar los ámbitos de aplicación de cada uno de los modelos analizados.

1.3. Aplicaciones de los modelos de atención visual a la visión por computador

A continuación se describirán algunas de las aplicaciones más relevantes que han surgido a partir de la publicación de los trabajos de Itti et al., y Koch y Ulman en los años ochenta [IKN98, IK00, IK01, CS85], así como otras nuevas metodologías aplicadas a diversos ámbitos

1.3.1. Detección y reconocimiento de objetos

Los modelos de atención visual han ido evolucionando y cada vez funcionan mejor en tareas de aprendizaje y reconocimiento de objetos debido a su flexibilidad y generalidad [FBR05]. Además con la incorporación de nuevas tecnologías tales como los mapas de profundidad, y el abaratamiento de las cámaras de rango (dispositivos que devuelven información de profundidad) es posible hacer más accesibles metodologías como la propuesta en la que se aborda el reconocimiento de objetos mediante el sistema VOCUS (detección visual de objetos con un sistema de atención computacional), que incorpora información de profundidad mediante el uso de un escáner láser [FBR05]. Un ejemplo de implementación de un algoritmo de entrenamiento genérico de detección de objetos es el de Seo y Milanfar, en el que se emplearon las características salientes como descriptores. Esto permite aprender nuevos objetos a partir de tan solo un ejemplo ofreciendo una eficiencia de reconocimiento muy alta y sin necesidad de realizar procesos de segmentación previa [SM10, SM11].

Walther et al. muestran cómo la saliencia puede acelerar considerablemente el proceso de aprendizaje y el rendimiento del proceso de reconocimiento, permitiendo además el aprendizaje simultáneo de múltiples objetos sobre una misma imagen [WRKP05].

Barrington et al. propusieron NIMBLE, un modelo de memoria visual basado en las fijaciones extraídas a partir de la información de saliencia. Dicho modelo también es capaz de reconocer caras, a partir de tan solo una fijación [BMHC08]. Kanan y Cotrel han propuesto posteriormente una

mejora del NIMBLE para la extracción de características, así como para el cálculo de la saliencia, empleando estadísticas de imágenes naturales [KC10].

Han y Vasconcelos emplean información de saliencia y relevancia *top-down* para la obtención de una aproximación bioinspirada plausible de reconocimiento de objetos [HV10]. Para ello reemplazan por una red de cálculo de saliencia la primera capa de la arquitectura HMAX [RP99]. Esta red emula el funcionamiento del sistema visual y los resultados muestran cómo la información de saliencia puede mejorar enormemente el proceso de reconocimiento de objetos.

Gao et al. muestran el uso de la saliencia en un modelo *top-down* para la localización de objetos y la clasificación de imágenes [GHV09]. Este modelo actúa como un selector de puntos de interés en función de la relevancia que estos tienen para el sistema de reconocimiento visual. El resultado demuestra un buen rendimiento durante la localización de objetos en escenas muy desordenadas y una gran capacidad para extraer información relevante para la clasificación de imágenes.

Alexe et al. muestran cómo un detector genérico de objetos basado en la medida de saliencia puede aprender nuevas categorías de objetos a partir de imágenes de videovigilancia [ADF10]. La misma metodología permite además mejorar el rendimiento durante la detección y localización de objetos de categorías conocidas.

En este ámbito también hay algún trabajo teórico como el de Harel y Koch en el que se valora cuan óptimo y en que medida es adecuado el uso de la atención visual para el aprendizaje y reconocimiento de objetos [HK09]. Independientemente de estas valoraciones teóricas, los resultados mostrados en las publicaciones aquí referenciadas que tratan sobre detección y reconocimiento de objetos, han demostrado ser novedosos y muy competitivos con respecto a los modelos existentes en el estado del arte. Con todo esto se evidencia la utilidad de la aplicación de los modelos de atención visual en tareas de detección y reconocimiento de objetos.

1.3.2. Segmentación

Al igual que en otros apartados de este trabajo, al revisar la bibliografía referida a la segmentación, nos encontramos con la problemática de la valoración de los resultados obtenidos. El caso de la segmentación es un ejemplo claro en el que la segmentación manual suele ser la referencia frente a la cual se mide la calidad del resultado obtenido por los diferentes modelos. Así es el caso de los trabajos de Hou y Zhang [HZ07] y Seo y Milanfar [SM10] en los que se muestra la aplicabilidad de sus modelos para segmentación genérica de objetos basada en información de saliencia. Igualmente, Achanta et al.

[AHES09] proponen un modelo de detección de regiones basado en la información de frecuencia para la segmentación de objetos. Este modelo obtiene buenos resultados sobre una base de datos de mil imágenes naturales cuando se comparan con segmentaciones realizadas manualmente por humanos. Además mejora varios de los métodos de saliencia basados únicamente en aproximaciones puramente computacionales o bio-inspiradas. Para conseguir una pre-segmentación automática de imágenes de alta resolución en teledetección, Sun et al. proponen una combinación de varios métodos *bottom-up* de cálculo de saliencia, un detector de aristas y un detector basado en grafos (GBVS) [SWZW10, HKP07]. De este modo se consigue reducir la carga computacional y mejorar la eficiencia del procesamiento con imágenes de gran resolución.

1.3.3. Visión en robótica

En tareas de visión activa, las cabezas robóticas con cámaras motorizadas precisan generar los desplazamientos que controlan la dirección de la mirada de la cabeza robotizada. La decisión de cual es el siguiente punto a seguir, observando una escena, es compleja y la atención visual ha demostrado ser una de las posibles vías para resolver este problema.

Un ejemplo de uso de la atención visual en este ámbito ha sido propuesta por Clark y Ferrier [CF88]. En el modelo propuesto se incluyen dos mecanismos independientes para la realización de movimientos de persecución y sacadas. Precisamente para las sacadas se emplea la extracción de mapas de características obtenidos mediante la correlación cruzada de las imágenes capturadas con imágenes muestra conocidas. De este modo, ponderando los mapas con coeficientes variables en el tiempo, se pueden ir recorriendo las regiones más salientes de la imagen.

Un sistema semejante que también implementa los mecanismos de persecución y sacadas ha sido desarrollado por el grupo de Eklundh [BEU96]. En este sistema el mecanismo de control de la mirada se basa en un modelo de atención visual que emplea información de profundidad y movimiento.

Otro trabajo en el que se integra la atención visual en un sistema de visión activa ha sido llevado a cabo por el laboratorio IMA, de la universidad de Hamburgo. El sistema completo conocido como NAVIS (Neural Active Vision) está constituido básicamente por dos componentes; un sistema de atención y un sistema de control de las cámaras. El módulo de atención emplea características como el color, simetrías de aristas o excentricidades de regiones. Todas estas características combinadas con información *top-down* acerca del entorno dan lugar a un mapa de atención que es empleado por el módulo de control de las cámaras para realizar desplazamientos hacia esas

posiciones[BMB01, BJM98].

Ballard y Brown, del grupo de Rochester, han trabajado en proyectos de visión activa empleando estrategias *top-down*, basadas en el conocimiento, para la detección de puntos de interés en una imagen y también han desarrollado metodologías combinadas *top-down* y *bottom-up* usando modelos de Markov [RB91]. En este trabajo las pistas que guían la atención provienen del mapa de saliencia obtenido al combinar información de intensidad, aristas, varianza y otras características simples.

Una de las dificultades de la navegación en entornos exteriores mediante robots autónomos es la necesidad de localizar marcas de modo fiable y rápido. Para resolver esta tarea Celaya et al. presentan una estrategia que consiste en seguir las regiones de alta saliencia a lo largo del tiempo [CAJT07]. Estas mismas regiones pueden ser almacenadas en una base de datos y ser utilizadas más tarde para identificar la localización del robot durante la navegación.

En el ámbito de la exploración planetaria, Privitera y Stark proponen el empleo de modelos de atención para la identificación de regiones relevantes desde el punto de vista geológico. Para ello emplean imágenes de satélite que son analizadas automáticamente, identificando localizaciones sobre la corteza de un planeta en las que podría ser interesante realizar una exploración mediante robots [PS03].

Otra de las aplicaciones de los modelos computacionales de atención visual a la visión en robótica es la de Heinen y Engel [HE09]. En este trabajo se propone el modelo de atención NLOOK, altamente independiente ante transformaciones bidimensionales de similaridad, como las traslaciones, rotaciones, reflexiones o variaciones de escala.

1.3.4. Compresión de imágenes o vídeos

Un ámbito de aplicación de la saliencia visual de especial interés es la compresión de imágenes. En función de las restricciones tecnológicas se podrán aplicar diferentes grados de compresión al contenido de las imágenes, ya sea para la creación de iconos de pequeño tamaño [LMLCBT06, HZ08], tan extendidos actualmente con la enorme expansión del uso de dispositivos móviles, o para el control de la compresión de imágenes durante su transferencia a través de la red [ZP90].

Se pueden encontrar múltiples ejemplos de compresión basada en la información de saliencia tanto aplicada sobre imagen [OBH+01] como aplicada sobre vídeo [Itt04]. También en este ámbito la combinación de metodologías ofrece buenos resultados. En el trabajo de Wang et al. se muestra cómo la combinación del cálculo del gradiente con la información de saliencia del modelo de Itti et al. puede dar lugar a un mapa de relevancia que permite

realizar reescalados de imágenes con buenos resultados [WTSL08].

Hua et al. muestran un ejemplo de conversión de formatos de vídeo, ya sea para visualización en pantallas de diferentes tamaños o para almacenamiento. En este trabajo se propone el seguimiento de regiones de alta saliencia como estrategia para determinar la cantidad de compresión aplicada localmente obteniendo de este modo una conversión libre de distorsiones [HZL+10].

En esta misma línea Hwang et al. proponen una metodología que evita distorsiones excesivas durante el proceso de muestreado de imágenes. Dicha metodología se puede enmarcar entre las más conocidas como técnicas de reescalado que tienen en cuenta el contenido. En este trabajo se combina la información del mapa de saliencia con detectores de caras para ayudar al proceso de remuestreo [HC08]. Un trabajo de Liu et al. añade además la información de movimiento a la información de saliencia [LSZ+07]. En la misma línea Rubinstein et al. [RSA08] proponen una metodología que realiza la segmentación de dominios volumétricos componiendo los diferentes fotogramas en una matriz tridimensional y calculando la intersección de esos volúmenes con cada fotograma para eliminar segmentos de tal modo que no se produzca un aumento de energía durante el proceso de eliminación. De este modo se consigue un reescalado que no produce distorsiones en la imagen.

1.3.5. Otras aplicaciones

Cuando a alguna de las aplicaciones anteriormente descritas se le añade la variable tiempo surgen nuevas aplicaciones tales como la detección y reconocimiento de gestos o cambios. Trabajos como el de Tian y Yue van en esa dirección permitiendo identificar cambios a lo largo del tiempo en imágenes de teledetección [TWY07]. Por otra parte Seo y Milanfar muestran cómo realizar el reconocimiento de acciones en vídeos sin la necesidad de segmentar los objetos del fondo, sin realizar estimaciones del movimiento ni seguimiento de puntos. Para ello se emplean descriptores espacio temporales aplicados sobre un vídeo consulta, que miden la probabilidad a priori de un vóxel en su entorno, y se comparan con los extraídos del vídeo objetivo [SM11].

Si en lugar del tiempo añadimos información espacial estaremos ante escenas tridimensionales. También ahí tiene cabida la atención visual, como muestran Xue et al. [XXL+13] empleando la información de saliencia para generar un efecto de profundidad en las escenas. Pero también puede ser aplicada para mejorar el aspecto final de la visualización de objetos tridimensionales como muestra el reciente trabajo de Dong et al. [DLF+14].

Otro ejemplo en el que se combinan varias metodologías existentes para obtener un modelo más elaborado es el de Michalke et al. en el que se muestra un sistema de asistencia a la conducción combinando información

de saliencia, seguimiento y reconocimiento de objetos [MGS+07]. Este sistema ofrece mensajes de alerta para situaciones potencialmente peligrosas durante la conducción.

También se emplea la información de saliencia para la creación de marcas de agua como método de protección de datos o firma digital de imágenes. En este caso los mapas de saliencia permiten insertar el contenido en zonas de la imagen poco salientes, de tal modo que estas marcas sean imperceptibles a primera vista [SSP+09]. Estas mismas regiones de baja saliencia también pueden ser empleadas para presentar información publicitaria con bajo impacto visual [LQH+09]. En otro ámbito esta misma información de saliencia puede ser analizada para identificar manipulaciones de imágenes para evaluaciones forenses [MDNBDN12].

Mediante una combinación en paralelo de una estrategia *bottom-up* bioinspirada y una elección entre mapas estáticos o dinámicos, Marat et al. proponen una metodología para la generación de resúmenes de vídeo. Este tipo de metodologías son válidas tanto para la creación de resúmenes mediante fotogramas relevantes del vídeo como para crear previsualizaciones rápidas del contenido de los mismos [MGP07]. En la misma línea Yu-Fei Ma et al. describen una metodología que permite extraer resúmenes de vídeo sin la necesidad de una interpretación semántica del contenido, en su propuesta incluyen información de vídeo, audio y lingüística [MLZL02].

El impacto de las aplicaciones de inspección visual de defectos en el ámbito industrial para la producción de objetos manufacturados es enorme. Existe una gran demanda de este tipo de sistemas ya que evitan el avance de objetos defectuosos dentro del sistema de producción. En este contexto un defecto es precisamente algo que atrae la atención rompiendo de algún modo las reglas de regularidad. Un ejemplo claro es la detección de cortes en las pistas de los circuitos impresos [BHBH92]. En esta misma línea existen otras posibles aplicaciones como la clasificación de objetos en sistemas en cadena o la separación en sistemas de reciclaje. Precisamente en estos entornos la atención visual muestra un buen comportamiento ya que las reglas de selección son ambiguas y difíciles de precisar mediante una estrategia puramente *top-down*.

Una novedosa línea de investigación enfocada hacia pacientes con retinitis pigmentosa o degeneración macular relacionada con la edad es la propuesta por Parikh. En este trabajo se emplean los algoritmos de detección de saliencia en una prótesis retiniana que se implanta a los sujetos que debido a su patología han perdido la visión. Los algoritmos empleados son una simplificación de los desarrollados por Itti et al. para obtener una mayor eficiencia ya que en estas aplicaciones es enormemente importante el tiempo real [PIW10].

Sharmili et al. han propuesto otra metodología para implantes biónicos en la que se realiza un reescalado de las imágenes adquiridas por cámaras

externas antes de enviar la información a las neuronas conectadas al implante intraocular y también emplea la información del mapa de saliencia [SRS11].

Huang et al. han desarrollado una herramienta para motores de búsqueda de imágenes que permite la reordenación de las imágenes respuesta en función de la relevancia [HYZF10]. Para ello emplean dos mecanismos diferentes, una medida de la relevancia multiescala y la medida de la saliencia. Precisamente esa relevancia multiescala viene dada por los propios iconos de las imágenes respuesta que seleccionan los sujetos y la medida de saliencia es la que permite predecir cuales serán los iconos más salientes y por ello los primeros en captar la atención del sujeto.

Cada vez son más las aplicaciones en las que un agente virtual necesita interaccionar con el entorno virtual que lo rodea. Kokkinara y Oyekoya enlazan la atención visual con la realidad virtual mediante la creación de modelos de atención que hacen creíble la interacción de estos avatares en comparación con la observada en sujetos reales en juegos como Second Life [OSS09, KOS11]. Otro ejemplo es el empleo de la realidad virtual para el estudio del comportamiento de los mecanismos de atención en entornos virtuales, en los que se pueden emplazar de modo controlado objetos y distractores y capturar simultáneamente toda la información visual que el sujeto está percibiendo en cada instante. Rothkopf y col [RBSdB05] muestran como un modelo bayesiano permite identificar la tarea que está realizando el sujeto en este tipo de entornos basándose en la información del contexto y en las características que atraen su atención.

En el ámbito de la creatividad artística también podemos encontrar aplicaciones, como la de DeCarlo et al. [DS02] en la que mediante el empleo de modelos de atención obtienen imágenes estilizadas de carácter modernista que recuerdan a las obras de Toulouse-Lautrec.

Además de las aplicaciones en ámbitos tecnológicos o industriales, mediante todos estos ejemplos se puede apreciar hasta que punto las aplicaciones posibles pueden llegar a ser innumerables y abarcar múltiples y tan variados aspectos de la vida cotidiana.

Capítulo 2

Modelo de Saliencia Dinámica (AWSD)

El modelado de la atención visual y, especialmente, el de la saliencia *bottom-up* orientada a imágenes, ha supuesto un gran esfuerzo de desarrollo durante los últimos 20 años. Por ello, existen múltiples modelos de saliencia como el presentado en este trabajo y se están llevando a cabo diversos intentos de unificación tanto de los bancos de pruebas, como de las metodologías de evaluación empleadas, para así poder facilitar la comparación de las nuevas propuestas.

La percepción y las respuestas neurofisiológicas a los estímulos visuales dependen tanto del contexto espacial, ¿qué rodea al objeto?, como del temporal, ¿qué cambia a lo largo del tiempo? De entre todas las claves visuales de bajo nivel, el movimiento representa una de las pocas características sobre las que existe un amplio consenso acerca de su correlación directa con la atención visual [WH04], [CI06b]. Una alteración repentina en la dinámica de los objetos presentes en la escena es detectada rápidamente por el sistema visual humano (SVH), causando un efecto *pop-out*. Las evidencias fisiológicas confirman la existencia de regiones responsables del procesamiento del movimiento [FVE91, CS01, TPL+02]. Por ello es necesario incorporar el movimiento en las arquitecturas de los modelos de atención estáticos para ampliar su capacidad selectiva en secuencias dinámicas de imágenes.

Partiendo de la hipótesis de que los mecanismos de adaptación al contexto espacial son aplicables al contexto temporal, se obtiene un modelo *bottom-up* unificado, capaz de detectar saliencia tanto en imágenes estáticas como en vídeo, al que hemos denominado AWSD. La idea básica sobre la que se sustenta el modelo AWSD es que la saliencia, tanto estática como dinámica, se produce en aquellos puntos donde la energía local espacio-tiempo (medida indirecta de la alineación de la fase) posee la máxima desviación respecto a

la distribución media de esta característica en un espacio multiescala.

La energía local constituye un estadístico de alto orden que concentra gran cantidad de la información perceptualmente relevante. Para acceder a ella, el modelo utiliza el blanqueado (descrito en el Apéndice A) como un mecanismo muy simple que condensa parte de las implicaciones de la hipótesis de Barlow [Bar61]. En nuestro modelo, la reducción de la redundancia juega dos papeles diferenciados: en una primera etapa elimina las redundancias de segundo orden de la cadena de procesado y permite una adaptación de corto rango a la estadística de las imágenes de entrada y, por otra parte, permite aprender la distribución media de la energía local en un espacio multirresolución y simplificar la medida de disimilaridad en este espacio normalizado y decorrelacionado. Como resultado, el AWSD utiliza las características óptimas para cada vídeo, al contrario que otros algoritmos que utilizan un espacio de características fijo y subóptimo para todos los vídeos de la base de datos. El modelo computacional propuesto supone una hipersimplificación de los mecanismos realmente involucrados en el SVH, pero los resultados alcanzados tanto en la predicción de fijaciones como de efectos *pop-out* en vídeo e imágenes [LAGDFVP13], nos permiten afirmar que el blanqueado adaptativo debe jugar un papel relevante en la percepción de la saliencia.

Organización del capítulo

El modelo planteado en este trabajo presenta dos rutas claramente diferenciadas, estática y dinámica, para el procesamiento de los vídeos de entrada. En la sección 2.1 se analizarán la plausibilidad biológica de los componentes básicos de dichas rutas de procesamiento, secciones 2.1.1 y 2.1.2 para la componente estática y dinámica respectivamente. Por otra parte, en la sección 2.2 se realizará la descripción detallada de la arquitectura de dicho modelo mostrando el formalismo y los parámetros elegidos para cada uno de los elementos del modelo.

2.1. Plausibilidad biológica del AWSD

El SVH presenta gran cantidad de ajustes dinámicos y también ajustes relacionados con la experiencia. Lo que se denomina adaptación a corto plazo no siempre es fácilmente separable de los cambios de sensibilidad o ajustes de ganancia, del mismo modo que la adaptación a largo plazo es difícilmente separable del aprendizaje. El SVH adapta sus respuestas a las características locales y globales de cada imagen, mostrando una adaptación de corto plazo al contraste, al contenido de color o a las estructuras espaciales. Este proceso

ha sido observado desde los fotorreceptores y las células ganglionares hasta las células corticales [RR09, Koh07, SHD07].

La atención visual modula y prioriza la representación codificada de los datos visuales entrantes en el SVH, permitiendo que solo los elegidos alcancen la consciencia del individuo. Este hecho puede ser una consecuencia de la hipótesis de Barlow [Bar61] y Attneave [Att54] del *sensory coding*, que afirma que las neuronas deben codificar la información sensorial reduciendo el alto grado de redundancia de la señal entrante. Dado que las imágenes son altamente estructuradas, esta operación elimina únicamente las correlaciones existentes entre píxeles con apenas contenido informativo perceptualmente importante, dejando al descubierto las estructuras visualmente relevantes asociadas a estadísticos de alto orden [HHH09] (alineamiento en la fase, orientación, movimiento, etc.). Esta característica puede ser vista como una potencial vía de adaptación sensorial. El blanqueado, y por ello, la minimización de la redundancia, está condicionado por la necesidad de suprimir el ruido de entrada [AR92].

La retina es el primer lugar de la cadena de procesado del SVH en donde aparece el mecanismo de blanqueado [AR92],[GCF06],[DL14]. Las células ganglionares exhiben campos receptores *center-surround* (CS) cuya misión, puede ser similar al blanqueado de los datos, produciendo una recodificación de la señal de los fotoreceptores para eliminar las correlaciones, tanto en el espacio como en el tiempo y en el color. Este hecho es aproximado en el modelo AWSD en su primera etapa, que produce el blanqueado de las bandas de color emulando el procesado llevado a cabo en la retina y permitiendo una representación menos redundante y adaptada a la estructura estadística de la señal entrante (adaptación de corto plazo) [SO01].

La adaptación a corto plazo y sus efectos pueden subclasificarse en adaptación al contexto espacial, relacionada con las interconexiones neuronales, y adaptación al contexto temporal, que implica en cierto modo a la memoria. Ambos tipos de adaptación están íntimamente relacionados funcionalmente y en sus consecuencias perceptivas.

Tanto la adaptación al contexto espacial, como temporal y el hecho de que ciertas áreas de la corteza reflejen características estadísticas de las entradas visuales, motiva la creación de una aproximación de primer orden en la que la adaptación de largo plazo se puede aplicar a pequeñas escalas. Un ejemplo de esa representación son los campos receptores que responden de modo selectivo ante ciertas orientaciones o frecuencias de los estímulos presentes en la imagen.

Desde el punto de vista biológico, existen evidencias de que la adaptación a corto plazo permite a las neuronas adaptar sus respuestas a cambios diferentes en periodos muy cortos de tiempo, lo que permite reducir las inter-

dependencias entre neuronas, recalibrando de modo activo la sensibilidad de las neuronas en función de la experiencia. El modelo AWSD intenta reflejar esta capacidad mediante el agrupamiento de las imágenes de entrada en pequeños bloques a los que le aplica el procedimiento de blanqueado, simulando esta adaptación temporal de corto alcance.

2.1.1. Adaptación al contexto espacial

Tanto el control de ganancia como la normalización del contraste son mecanismos conocidos que dan soporte a la hipótesis de la decorrelación y blanqueado de las respuestas neuronales. Puesto que el color y la codificación espacial están relacionados con mecanismos neuronales diferenciados, la componente estática del modelo propuesto refleja ese mismo criterio procesándolos por separado.

En la parte estática del modelo, basada en el empleo de la variabilidad óptica, se realiza un blanqueado adaptativo de las diferentes escalas para cada una de las orientaciones elegidas de todo el campo visual [GDFVPD12]. De este modo se emplea un conjunto de componentes adaptado para cada escena específica. Esta aproximación se basa en características físicas como las que se describen a continuación.

Codificación del color

En estudios realizados con primates se ha observado que las respuestas de las neuronas de la región V1 presentan una gran selectividad al color [SH02, WSA03]. La representación mediante componentes de oposición de color, ya observada en LGN [RCC98], se recodifica posteriormente en la región V1. Esta nueva codificación está influenciada por los cambios del color del fondo y por ello modifica la percepción del color en función de la diferencia entre el contraste cromático del estímulo y el fondo.

En imágenes naturales en las que el color presenta una distribución muy restringida, se ha observado una evidente actuación de los mecanismos de adaptación, siendo dicha actuación claramente dependiente del observador [WM97, Web11]. En experimentos psicofísicos se ha observado que la adaptación al contexto produce cambios de sensibilidad que alteran la apariencia del color mediante la reducción del contraste entre colores que son semejantes a los ejes adaptados y separando los colores de aquellos estímulos que están alejados de los ejes adaptados. Un ejemplo de este comportamiento es la alteración que se produce en búsquedas de asimetrías de color en función del color del fondo [SH02].

En [GDLFVP12] se analizan las correlaciones existentes entre las diferentes componentes de varios espacios de color, RGB (rojo, verde, azul), LAB (luminancia y oposición de color), HSV (matiz, saturación, valor) y LMS (longitud de onda larga, media y corta). Para imágenes naturales se observa una gran correlación entre las tres componentes del espacio LMS, por lo que se sugiere como alternativa la obtención de las componentes decorrelacionadas mediante el empleo de PCA o ICA. Los mecanismos de plasticidad presentes durante la codificación de los *spikes*[1], parece estar relacionada con el análisis de componentes principales (PCA), que permite la obtención de la imagen blanqueada en el cerebro. Estos mecanismos permitirían realizar el análisis de componentes independientes (ICA), mediante la selectividad de grupos parciales de neuronas [SJT10].

En resumen el blanqueado de las componentes de color constituye un mecanismo biológicamente plausible de adaptación contextual que contribuye a una codificación eficiente del rango dinámico disponible a través de una representación adaptativa.

Codificación de las estructuras espaciales

Múltiples estudios demuestran la participación de la ruta *qué* en el procesamiento de la forma, dicha ruta va desde la corteza visual primaria (V1), a la secundara (V2), las áreas intermedias V3 y V4 y se dirige a la corteza inferotemporal (IT) [FVE91]. En estas regiones se observa una clara selectividad a diferentes orientaciones con patrones sencillos como líneas rectas o con formas compuestas más o menos complejas [NSRM13].

La adaptación contextual de las respuestas neuronales a las frecuencias espaciales y las orientaciones también ha sido observada en varios estudios psicológicos [Koh07, SHD07]. Existen variados ejemplos que muestran evidencias de la influencia del contexto más allá de los campos receptores. El ejemplo clásico de contornos ilusorios se ha visto que da lugar a una adaptación mediante un aumento de la selectividad en la orientación en múltiples áreas cerebrales. Otro ejemplo puede ser la exposición prolongada a un patrón vertical con un gran contraste, que da lugar a una disminución de la sensibilidad al contraste para dicha orientación durante un corto periodo de tiempo (segundos o incluso minutos). Esta adaptación ha sido observada incluso cuando los estímulos de referencia y test son percibidos por ojos diferentes. Esto sugiere que el proceso de adaptación se produce en áreas posteriores a V1, región en la que ya se da un procesamiento binocular.

[1] También conocidos como potenciales de acción, son los impulsos eléctricos que viajan a través de las membranas celulares para trasladar la información de un lugar a otro.

Todo esto parece sugerir que los mecanismos de adaptación no se limitan a un control de la ganancia, como el descrito en la retina, sino que las diferentes regiones que participan en el procesamiento de las estructuras espaciales, pueden realizar ajustes adaptando su rango dinámico a las características de los estímulos percibidos [Koh07].

Separación de características

Desde el punto de vista psicofísico y neurofisiológico se sugiere la presencia de mecanismos de procesamiento del color y estructuras espaciales independientes. Mientras que la codificación del color tiene su origen en las estructuras de la retina y LGN [WSA03], para el caso de las estructuras espaciales se cree que ocurre principalmente en la corteza visual [GKS05]. De ahí que la componente estática del modelo AWSD asuma el procesamiento por separado del color y las características espaciales. Puesto que la técnica de blanqueado es altamente dependiente del número de componentes blanqueadas, esta separación además de fundamentarse en la semejanza con la biología, también supone una ventaja en cuanto al coste computacional de la misma [GDLFVP12].

2.1.2. Adaptación al contexto temporal

Está muy documentada la existencia de regiones cerebrales relacionadas con la codificación del movimiento, en cuanto a la acción, o sea la codificación de los movimientos que el cerebro programa [PIIK05], y en particular la codificación de los movimientos oculares. Pero en cuanto a la codificación de los movimientos presentes en una escena vista por un observador, al igual que la codificación de otras múltiples características como el color o la orientación, todavía no hay una clara asignación del papel de cada una de estas áreas, y son varias las hipótesis que se barajan. La relación entre los mecanismos responsables de guiar la atención y los de control de los movimientos oculares se sustentan en la teoría premotora de la atención de Rizzolatti [RRDU87].

Para que un humano perciba el movimiento no es necesario que el estímulo se mueva de modo continuo, sino que bajo determinadas condiciones, un estímulo puede parecer en movimiento ya que el SVH rellena la información de esta presentación discreta, pudiendo existir hasta varios grados de ángulo visual entre posiciones consecutivas, y grandes intervalos temporales de hasta 400ms. Esto es lo que se denomina movimiento aparente [SHD07]. La componente dinámica del modelo AWSD identificará como movimiento todo aquello que esté dentro de la ventana temporal asociada a los bloques procesados (p. ej. en vídeos de 15fps (66ms), un objeto que se desplace en los 7

fotogramas siguientes, o sea 462ms. será detectado).

La adaptación contextual asociada al movimiento ya fue descrita en experimentos realizados con células ganglionares de diferentes mamíferos por Goldstein (1957), Hubel y Wiesel (1959) o Barlow y Hill (1963). En esos trabajos se describe la selectividad de dichas células ante diferentes tipos de movimiento y a su vez el efecto de relajación extrema que sucede tras la activación continuada de las mismas, dando lugar a cambios en la percepción del movimiento. En algunos casos el resultado es la percepción de movimiento en la dirección contraria [SP67, MZH12], o en otros, como el descrito por Goldstein, se produce una disminución de la velocidad aparente. De modo semejante a la descripción de las características espaciales (sección previa), los mecanismos de adaptación también parecen poder realizar ajustes adaptando su rango dinámico a características asociadas al movimiento [Koh07]. El modelo AWSD refleja esta selectividad mediante el empleo de los filtros 3D descritos por [DFVP08, DPFV06]. Inicialmente estos filtros fueron empleados para la extracción de regiones o la segmentación de objetos. Mediante el cálculo de la amplitud de las respuestas de los filtros espacio temporales, sintonizados a diferentes escalas y orientaciones, se puede obtener una estimación del movimiento con una gran sensibilidad a la orientación espacial, velocidad y dirección del movimiento. Estas son precisamente las características que el modelo AWSD pretende reflejar del SVH.

Todo esto plantea la posibilidad de una inhibición lateral de las componentes de movimiento al igual que sucede con la inhibición lateral de las respuestas neuronales en el dominio espacial para la intensidad, la inhibición de orientaciones o la inhibición de las componentes de frecuencia. Esta inhibición de nuevo permitiría optimizar el limitado rango dinámico de las neuronas [PIIK05]. El modelo integra esta estimulación lateral mediante la construcción de filtros espaciotemporales con un solape de la información.

Codificación del movimiento

El movimiento es una de las características preatentivas que ha sido ampliamente utilizada en estudios psicofísicos. En ellos se pueden distinguir tres propiedades asociadas al movimiento; el parpadeo, la dirección de movimiento y la velocidad de movimiento [HH05b]. Ya en la retina, existe una diferenciación entre las estructuras vinculadas al procesamiento de las características espaciales, con mayor resolución espacial, descritas como células P -Parvo- y que dan lugar al flujo parvocelular descrito en la sección 1.1.1 frente a las estructuras vinculadas al procesamiento de características temporales, células M de la retina -Magno- con una mayor resolución temporal y que dan lugar al flujo magnocelular. En cuanto a la distribución de dicha resolución

temporal, presenta una estructura compleja. Por una parte la sensibilidad al movimiento decrece a medida que nos separamos de la fóvea mientras que la sensibilidad ante los parpadeos es mayor en la periferia [JB85].

En estudios de la corteza primaria realizados en primates se ha observado la participación de varias áreas en el procesamiento del movimiento. Dichas áreas presentan una conectividad jerárquica que se extiende desde V1 hacia MT, MST y 7a [FVE91, CS01, TPL+02]. Toda esta estructura jerárquica destinada específicamente a la identificación del movimiento, aporta una idea de la relevancia del movimiento para el complejo SVH.

Por ejemplo, el área MT presenta la misma selectividad a la velocidad que V1, pero con mayores campos receptores. Además MT es selectiva al ángulo entre el movimiento local y el gradiente de la velocidad. MST es selectiva a cambios complejos como la expansión o contracción y la rotación. La región 7a es selectiva ante traslaciones, movimientos espirales, rotación completa de la escena y movimientos radiales de expansión y contracción pero con mayores campos receptores.

Las células del cortex extriado de V1 muestran una selectividad hacia la velocidad local y la dirección del movimiento. El modelo AWSD intenta reflejar esta estructura, de un modo muy simplificado, mediante el empleo de bloques de imágenes, que se procesan mediante filtros espaciotemporales para determinar la contribución de las características de movimiento al mapa de saliencia global [Hee88]. La elección del tamaño del bloque, el número de orientaciones y el número de escalas determinará la selectividad ante los cambios locales de velocidad[2].

En cuanto a la integracion de esta última característica, el movimineto, con las previamente descritas, de nuevo, basándose en la separación de procesamiento que realiza el cerebro, la propuesta del modelo AWSD opera de modo análogo. Desde el punto de vista biológico, por una parte está el denominado flujo *qué*, responsable del procesamiento de color y estructuras espaciales, mientras que por otro lado está el flujo *dónde* responsable del procesamiento del movimiento. Asimismo, la componente estática del modelo es la responsable de analizar color y características espaciales, y será la componente dinámica la responsable del procesamiento de las características de movimiento.

[2]El valor empleado por Tsostos [TPL+02] es de 5 fotogramas, para el modelo AWSD se ha estimado el valor óptimo de $N = 7$.

2.2. Arquitectura del modelo AWSD

La propuesta que aquí se plantea consiste en la combinación de una componente estática, basada en la adaptación contextual estimando la variabilidad óptica en un dominio atemporal, con una componente dinámica, basada en la adaptación contextual a corto plazo teniendo en cuenta la variable temporal. El modelo propuesto pretende ofrecer una solución eficiente para sistemas basados en atención visual. Por ello es un sistema modular, con una estructura jerárquica y con un número limitado de etapas, características que lo hacen biológicamente plausible teniendo en cuenta las capacidades del SVH. Es importante indicar que hay múltiples partes que no está clara todavía su implementación biológica y por ello no pretende ser un fiel reflejo de la estructura del SVH.

La arquitectura del modelo AWSD (figura 2.1) está dividida en dos caminos procesales: 1) la *ruta espacial* destinada a la obtención de la saliencia estática de cada fotograma y, 2) la *ruta temporal* que obtiene la saliencia dinámica de cada fotograma a partir de un bloque de n fotogramas consecutivos. Ambas cadenas procesales comparten el objetivo de explicitar la información perceptualmente relevante y ecualizar los espacios resultantes sobre los que se obtiene la saliencia. Las dos vías procesales comparten dos etapas, la etapa cromática y la etapa de salida. La primera se encarga de blanquear las bandas de color, lo que permite una adaptación del espacio cromático a la estructura estadística propia de cada imagen de entrada y la segunda integra los mapas estáticos y dinámicos para cada fotograma. A continuación detallamos cada una de ellas:

2.2.1. Etapa de entrada

Su misión es construir volúmenes de datos espacio-temporales de n fotogramas consecutivos $\{\mathbf{F}_f^{rgb} \mid f \in \{1, \cdots, n\}\}$, donde asumimos, sin pérdida de generalidad, que cada fotograma $\mathbf{F}_f^{rgb}$ es una imagen RGB con una resolución espacial $N \times N$. Elegimos un valor de $n = 7$ como un buen compromiso entre la *causalidad* del modelo ($n \downarrow$) y la necesidad de lograr un análisis global y estable del movimiento ($n \uparrow$)[3].

Esta estructura de bloques consecutivos da lugar a transiciones entre bloques (bordes temporales). Para minimizar el efecto de esta concatenación, durante el procesamiento de cada bloque, se emplean varios fotogramas previos y posteriores al mismo, que son eliminados al final del procesamiento.

[3]El tamaño bloque óptimo elegido es similar a los 5 fotogramas empleados en [TPL+02].

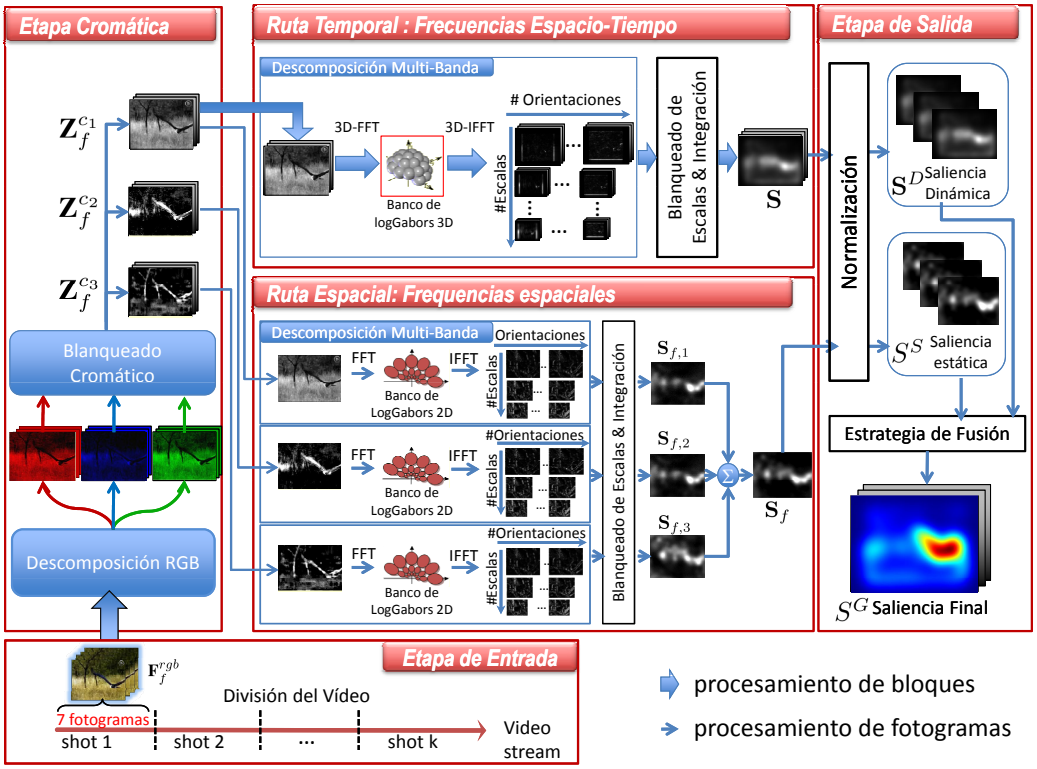


Figura 2.1: Diagrama general del flujo de datos a través del modelo AWSD.

2.2.2. Etapa de procesado cromático

Esta etapa implementa un mecanismo de adaptación contextual para lograr una representación más eficiente de la información cromática. Para cada fotograma $\mathbf{F}_f^{rgb}$ $(f \in \{1, \cdots, n\})$ se blanquean las bandas cromáticas RGB, logrando un nuevo espacio denotado por $\mathbf{Z}_f^{c_i}$ con $c_i \in \{1, 2, 3\}$ (ver apéndice A).

En este punto, la cadena procesal se divide en dos partes: el flujo responsable de la saliencia dinámica (denotado en la figura 2.1 como *procesamiento de bloques*), que analiza el bloque completo de la componente cromática principal $\{\mathbf{Z}_f^{c_1} \mid f \in \{1, \cdots, n\}\}$ y el flujo responsable de la saliencia estática (denotado como *procesamiento de fotogramas*), que procesa individualmente las bandas cromáticas de cada fotograma f del bloque $(\mathbf{Z}_f^{c_1}, \mathbf{Z}_f^{c_2}, \mathbf{Z}_f^{c_3})$. Esta etapa se repite con la llegada de cada nuevo bloques de imágenes, distribuyendo los nuevos datos hacia los dos flujos correspondientes.

2.2.3. Construcción del mapa de saliencia espacial

El mapa global de saliencia estática se obtiene integrando los mapas de saliencia estáticos de cada una de las componentes de color $(\mathbf{Z}_f^{c_1}, \mathbf{Z}_f^{c_2}, \mathbf{Z}_f^{c_3})$. Los siguientes pasos explicitan el proceso:

1) Construcción del espacio multiescala de la energía local en cada banda cromática $(\mathbf{Z}_f^{c_1}, \mathbf{Z}_f^{c_2}, \mathbf{Z}_f^{c_3})$: el espacio multirresolución de la energía local se logra mediante un banco de filtros logGabor (con una gran sensibilidad a la orientación espacial, velocidad y dirección del movimiento). La función de transferencia de cada filtro, en el dominio de la frecuencia, es una función real:

$$\Upsilon\left(r_o, \theta_o, \sigma_r, \sigma_\theta\right) = \exp\left\{-\frac{\left(\log\left(r/r_o\right)\right)^2}{2\left(\log\left(\sigma_r/r_o\right)\right)^2}\right\} \cdot \exp\left\{-\frac{\left(\theta-\theta_o\right)^2}{2\sigma_\theta^2}\right\} \tag{2.1}$$

donde r_o es la frecuencia central del filtro, θ_o es la orientación del filtro, σ_r y σ_θ son las anchuras radial y angular de la envolvente de la función en coordenadas polares, respectivamente.

La convolución del filtro $\Upsilon_j\left(r_o, \theta_o, \sigma_r, \sigma_\theta\right)$ con una imagen real $Z_f^{c_i}(x,y)$ produce una respuesta compleja $(O_j^{par}(x,y) + i \cdot O_j^{impar}(x,y))$ cuya norma es la energía local en cada punto:

$$E_j^{\theta_o}(x,y) = \sqrt{O_j^{par}\left(x,y\right)^2 + O_j^{impar}\left(x,y\right)^2} \tag{2.2}$$

donde:

$$O_j^{par}\left(x,y\right) = Re\left(\mathcal{F}^{-1}\left\{\Upsilon_j\left(r_o, \theta_o, \sigma_r, \sigma_\theta\right) \cdot \mathcal{F}\left\{Z_f^{c_i}(x,y)\right\}\right\}\right)$$
$$O_j^{impar}\left(x,y\right) = Img\left(\mathcal{F}^{-1}\left\{\Upsilon_j\left(r_o, \theta_o, \sigma_r, \sigma_\theta\right) \cdot \mathcal{F}\left\{Z_f^{c_i}(x,y)\right\}\right\}\right)$$

donde $\mathcal{F}\{\cdot\}$ denota la transformada de Fourier.

Dado que la componente cromática principal $\mathbf{Z}_f^{c_1}$ contiene la información con máxima variabilidad (contenido en frecuencias altas) con respecto a $\mathbf{Z}_f^{c_2}$ y $\mathbf{Z}_f^{c_3}$ (contenido en bajas frecuencias), se ha decidido utilizar dos bancos de filtros diferenciados. Para la componente $\mathbf{Z}_f^{c_1}$ la parametrización es:

- El plano de frecuencia está dividido en $n_o = 4$ orientaciones y $n_s = 7$ escalas.
- El primer filtro de cada orientación se sitúa con una longitud de onda de $\lambda_{min} = 3$ píxeles, el resto se distribuye a incrementos de una octava $\lambda_i = 2^{(i-1)} \cdot \lambda_{min}$ con $i \in \{1, \cdots, n_s\}$.

- La σ_r se determina, para cada banda, para lograr un ancho de banda de dos octavas sobre una escala logarítmica.
- La σ_θ es de 22.5° para todas las orientaciones.

De modo análogo, para las componentes $\mathbf{Z}_f^{c_2}$ y $\mathbf{Z}_f^{c_3}$:

- El plano de frecuencia está dividido en $n_o = 4$ orientaciones y $n_s = 5$ escalas.
- El primer filtro de cada orientación se sitúa con una longitud de onda de $\lambda_{min} = 6$ píxeles. El resto se distribuye a incrementos de una octava $\lambda_i = 2^{(i-1)} \cdot \lambda_{min}$ con $i \in \{1, \cdots, n_s\}$.
- La σ_r se determina para cada banda para lograr un ancho de banda de dos octavas sobre una escala logarítmica.
- La σ_θ es de 13.2° para todas las orientaciones.

2) Para cada banda cromática c del fotograma f, el mapa de saliencia se cuantifica como:

$$\mathbf{S}_{f,c} = \sum_{o=1}^{n_o} \left(\frac{\sum_{s=1}^{n_s} \left(\mathbf{Z}_{f,c,o,s}^{E_\theta} - \bar{\mathbf{Z}}_{f,c,o}^{E_\theta} \right)^2}{\max_{N_{Pix}} \left\{ \sum_{s=1}^{n_s} \left(\mathbf{Z}_{f,c,o,s}^{E_\theta} - \bar{\mathbf{Z}}_{f,c,o}^{E_\theta} \right)^2 \right\}} \right) \tag{2.3}$$

donde el conjunto de matrices $\left\{ \mathbf{Z}_{f,c,o,s}^{E_\theta} \mid s = 1 \dots n_s \right\}$ denota al espacio multiescala blanqueado de la energía local para una determinada orientación o, en la banda cromática c, correspondiente al fotograma f. Siendo $\bar{\mathbf{Z}}_{f,c,o}^{E_\theta} = \sum_{s=1}^{n_s} \mathbf{Z}_{f,c,o,s}^{E_\theta} / n_s$, la distribución media de la energía local en dicha orientación y $\{N_{Pix}\}$ denota el conjunto de píxeles del fotograma f ($card\,(N_{Pix}) = N^2$). En la figura 2.2 se muestra este procedimiento sobre un fotograma de ejemplo para la componente cromática $\mathbf{Z}_f^{c_2}$.

Llegado este punto del procesamiento ya disponemos de un mapa de saliencia para cada uno de los tres canales cromáticos blanqueados de cada fotograma.

3) Obtención del mapa de saliencia estático global. Se consigue mediante la integración de los tres mapas $\mathbf{S}_{f,c}$ previo suavizado gaussiano:

$$\mathbf{S}_f = \sum_{c=1}^{3} \mathcal{G}(x, y, \sigma) \otimes \mathbf{S}_{f,c} \tag{2.4}$$

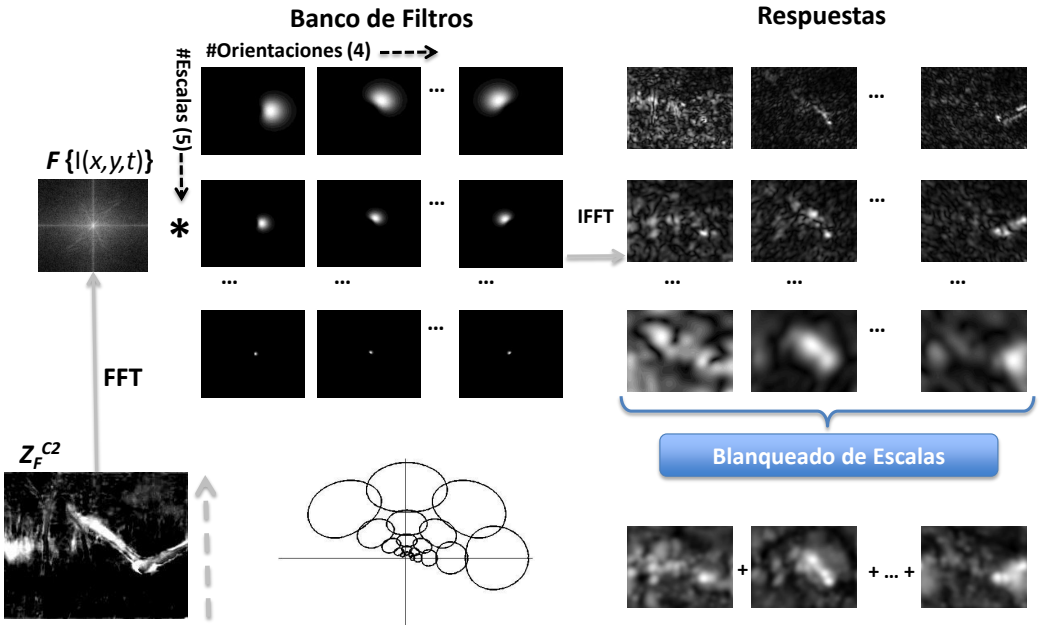


Figura 2.2: Detalle del blanqueado de escalas y suma de orientaciones para la componente cromática $\mathbf{Z}_f^{c_2}$ de la ruta estática del AWSD con el vídeo BBC-wildlife-eagle. El esquema central muestra la configuración de los filtros.

Finalmente, la función $S^S(x, y, f)$ denota al conjunto $\{\mathbf{S}_f \mid f \in \{1, \cdots, n\}\}$.

2.2.4. Construcción del mapa de saliencia temporal

El mapa de saliencia dinámico se obtiene sobre el espacio blanqueado de la energía local aplicando la norma $\mathcal{L}^2$, respecto de la distribución media, en cada *voxel*. Sólo se utiliza la componente cromática de mayor contenido frecuencial $\mathbf{Z}_f^{c_1}$ para obtener la energía local espacio-tiempo, (tal como ya se ha descrito, $\mathbf{Z}_f^{c_1}$ es el resultado del blanqueado cromático mediante el análisis PCA que devuelve las componentes ordenadas, siendo la primera la que contiene mayor variabilidad). Los pasos seguidos son:

1) Construcción del espacio multiescala $\{\mathbf{Z}_f^{c_1} \mid f \in \{1, \cdots, n\}\}$ de la energía local del bloque: la obtención de la energía local, en un volumen espacio-tiempo, se logra mediante un banco de filtros logGabor 3D. La función de transferencia de estos filtros, T, se puede diseñar como el producto de una función logGabor unidimensional sobre la frecuencia radial y una Gaussiana en la *distancia angular*, la cual es rotacionalmente simétrica en

coordenadas esféricas [DFVP08]:

$$T(\rho_i, \phi_i, \theta_i, \sigma_\rho, \sigma_\alpha) = \exp\left\{-\frac{(log(\rho/\rho_i))^2}{2(log(\sigma_\rho/\rho_i))^2}\right\} \cdot \exp\left\{-\frac{\alpha(\phi_i, \theta_i)^2}{2\sigma_\alpha^2}\right\} \quad (2.5)$$

donde $(\rho_i, \phi_i, \theta_i)$ indican la frecuencia central del filtro y σ_ρ, σ_α son las desviaciones estándar, $\alpha(\phi_i, \theta_i) = acos(\mathbf{f}\cdot\mathbf{v}/\|\mathbf{f}\|)$ con $\mathbf{v} = (\cos\phi_i \cdot \cos\theta_i, \cos\phi_i \cdot \sin\theta_i, \sin\phi_i)$ y $\mathbf{f}$ representa las coordenadas Cartesianas de un punto en el espacio frecuencial 3D. La componente azimutal del filtro, ϕ_i refleja la orientación espacial en un determinado fotograma mientras que θ_i está relacionada con la velocidad y dirección del movimiento.

La convolución del volumen espacio-tiempo con el filtro $T_i(\rho_i, \phi_i, \theta_i, \sigma_\rho, \sigma_\alpha)$ i-ésimo produce una respuesta compleja cuya norma es la energía local del volumen 3D en cada voxel, que denotaremos por $E_i^\alpha(x, y, t)$ (en analogía con la ec. 2.2).

Los parámetros del banco de filtros log-Gabor 3D se han elegido para lograr una amplia cobertura del espacio de escalas, un muestreo uniforme del espacio de orientaciones y filtros selectivos en orientación (velocidades y direcciones del movimiento):

- El volumen frecuencial se dividió en $n_o = 23$ orientaciones (o 23 $\alpha_i = \alpha(\phi_i, \theta_i)$) y $n_s = 6$ escalas.
- El primer filtro de cada orientación se sitúa en una longitud de onda $\lambda_{min} = 3$ píxeles y el resto se distribuye a incrementos de una octava $\lambda_i = 2^{(i-1)} \cdot \lambda_{min}$ con $i \in \{1, \cdots, n_s\}$.
- θ_i se muestrea uniformemente, mientras el número de ϕ_i decrementa con la elevación para mantener la densidad de filtros constante imponiendo la condición igualdad en longitudes de arco entre muestras adyacentes de ϕ_i sobre una esfera de radio unidad.
- La σ_ρ se determina para cada banda logrando un ancho de dos octavas en escala logarítmica.
- La σ_α es de 25° para todas las orientaciones.

En la figura 2.3 se muestra este procedimiento sobre un bloque de fotogramas de ejemplo para la ruta dinámica partiendo de $\mathbf{Z}_f^{c_1}$.

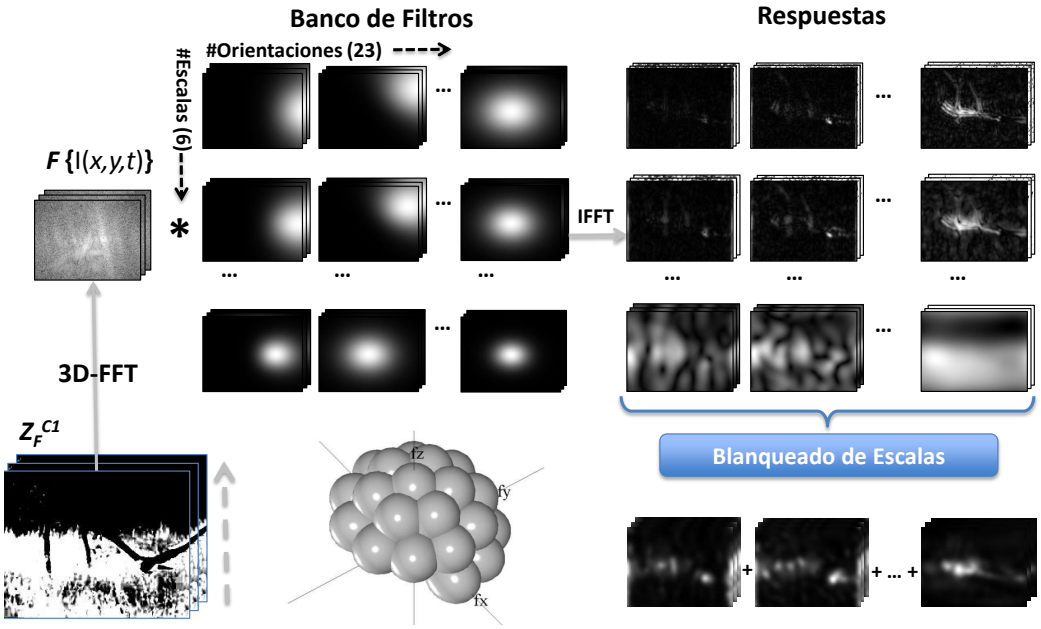


Figura 2.3: Detalle del proceso de blanqueado de escalas y suma de orientaciones de la ruta dinámica del AWSD con el vídeo BBC-wildlife-eagle. El esquema central muestra la configuración de los filtros.

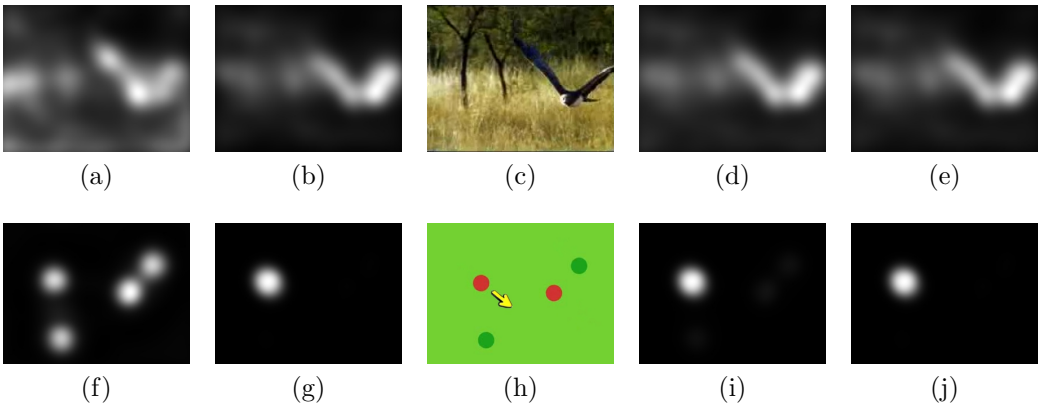
2) La obtención del mapa de saliencia dinámica se cuantifica como:

$$\mathbf{S} = \sum_{o=1}^{n_o} \left(\frac{\sum_{s=1}^{n_s} \left(\mathbf{Z}_{o,s}^{E_\alpha} - \bar{\mathbf{Z}}_{o}^{E_\alpha} \right)^2}{\underset{N_V}{max} \left\{ \sum_{s=1}^{n_s} \left(\mathbf{Z}_{o,s}^{E_\alpha} - \bar{\mathbf{Z}}_{o}^{E_\alpha} \right)^2 \right\}} \right) \tag{2.6}$$

donde el conjunto de matrices $\left\{ \mathbf{Z}_{o,s}^{E_\alpha} \mid s = 1 \dots n_s \right\}$ denota el espacio de escalas de la energía local blanqueado para una determinada orientación o, $\bar{\mathbf{Z}}_{o}^{E_\alpha} = \sum_{s=1}^{n_s} \mathbf{Z}_{o,s}^{E_\alpha} / n_s$ es la distribución media de la energía local en dicha orientación, y $\{N_V\}$ denota el conjunto de voxels del volumen ($Card\,(N_V) = 7 \cdot N^2$).

3) Obtención del mapa de saliencia dinámico global: Finalmente, la función $S^D(x, y, f)$ denota la matriz $\mathbf{S}$, previo suavizado gaussiano 3D. Representa los mapas finales de saliencia dinámica de todos los fotogramas del bloque.

$$\mathbf{S}^D = \mathcal{G}^{3D}(x, y, t, \sigma) \otimes \mathbf{S} \tag{2.7}$$



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

(f) (g) (h) (i) (j)

Figura 2.4: Fotograma (c) 218 del BBC-wildlife-eagle y (h) 210 del vídeo SCE-Bur5. Los correspondientes mapas de saliencia para los modelos (a)/(f) estático, (b)/(g) dinámico, (d)/(i) ponderado y (e)/(j) prioritario.

2.2.5. Etapa de fusión

Para producir los valores finales de saliencia, $S^G(x, y, f)$, los mapas de saliencia estáticos y dinámicos deben colaborar para producir un mapa de saliencia único. Adoptamos un enfoque semejante al de A. Bur [Bur09], en el que se prioriza al mapa con un contenido informativo claramente superior al resto. En caso contrario, ambos mapas se integran mediante un esquema competitivo implementado mediante una suma ponderada. El factor de peso es proporcional a la relevancia informativa que porta cada mapa y lo obtenemos como la inversa de número de máximos locales del mapa de saliencia que superan un determinado umbral $T = m + \sigma$, donde m y σ son estimadas como la media y la desviación típica del correspondiente mapa. Formalmente, este modelo viene dado por:

$$\begin{cases} S^S(x,y,f) & \text{si} \quad \frac{n_d}{n_s} > \varepsilon \\ S^D(x,y,f) & \text{si} \quad \frac{n_s}{n_d} > \varepsilon \\ \frac{1}{n_s}\mathcal{N}(S^S(x,y,f)) + \frac{1}{n_d}\mathcal{N}(S^D(x,y,f)) & \text{otro caso} \end{cases} \tag{2.8}$$

donde $\mathcal{N}(\cdot)$ denota una operación de normalización del rango dinámico entre $[0, 1]$, n_s y n_d son el número de máximos locales en los mapas $S^S(x, y, f)$ y $S^D(x, y, f)$, respectivamente. El umbral ϵ se ha determinado como el promedio de máximos estáticos/dinámicos sobre la base de datos CITIUS-VDB ($\epsilon = 1.6$). Este valor es fijo y no se ha modificado durante la aplicación del modelo a otras bases de datos.

La figura 2.4 muestra el resultado obtenido empleando solamente la componente estática del modelo AWSD, figuras 2.4a/2.4f y la componente dinámi-

ca, figuras 2.4b/2.4g, frente al resultado de emplear dos estrategias de integración diferentes, la ponderada, figuras 2.4d/2.4i y la estrategia prioritaria, 2.4e/2.4j, aplicadas sobre dos vídeos de ejemplo en los que están presentes múltiples características, incluido el movimiento.

En el segundo ejemplo la componente estática del modelo, figura 2.4f, extrae la saliencia asociada a las características de color, que compiten con el desplazamiento de los objetos, identificados claramente por la componente dinámica, figura 2.4g. Tanto el esquema ponderado como el prioritario devuelven resultados semejantes, figuras 2.4i y 2.4j.

En la ecuación 2.8 se incluye un proceso de normalización, ya que los rangos dinámicos pueden ser diferentes para las características estáticas y dinámicas. El empleo de una estrategia de normalización u otra puede amplificar pequeños valores sobresalientes de un nivel de ruido bajo, mientras que un proceso competitivo tenderá a sobreponderar aquellos mapas con niveles más altos de saliencia. Tal como muestran varios autores [IK99][OBH06], este paso determinará la contribución de las características empleadas. Durante este trabajo se han probado dos metodologías diferentes, la normalización iterativa y la normalización lineal integral, optando finalmente por la segunda por varios motivos; el coste computacional de la normalización integral es menor que el de la normalización iterativa, y al contrastar los modelos con la base de datos de prueba descrita en la sección 3.2, la normalización integral ofrecía mejores resultados.

En la figura 2.5 se muestra un fotograma de ejemplo en el que se han empleado ambas normalizaciones, se muestra también la imagen obtenida tras el proceso de blanqueado y *suma* de orientaciones, figura 2.5b.

Tal como se aprecia en la figura 2.5d el modelo que emplea la normalización integral es más semejante al modelo de atención humano de ese mismo fotograma, figura 2.5c, que el de la figura 2.5e. En ambos casos se ha aplicado además un filtro de suavizado Gaussiano a los mapas. Esta mejor adecuación de los modelos que emplean filtros de suavizado ya ha sido descrita por Judd et al. [JDT12].

Tanto durante la construcción del mapa de saliencia espacial como del temporal, los mapas que surgen del proceso de blanqueado han de ser integrados antes de proceder a la etapa de fusión. El empleo de la suma de las componentes de orientación, o del máximo se relaciona con las denominadas *hipótesis de suma de características* y la *hipótesis de V1*, descritas por Koene et al. [KZ07]. La primera afirma que la información de los estímulos es procesada en diferentes mapas asociados a características visuales únicas, que a continuación son sumadas. En contraposición, conforme a la *hipótesis de V1*, la información de diferentes características no se suma, sino que se determina qué agrupaciones de características son identificables, y a continuación, se

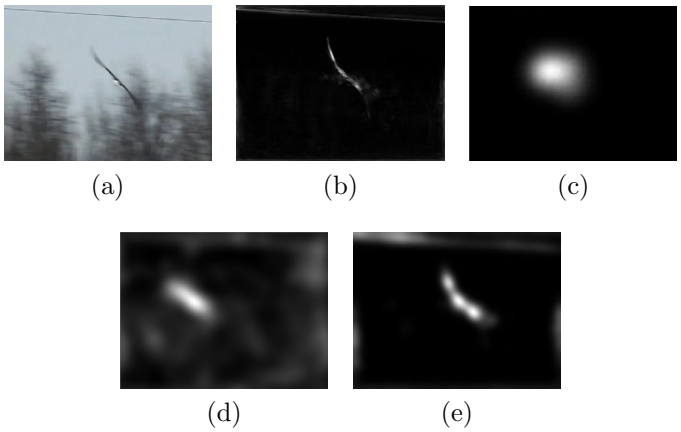


Figura 2.5: (a) Fotograma 15 del vídeo RCD-Eagle de la BD de USC, (b) el mapa obtenido tras el blanqueado, (c) modelo de saliencia humano para ese fotograma y los mapas tras la normalización (d) integral o (e) iterativa.

crea el mapa global mediante un proceso competitivo [TZKM11]. Koene et al. plantean además el empleo de mapas asociados a dobles características antes del proceso de integración global (color-orientación, movimiento-orientación).

En este trabajo se ha optado por realizar dos etapas de integración independientes, para las características estáticas y dinámicas respectivamente, y posteriormente unificarlos al final mediante el esquema prioritario.

Ejemplos de aplicación directa del AWSD

Aunque, en general, las técnicas de identificación de regiones salientes están enfocadas hacia la creación de preprocesamientos para la realización de tareas más complejas, es posible mediante la aplicación directa de la metodología descrita en este capítulo obtener, de modo inmediato, aplicaciones básicas usando la información de saliencia. Un ejemplo puede ser el reencuadre de escenas en las que predomina el movimiento.

La metodología empleada consiste en identificar la parte de la imagen original que contenga la mayor cantidad de información *relevante*, descartando aquellas regiones de contenido uniforme carentes de información, apoyandose para ello en el mapa de saliencia del modelo AWSD. Este ejemplo se muestra en los fotogramas de la figura 2.6 en los que se ha aplicado esta técnica sobre el vídeo RCD-Eagle.

La implementación del algoritmo de reencuadre, a pesar de la sencillez, muestra el gran potencial de las metodologías que hacen uso de la información de saliencia. En el ejemplo anterior se ha partido de un vídeo de resolución

Figura 2.6: Empleo del modelo AWSD para reencuadre de escenas. Se muestra un fotograma de ejemplo para el vídeo de prueba RCD-Eagle

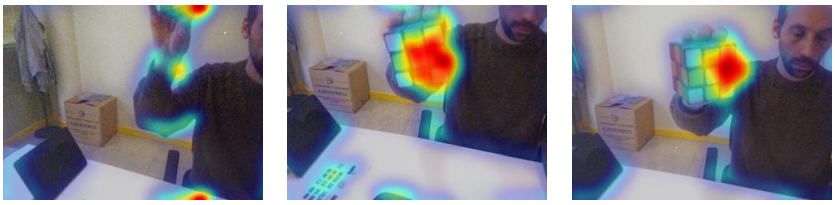


Figura 2.7: Funcionamiento del modelo AWSD durante el seguimiento de un objeto en movimiento con una camara web.

320x240 y se le ha impuesto el criterio de que la región enfocada fuese de 160x120 empleando la información del mapa de saliencia del modelo AWSD. El resultado se consigue mediante un procedimiento de maximización de la información contenida en el rectángulo del tamaño deseado. Para evitar pequeños saltos locales se ha realizado también un filtrado anisotrópico [LDP04] sobre las posiciones XY determinadas por el algoritmo de maximización. Esto mejora la apariencia final del resultado, dando lugar a una sensación de continuidad durante la persecución del objetivo.

Otro ejemplo de aplicación directa del modelo AWSD es el *seguimiento mediante cámara web* de objetos *salientes*, figura 2.7. Para esta implementación ha sido necesario relajar los parámetros del modelo AWSD. Se han reducido el número de orientaciones y escalas, y se ha empleado la componente dinámica para así poder responder ante los cambios en la escena en tiempos razonables. El modelo es capaz de analizar en estas condiciones un fotograma por segundo a baja resolución. El funcionamiento del modelo se basa en la identificación de la región de mayor saliencia sobre una malla de nueve elementos y, en función de la región con mayor saliencia, realiza los movimientos de la cámara web en esa dirección.

Capítulo 3

Bases de datos públicas, nuestra propuesta CITIUS y metodologías de evaluación

La evaluación de la calidad de los diferentes algoritmos empleados para la obtención de la saliencia visual es, en sí misma, una tarea compleja. Los mecanismos *top-down* pueden ser modulados por factores como la memoria, la edad, el sexo, factores culturales o la experiencia, afectando de modo diferente durante la realización de diferentes tareas; observación libre, identificación, búsqueda o reconocimiento. No hay una única medida preestablecida para determinar la calidad del resultado de un algoritmo [AM09]. En general para obtener la validez de un modelo de saliencia *bottom-up*, dicho modelo se compara con los resultados obtenidos experimentalmente, asumiendo la *hipótesis de saliencia visual*. Para ello se diseñan tareas en las que se procura minimizar los efectos *top-down* y se evalúa la capacidad de los modelos para predecir las fijaciones de los sujetos durante la observación libre de vídeos, o sea sin ningún objetivo específico.

La bibliografía reciente está repleta de diversos métodos y técnicas para realizar estas comparaciones [SM09b, PS00, BI13]. Pero, en muchos casos, no es posible reproducir los experimentos, ya que no están disponibles el código, las bases de datos o los parámetros que se han empleado para obtener los resultados mostrados en la publicación correspondiente.

En este trabajo se ofrece, tanto un conjunto de datos organizado y estructurado de modo análogo para todas las BD, como un entorno sobre el que poder añadir nuevos modelos computacionales; de modo que sea realmente sencilla su comparación con los resultados de modelos previamente evaluados.

Organización del capítulo

Este capítulo se centra en la descripción de las BD que han sido empleadas en este trabajo, sección 3.1. A continuación se detallan los problemas encontrados en estas BD externas y se describe nuestra propuesta, la BD CITIUS, sección 3.2 y las particularidades de la misma que la hacen diferente. También se muestra la clasificación de los movimientos oculares desde un punto de vista computacional, así como el procedimiento empleado para crear los mapas de atención visual humanos. A continuación, en la sección 3.3, se describen las metodologías candidatas para la evaluación del modelo detallado en el capítulo 2 frente a las bases de datos anteriormente descritas y las que finalmente han sido seleccionadas. En la última parte, sección 3.4, se indicará cómo se han evitado los efectos debidos al sesgo central, bordes y la evaluación de la fiabilidad temporal.

3.1. Metodología experimental

El procedimiento de creación de las bases de datos para su posterior análisis, siguiendo la descripción de los propios autores, ha sido semejante para las seis BD empleadas en este trabajo. A continuación detallaremos las condiciones en las que los sujetos han realizado los experimentos para la primera base de datos de prueba, la BD CRCNS.

3.1.1. Base de datos de vídeos CRCNS

La base de datos CRCNS, así como sus resultados y el código fuente, están disponibles a través de la web del ILAB de la Universidad del Sur de California [UoSC12]. Los vídeos de esta base de datos tienen muy baja calidad y presentan un claro efecto de entrelazado, ya que han sido inicialmente utilizados en un experimento cuya finalidad era observar la influencia de estos artefactos sobre un modelo de saliencia dinámica [CI06a, CI06b, IB05, IB09, Itt05]. El experimento demostraba que los resultados no se veían significativamente afectados por estos artefactos [Itt04].

Para la creación de la base de datos CRCNS se empleó el protocolo siguiente: Al inicio de cada tarea el sujeto debía pulsar una tecla y a continuación se le presentaba un punto de fijación parpadeando en el centro de la pantalla, tras desaparecer dicho punto comenzaba la reproducción de un vídeo. La grabación de los movimientos de los ojos mediante un seguidor ocular comenzaba en el instante en el que el sujeto pulsaba una tecla.

La visualización de los experimentos se realizó en un monitor CRT de 22 pulgadas situado aproximadamente a 80 cm. de distancia del sujeto. El

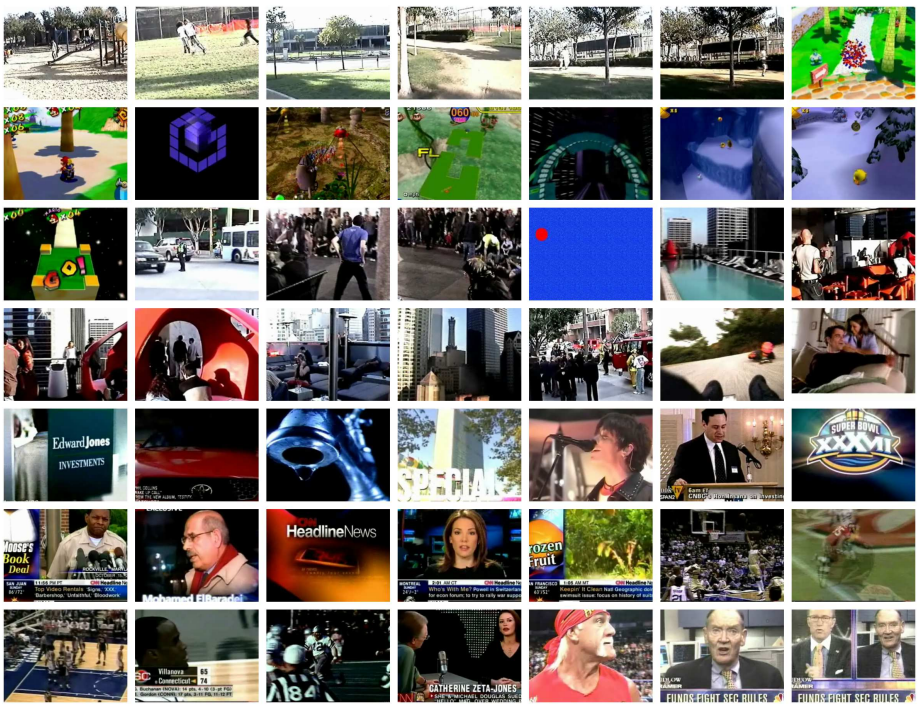


Figura 3.1: Fotogramas iniciales de los vídeos de la BD CRCNS.

ángulo subtendido a esa distancia es de 28 grados en horizontal por 21 grados en vertical, con una resolución de 23 píxeles por grado. En la figura 3.1 se incluyen los fotogramas iniciales de los vídeos del experimento original que componen esta BD. Los vídeos fueron almacenados a 30 fps.

Tal como se describe en las primeras publicaciones de Itti et al. [IKN98, IK00, IK01], a los sujetos se les indicaba al inicio de las tareas que procurasen seguir a los actores principales de cada escena. Con estos vídeos se realizaron inicialmente dos experimentos:

- `Experimento ORG (original)`: Consta de 50 vídeos, vistos por ocho sujetos, tres mujeres y cinco hombres con edades comprendidas entre los 23 y los 32 años. Incluyen de 4 a 50 fijaciones útiles por sujeto con un total de 235 archivos de fijaciones oculares.

- `Experimento MTV`: Consta de 50 vídeos obtenidos cortando los vídeos originales en trozos de 1 a 3 segundos y componiéndolos de nuevo. Fueron vistos por ocho sujetos, entre 4 y 7 sujetos por cada vídeo, tres mujeres y cinco hombres con edades comprendidas entre los 23 y los 32 años. Incluyen entre 6 y 50 fijaciones oculares útiles con un total de 285 archivos de fijaciones oculares.

Figura 3.2: Fotogramas en los que se aprecia el cambio del fondo y el desplazamiento del objeto perseguido; los jugadores y el balón de fútbol.

Figura 3.3: Fotogramas elegidos manualmente, la cámara persigue a los jugadores y al balón durante un partido de fútbol americano.

En la figura 3.1 no se incluyen muestras del experimento MTV, ya que son fragmentos concatenados de los vídeos del experimento original. En la figura 3.2 se muestran varios fotogramas significativos seleccionados manualmente sobre una de las secuencias de esta BD.

En los fotogramas mostrados se aprecia cómo se producen cambios del fondo durante la persecución de un objeto, en este caso el balón de fútbol. Esta BD contiene vídeos en los que la cámara se mantiene estática y otros en los que tanto la cámara como el contenido de la escena se mueven. El hecho de que ambos sistemas se muevan simultáneamente dificulta el proceso de extracción del fondo de las escenas. La figura 3.3 muestra otro evento deportivo en el que la cámara persigue a varios jugadores por el campo durante un partido de fútbol americano. De nuevo el fondo cambia completamente durante el movimiento de la cámara.

Otra de las categorías de las que se han incluido varios vídeos en esta BD es la de videojuegos, o sea vídeos capturados directamente de un monitor durante una partida de un jugador. El ejemplo de la figura 3.4 es de uno de los juegos de la consola GameCube de Nintendo. En estos juegos el foco está situado en el avatar del jugador y van apareciendo diferentes objetos al avanzar por las diferentes pantallas.

Resumiendo, los vídeos del experimento original de la BD CRCNS se pueden incluir en las siguientes categorías: escenas urbanas en parques o avenidas con personas y coches en movimiento, vídeos tomados de videojuegos de ordenador, programas de noticias, anuncios televisivos, escenas extraídas de eventos deportivos variados y alguna entrevista televisiva.

Figura 3.4: Fotogramas elegidos manualmente, la cámara se sitúa tras el avatar del jugador que avanza por las diferentes pantallas.

Figura 3.5: Varios fotogramas elegidos manualmente de una secuencia dinámica modificada digitalmente.

3.1.2. Base de datos de vídeos DIEM

La base de datos CARPE (Computational and Algorithmic Representation and Processing of Eye-movements) forma parte del proyecto DIEM (Dynamic Images and Eye Movements) dirigido por el Prof. John M. Henderson de la Universidad del Sur de Carolina [MSHH11]. El objetivo de este proyecto es investigar cómo ven los humanos. Para ello se han adquirido un conjunto de 85 vídeos de prueba que han sido vistos por 250 sujetos. Todos estos datos están disponibles vía web bajo licencia de uso no comercial Creative Commons CC-NC-SA 3.0 [Pro11].

Durante la realización de este experimento se ha incluido también el audio que se encuentra almacenado junto a los vídeos de la BD, salvo alguno de los ejemplos en los que el mismo vídeo se ha presentado a diferentes sujetos, con y sin audio. Puesto que los sistemas de control de la atención se ven afectados por los procesos cognitivos relacionados con el habla [HF04], debido a la presencia de la información auditiva, pueden surgir efectos que afecten al rendimiento de los modelos empleados.

A continuación, en la figura 3.5, se muestran varios fotogramas significativos de alguno de los vídeos de esta BD. En ellos se puede ver una secuencia modificada mediante ordenador añadiendo efectos de fuegos artificiales con pinturas de colores vistosos que tienden a atraer la atención de los sujetos.

Tal como se muestra en la figura 3.6, en esta base de datos también abundan vídeos de entrevistas en la calle o programas televisivos en los que tan solo se muestran uno o varios sujetos durante una conversación. En estos casos la presencia de la información de audio produce una evidente alteración

Figura 3.6: Varios fotogramas elegidos manualmente de una grabación en la que se realizan entrevistas en la calle.

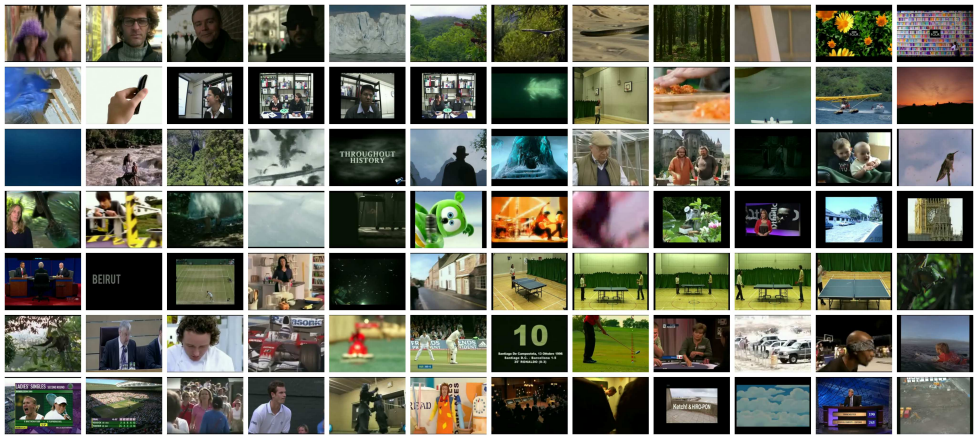


Figura 3.7: Fotogramas iniciales de los vídeos de la BD DIEM.

en la dirección del proceso de atención, dirigiéndola hacia la persona que está hablando. En la figura 3.7 se incluyen los fotogramas iniciales de los vídeos de esta base de datos. Los vídeos fueron almacenados a 30 fps.

En resumen, los vídeos que contiene la BD DIEM se pueden incluir en las siguientes categorías: entrevistas, documentales acerca de animales y plantas, programas de noticias, anuncios televisivos, escenas extraídas de eventos deportivos variados, vídeos tomados de videojuegos de ordenador, concursos televisivos e incluso algún capítulo de dibujos animados. En general se mezclan escenas exteriores e interiores, así como urbanas y naturales.

Figura 3.8: Fotogramas iniciales de los vídeos de cada una de las categorías de la BD AE-UCFS.

Figura 3.9: Varios fotogramas elegidos manualmente de la categoría de saltos.

3.1.3. Base de datos de vídeos AE-UCFS

La base de datos de vídeos de deportes de la Universidad de Florida Central incluye vídeos que contienen acciones de deportes variados [RAS08, SZ14]. Han sido recopilados de múltiples sitios web, como la galería de vídeos de la BBC y la web GettyImages.

En la figura 3.8 se incluyen los fotogramas iniciales representativos de cada una de las categorías que componen esta BD. Incluye un total 150 vídeos tomados desde diferentes ángulos, con una duración promedio de 6.4 seg. y con una resolución de 720x480. El objetivo de la creación de esta BD era analizar el comportamiento de algoritmos de detección de acciones en entornos no restringidos [SZ14], aunque ha sido empleada también para detección de saliencia [MS12]. Los vídeos fueron almacenados a 10 fps.

Contiene vídeos de nueve categorías claramente diferenciadas; salto de trampolín (14), golf-swinging (18), fútbol (18), levantamiento de pesas (6), paseo a caballo (12), personas corriendo (13), skateboarding (12), caballo con arcos (20), barras paralelas y asimétricas (13) y personas caminando (22).

A continuación, en la figura 3.9, se muestran varios fotogramas significativos de alguno de los vídeos de esta BD. En el vídeo se muestra un salto de trampolín en el que la cámara persigue al saltador. Los vídeos de esta base de datos son en general muy cortos, el actor principal es evidente y en muchos de ellos el fondo se mantiene estático.

Figura 3.10: Fotogramas iniciales de todos los vídeos de la BD ASCMN.

Figura 3.11: Varios fotogramas elegidos manualmente de la categoría de aglomeraciones de personas.

3.1.4. Base de datos de vídeos ASCMN

Esta base de datos referenciada como ASCMN por las iniciales de las categorías que incluye (anómalos, videovigilancia, multitudes, movimiento y ruido), fue publicada por Riche el al. en 2012 [RMC+12]. Debido a esta misma publicación también se referencia la BD como ACCV. Para su construcción se han adquirido un conjunto de 24 vídeos de prueba que han sido vistos por 10 sujetos. Todos los vídeos han sido almacenados a 15 fps.

En la figura 3.10 se incluyen los fotogramas iniciales de los vídeos que componen esta BD. Esta base de datos contiene vídeos con objetos con movimiento sorprendente, movimiento normal, agrupaciones de personas en movimiento, vídeos con la cámara en movimiento y largos periodos de tiempo precedentes a la aparición de objetos salientes.

A continuación, en la figura 3.11, se muestran varios fotogramas significativos de uno de los vídeos de esta BD. En el vídeo se muestra una conglomeración de personas, y ciertas regiones en las que los movimientos de algunas de ellas son totalmente diferentes.

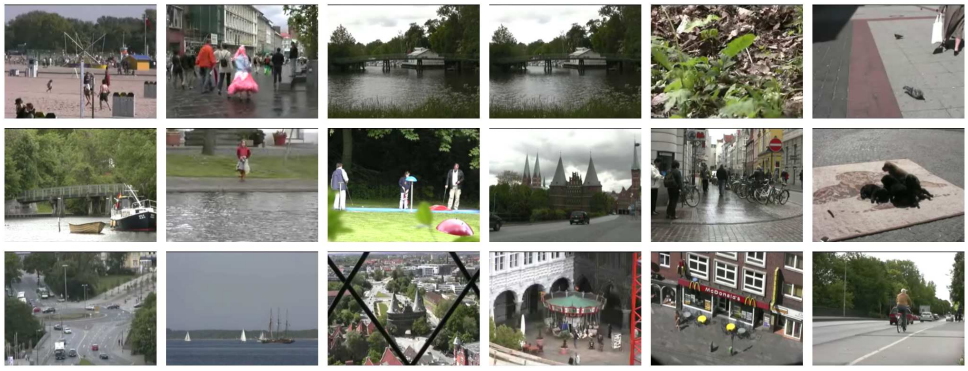


Figura 3.12: Fotogramas iniciales de todos los vídeos de la BD GC.

3.1.5. Base de datos de vídeos GC

La base de datos GC también referenciada como INB por las iniciales del instituto en el que ha sido creada, *Institute for Neuro and Bioinformatics*, contiene múltiples escenas adquiridas en las proximidades de la ciudad alemana de Lübeck, por lo que también se referencia como BD de Lübeck. Fue creada por Dorr et al. para la realización de un experimento en el que se evalúa la semejanza de los patrones oculares de los sujetos durante la observación libre de escenas naturales [DMGB10]. Consta de 18 vídeos de 20 seg. vistos por 54 sujetos y se ha empleado un seguidor ocular de resolución temporal 250Hz (4ms/dato).

En la figura 3.12 se incluyen los fotogramas iniciales de los vídeos que componen esta BD. Ocho vídeos muestran personas en áreas peatonales, en la playa y jugando al mini golf; tres muestran animales y coches cruzando; otros tres son escenas casi estáticas con objetos distantes; uno es un vídeo tomado desde una torre de una iglesia. Los vídeos fueron almacenados a 30 fps. con una resolución de 1280x720 píxeles.

Figura 3.13: Fotogramas iniciales de los vídeos de la BD Hollywood2.

3.1.6. Base de datos de vídeos Hollywood2

Esta base de datos de los mismos autores de la AE-UCFS es una de las mayores bases de datos de vídeos, con información de fijaciones oculares, disponibles hasta el momento. Contiene escenas de doce categorías diferentes; contestando el teléfono, conduciendo, comiendo, peleando, saliendo del coche, agitando las manos, abrazando, besando, corriendo, sentándose, levantándose y manteniéndose en pie. Todas estas acciones se han tomado de 69 películas de Hollywood. En total esta BD contiene 20 horas de vídeo que han sido divididas en dos conjuntos de escenas diferentes, uno de entrenamiento con 823 vídeos y otro de test con 884 vídeos. Los vídeos de la base de datos Hollywood2-Train fueron almacenados a 23, 25 y 29 fps, y los vídeos de la Hollywood-Test a 25 y 29 fps.

En la figura 3.13 se muestran una selección aleatoria con 80 de los 1.707 vídeos que componen esta BD. El experimento fue realizado por 16 sujetos (9 hombres y 7 mujeres) e incluye un total de 669.187 fijaciones y 92 horas de visualización por parte de los sujetos.

Para realizar los experimentos se ha empleado un seguidor ocular de alta resolución temporal (500Hz) con un error de calibración promedio de menos de 0.45º [MS12, MS15]. Los vídeos han sido presentados en un monitor de 1280x1024 situado a 60 cm. del observador. Cuatro de los sujetos han visto la tarea sin objetivo (observación libre), mientras que al resto se le planteó la tarea de reconocer la acción presente en el vídeo [MS15].

Nombre DB	Seguidor Ocular	Sujetos	Vídeos	Fotog.	Condición
CITIUS	SMI RED	22	72	14.247	Free-V
CRCNS-ORG	ISCAN RK-464	8	50	46.489	Task-G
CRCNS-MTV	ISCAN RK-464	8	50	32.749	Task-G
DIEM	SR-EyeLink1000	250	84	240.452	Free-V
AE-UCFS	SMI-X HiSpeed	16	150	8.262	Free-V
ASCMN	FaceLab 5	10	24	10.180	Free-V
GC	SR-EyeLink II	54	18	10.049	Free-V
HollyTrain	SMI-X HiSpeed	16	823	229.823	Ambas
HollyTest	SMI-X HiSpeed	16	884	257.733	Ambas

Tabla 3.1: Características de las BD, y tipo de tarea; observación libre (Free-V), guiada por un objetivo (Task-G) o ambas.

3.2. Nuestra propuesta. BD CITIUS

En la tabla 3.1 se muestra un resumen de las características de las bases revisadas y las de nuestra propuesta, la BD CITIUS. Al revisar las bases de datos descritas en la secciones 3.1.1-3.1.5, se han observado varias carencias que han motivado la creación de una nueva BD. En el caso de la BD CRCNS, sección 3.1.1, el número de sujetos experimentales es insuficiente para poder crear un mapa de atención humano dinámico completo. Existen largos periodos de tiempo en los cuales el número de observadores simultáneos que se encuentran realizando alguna fijación es muy bajo o nulo. Con respecto a la BD CARPE, 3.1.2, aunque el número de sujetos es mucho mayor y el número de fijaciones es en principio suficiente, presenta gran cantidad de vídeos urbanos. Por otra parte añade una característica que no contemplan los modelos, la intervención del audio en el proceso de atención. Además esta BD no contiene suficientes vídeos sintéticos con los que poder validar los modelos en situaciones controladas, en los que el foco de atención sea conocido a priori. La BD AE-UCFS, sección 3.1.3, está especialmente diseñada para la identificación de gestos o acciones, y contiene vídeos muy cortos en los que siempre hay personas realizando alguna actividad. La BD Holly2, sección 3.1.6 también carece de ejemplos sintéticos. La BD ASCMN, sección 3.1.4, se centra en el estudio de comportamiento de masas y contiene pocos vídeos, al igual que la BD INB, sección 3.1.5, que también se centra en escenas urbanas.

Los vídeos de la base de datos creada en el CITIUS han sido recopilados de múltiples fuentes de modo manual y clasificados en cuatro grandes categorías con diferente nivel de dificultad:

- `Reales estáticos (RCE)`: Vídeos de escenas de la naturaleza o urbanas, en los que la cámara se mantiene fija en una posición y el contenido

Figura 3.14: Varios fotogramas elegidos manualmente, el fondo permanece estático mientras se mueven múltiples objetos en la escena.

Figura 3.15: Varios fotogramas elegidos manualmente, se aprecia el movimiento del fondo y a su vez el desplazamiento del objeto perseguido, en este caso un águila.

de la escena varía con el tiempo.

- `Reales dinámicos (RCD)`: Vídeos de escenas de la naturaleza o urbanas, en los que la cámara se mueve de algún modo y además el contenido de la escena varía con el tiempo.
- `Sintéticos estáticos (SCE)`:Vídeos sintéticos generados mediante simuladores, en los que la cámara se mantiene fija en una posición y el contenido de la escena varía con el tiempo.
- `Sintéticos dinámicos (SCD)`: Vídeos sintéticos generados mediante simuladores, en los que la cámara se mueve de algún modo y además el contenido de la escena varía con el tiempo.

A continuación se muestran algunas secuencias de fotogramas significativos de cada una de las categorías descritas para esta base de datos. En la figura 3.14 se puede apreciar un ejemplo de la primera categoría (RCE), en el que una cámara fija percibe los cambios que se producen en una avenida en la que personas y coches se mueven libremente por la escena. El fondo prácticamente no varía en todo el vídeo, salvo cambios de iluminación debidos a modificaciones en las condiciones externas.

En la figura 3.15 podemos ver varios fotogramas para una secuencia dinámica de la segunda categoría (RCD). En este vídeo la cámara persi-

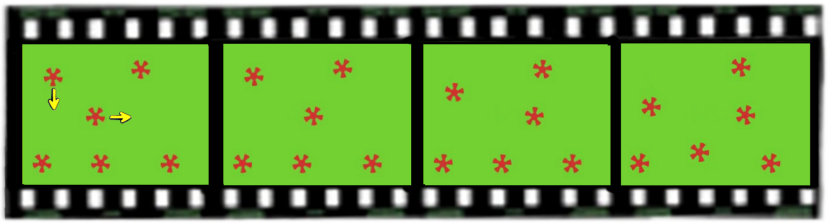


Figura 3.16: Varios fotogramas elegidos manualmente en los que varios objetos sintéticos se desplazan por la pantalla.

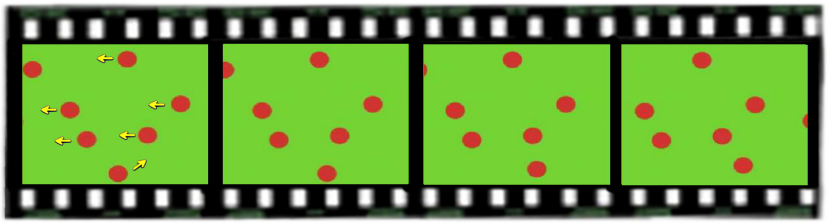


Figura 3.17: Varios fotogramas elegidos manualmente en los que varios objetos sintéticos se mueven por la pantalla y uno de ellos se desplaza en sentido contrario.

gue a un águila que sobrevuela un bosque, produciéndose una composición de movimientos, el de la cámara sobre el bosque y el del águila que no siempre está centrado en el fotograma.

En los vídeos pertenecientes a esta categoría, tanto el fondo como los objetos presentes en la escena pueden presentar movimiento.

En la figura 3.16 podemos ver varios fotogramas para una secuencia de la tercera categoría (SCE), varias estrellas que están girando sobre si mismas se desplazan por la pantalla en diferentes instantes[1].

En la figura 3.17 se muestra un ejemplo de la última categoría (SCD), en este vídeo todos los objetos se mueven hacia la izquierda menos uno de ellos que se desplaza en sentido contrario.

En general en estas dos últimas categorías, SCE y SCD, los vídeos presentan grupos de objetos con un comportamiento uniforme y, uno o varios de ellos, presenta alguna característica diferencial, ya sea el color, la forma o el movimiento que realizan.

Partiendo de la selección de vídeos, mostrada en la figura 3.18, se ha diseñado una tarea de observación libre con la finalidad de estudiar cómo se comportan los modelos ante diferentes efectos *bottom-up*. Se ha evitado dar ninguna información a los sujetos acerca del objetivo del estudio. Los 22 sujetos elegidos para realizar la observación libre de los vídeos de la BD

[1]Las flechas amarillas de las figuras son informativas, no existen en los vídeos reales.

Figura 3.18: Fotogramas de muestra de los vídeos naturales (arriba) y de los sintéticos (abajo) para la BD CITIUS.

han sido voluntarios de secundaria, alumnos y profesores de universidad con edades comprendidas entre los 11 y los 43 años ($\mu = 29.7$, $\sigma = 6.8$).

Para grabar las posiciones de ambos ojos durante la realización del experimento, se ha utilizado el seguidos ocular SMI RED EyeTracker (50Hz), mostrado en la figura 3.19, y el software de procesamiento BeGaze^{TM}de la empresa Sensomotoric Instruments.

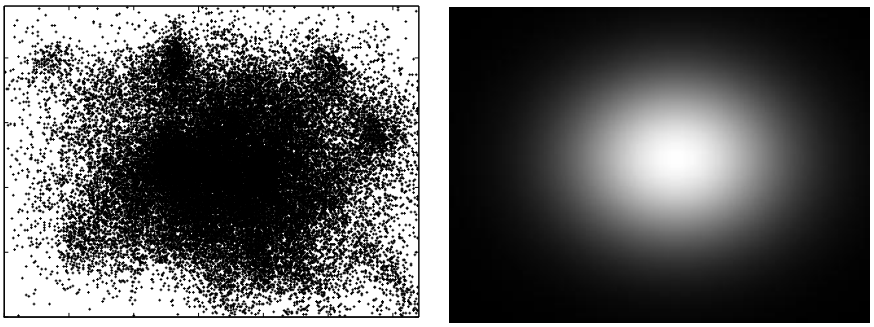
Todos los vídeos han sido reescalados a una resolución de 1280x1024 píxeles durante la presentación a los sujetos. Tanto el código como los vídeos y las fijaciones de los sujetos están disponibles a través de la web del CITIUS[2].

En la figura 3.20 se muestran las fijaciones de todos los sujetos superpuestas para todos los vídeos de esta base de datos. Se aprecia claramente el efecto del sesgo central descrito en la sección 3.4.1 común a todas las BD revisadas.

Para comprobar en qué medida este efecto por sí sólo puede explicar las fijaciones humanas, se ha construido un histograma con las posiciones XY de las fijaciones y se ha realizado un ajuste del mismo a una función

[2]La BD CITIUS con los vídeos y fijaciones pueden ser descargadas desde la URL: https://wiki.citius.usc.es/inv:downloadable_results

Figura 3.19: Esquema del seguidor ocular SMI EyeTracker. El sujeto se sitúa frente al monitor a una distancia de 60cm.



(a) (b)

Figura 3.20: (a) Superposición de las fijaciones de todos los sujetos para todos los vídeos de la BD CITIUS (40.558 fijaciones) y (b) modelo obtenido ajustando los histogramas de las posiciones XY a una función Gaussiana.

Gaussiana. El mapa resultante se ha normalizado y tomado como modelo válido para todos los fotogramas de todos los vídeos de esta BD, figura 3.20b. Con este modelo también se han realizado todos los análisis de modo análogo al empleado con los demás modelos probados.

Se han medido los tiempos promedio de duración de las fijaciones para todos los sujetos de esta BD, obteniendo un valor global de $309 \pm 237ms$. Este valor y la desviación son semejantes a los $330ms$ medidos en trabajos previos [Dor10] y es comparable a los valores obtenidos para las BD de prueba descritas en las secciones 3.1.1-3.1.5.

En la figura 3.21 se muestran las distribuciones de tiempos de fijación para las cuatro categorías de la BD CITIUS. Para los vídeos sintéticos con cámara en movimiento (SCD) los tiempos de fijación son claramente menores ($214 \pm 106ms$), este hecho se puede justificar en base a que al estar moviéndose

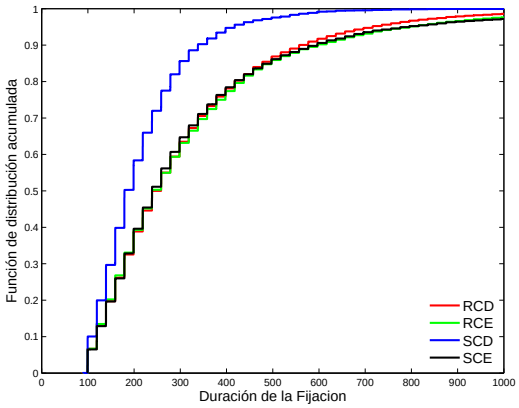


Figura 3.21: Función de distribución acumulativa empírica para la duración de las fijaciones con las cuatro categorías de la BD CITIUS.

tanto los objetos como el fondo, los sujetos tienen la necesidad de mover el foco de atención rápidamente de unas posiciones a otras. Por otra parte, las demás categorías presentan unos tiempos promedio semejantes $307 \pm 211ms$ para RCD, $321 \pm 268ms$ para SCE y $320 \pm 248ms$ para RCE .

3.2.1. Clasificación de los movimientos oculares

Los seguidores oculares a los que se refiere este trabajo, listados en la tabla 3.1, generan una serie de archivos para cada tarea realizada por cada uno de los sujetos. En dichos archivos se almacena, etiquetada temporalmente, toda la información relevante relacionada con los movimientos oculares del sujeto durante la presentación de los diferentes estímulos. Los tipos de datos que podemos encontrar en estos archivos son los siguientes:

- `Fijaciones.` Son las posiciones en las que el globo ocular se mantiene durante cierto tiempo realizando movimientos a velocidades bajas, inferiores a los $100°/s$. Existen múltiples estrategias para filtrar estos movimientos, basadas tanto en criterios espaciales como temporales, siendo las más sencillas las que emplean el perfil de velocidad para filtrar estos movimientos [SG00].

- `Sacadas.` Son movimientos bruscos cuya duración oscila entre los 20 y 80 ms. [Kow11]. En ellos la mirada cambia de un punto a otro a unas velocidades del orden de $400°/s$ con valores de aceleraciones que pueden alcanzar picos de $20.000°/s^2$ para pequeñas variaciones angulares [KAA03, AMK89].

- `Parpadeos.` Son movimientos de los párpados, reflejos o controlados, en los que los seguidores oculares no invasivos no son capaces de identificar la posición real de la mirada del sujeto, ya que los párpados impiden la visión de los sistemas de adquisición. Esto no sucede con otros sistemas más invasivos como los COIL (pequeñas espiras que inducen un campo, en los que la posición del ojo se obtiene por la medida de un campo magnético) o los electro oculogramas, en los que unos electrodos permiten identificar la situación del globo ocular aún cuando los párpados están cerrados.

- `Movimientos de persecución.` Son movimientos en los que el globo ocular realiza rotaciones suaves. A diferencia de los movimientos voluntarios, los movimientos de persecución se realizan de modo continuo por lo que no dan lugar a fijaciones.

Los tiempos promedio de duración de las fijaciones de los sujetos de las BD de prueba son: $489 \pm 410ms$ con CRCNS, $416 \pm 496ms$ con DIEM, $398 \pm 433ms$ con Holly2Train, $397 \pm 442ms$ con Holly2Test, $458 \pm 358ms$ con CRCNS-ORG, $536 \pm 451ms$ con CRCNS-MTV, $380 \pm 394ms$ con AE-UCFS, siendo estos valores semejantes a los mencionados en trabajos previos [Und98, Hen03].

3.2.2. Creación del mapa de atención visual humano

Conforme a la *hipótesis de saliencia visual*, los movimientos oculares están íntimamente relacionados con la atención visual. Esta suposición, y el hecho de que la atención está guiada por la saliencia, permite que podamos emplear las posiciones de las fijaciones oculares como referencia de la distribución espacial del mapa de atención de un sujeto.

Para cada una de las diferentes bases de datos empleadas, tomando como punto de partida las posiciones de las fijaciones oculares de todos los sujetos, usando el valor de la contribución de cada fijación obtenido mediante la ecuación 3.1 y el algoritmo 3.1 que se muestra en pseudocódigo, se obtiene el mapa de atención asociado a cada fotograma del vídeo. Dicho mapa es el resultado de la convolución de un filtro gaussiano sobre las posiciones de fijación de todos los observadores.

$$S(x, y, t) = (\alpha \cdot t + (1 - \alpha)) \cdot exp\left(-\frac{(x - x_i)^2 + (y - y_i)^2}{2\ \sigma^2}\right) \tag{3.1}$$

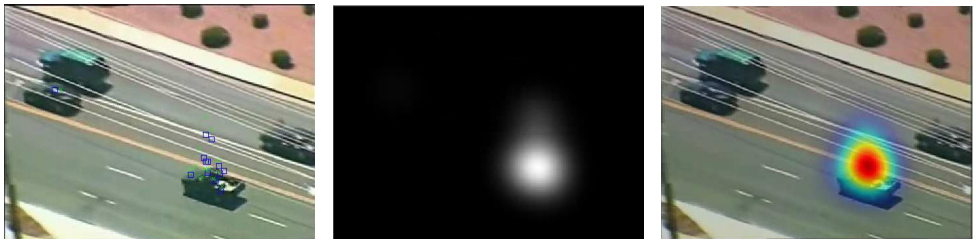
Siendo x, y las posiciones promedio asociadas a la posición en la que se produce la fijación, t la duración de la fijación, x_i, y_i las coordenadas de cada píxel de la imagen, σ el parámetro de ancho que aproxima el tamaño de la

Algoritmo 3.1: Mapa de atención a partir de las fijaciones humanas

```
/* Entrada: N fijaciones Fi(x,y,t).              */
/*  x e y son coordenadas espaciales.            */
/*  t es la duración de la fijación.             */
/* Salida: ImSAL(w,h), mapa de atención visual. */

for Fi = 1:N do
  (x,y) = Coord(Fi)
  t = Duración(Fi)
  ImSAL = ImSAL + S(x,y,t);
end for

Normalize(ImSAL,0,255);
```



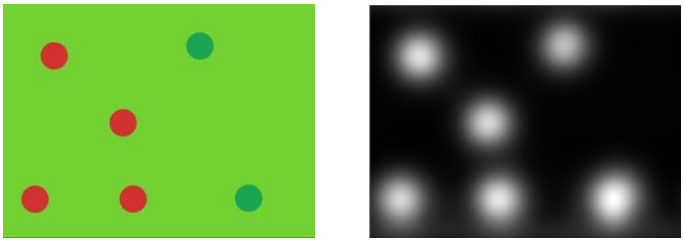
(a) (b) (c)

Figura 3.22: (a) Fotograma de una secuencia dinámica de la BD CITIUS, (b) su correspondiente mapa de atención y (c) el mapa en falsocolor superpuesto a la imagen original.

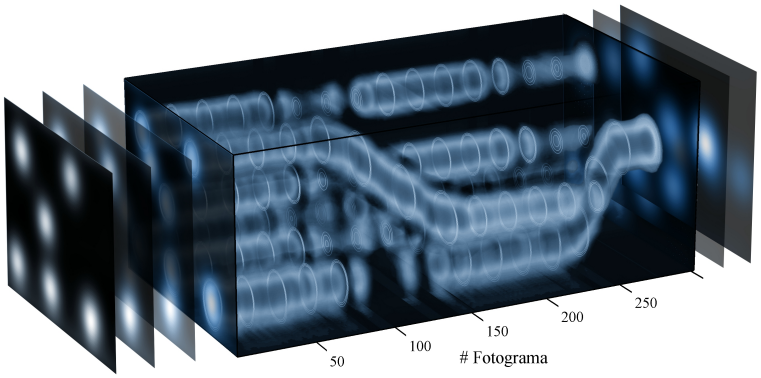
fóvea y α un parámetro que permite modular la contribución de la duración de las fijaciones. En la práctica cada localización en la que hay una fijación dará lugar a una función de distribución gaussiana. Con $\alpha = 0$ la amplitud de la gausiana no se verá afectada por la duración. Mientras que para $\alpha = 1$ la amplitud de la gaussiana será proporcional a la duración de la fijación.

En la figura 3.22a se muestra un fotograma de un vídeo en el que una cámara en movimiento persigue a un vehículo a gran velocidad. Se han marcado con cuadrados las coordenadas de posición x, y de las fijaciones realizadas por varios sujetos. En la figura 3.22b se muestra el correspondiente mapa de atención obtenido a partir de dichas fijaciones y por último en 3.22c se representa el mapa de atención en falsocolor superpuesto a la imagen original para resaltar las regiones con mayor valor de saliencia del mapa de atención.

Para reflejar en una sola imagen el resultado de la concatenación de todos



(a) Fotograma (Vídeo) (b) Fotograma (Mapa)



(c) Composición 3D

Figura 3.23: (a) Fotograma de una secuencia de ejemplo, (b) su correspondiente mapa de saliencia, (c) representación de todos los fotogramas de la secuencia en una representación espacio-temporal única.

los mapas de atención asociados a todos los fotogramas de un vídeo, se puede emplear una representación como la de la figura 3.23. Para su construcción se han concatenando imágenes de niveles de gris como la mostrada en la figura 3.23b y se ha añadido transparencia a las regiones oscuras, de este modo las regiones más salientes se pueden ver en una representación espaciotemporal tridimensional única como la de la figura 3.23c. Esta misma representación se ha utilizado en la sección de resultados para mostrar los mapas de atención creados a partir de las fijaciones de los sujetos y también para mostrar los mapas de saliencia de los modelos de atención. Aunque en esta figura los fotogramas iniciales y finales aparecen separados del resto a modo explicativo, en este trabajo se presentarán integrados en el volumen con los demás fotogramas de la secuencia.

Determinar cuál es el verdadero mapa de atención asociado a cada fotograma de un vídeo es complejo, y por ello se ha tomado como verdad absoluta (*ground truth*) el mapa de atención generado a partir de las fijaciones reales de los sujetos durante la realización de las tareas de atención visual en ese

mismo instante. Esta comparación está sujeta a ciertos sesgos que no se presentan en la realidad cotidiana. La limitación práctica de la realización de las tareas frente a un monitor no simula con precisión la realidad que un sujeto percibe a través de su sistema de visión en una escena real tridimensional, pero es una buena aproximación [ZTW09].

Es también importante resaltar que el mapa de atención obtenido de este modo para cada fotograma no devuelve información acerca del orden en el que los sujetos realizan las fijaciones sino acerca de la distribución espacial de las regiones a las que un grupo de sujetos presta mayor atención.

3.3. Métricas de evaluación

Cuando surgen nuevas metodologías, o se proponen mejoras en las existentes, uno de los hitos con los que se enfrenta el evaluador es el de valorar cuán buena y novedosa es una aportación. Uno de los problemas que surge al realizar esta comparación es el hecho de que diferentes modelos suelen estar diseñados para la resolución de tareas diferentes, por lo cual la comparación no resulta obvia. En algunos casos se puede usar como criterio la simplicidad del modelo, o incluso la capacidad de explicar el mayor número de resultados psicofísicos. También existen métricas, aunque están más orientadas hacia la comparación de modelos *top-down*, que valoran el número promedio de fijaciones o el tiempo necesario hasta localizar el objeto buscado.

Para una comparación cuantitativa de los resultados obtenidos por los modelos de saliencia S_M con respecto a las fijaciones humanas H_M se pueden encontrar varias aproximaciones: Se puede pensar en S_M y H_M como distribuciones de probabilidad y emplear las técnicas estadísticas habituales para realizar esta comparación, tales como la distancia KL [Kul59]. También se puede pensar en S_M y H_M como distribuciones aleatorias y calcular su relación mediante el coeficiente de correlación; o se pueden tomar los diferentes S_M de cada cada algoritmo como clasificadores binarios y emplear técnicas de la teoría de detección de señales como la ROC para valorar el rendimiento del clasificador [BI13]. También es posible comparar directamente las trayectorias estimadas por el modelo con las realizadas por los sujetos, por ejemplo asociando cada posición a una determinada región de la imagen y calculando la distancia entre regiones [PS00, CBN04]. Pero también se pueden comparar directamente las posiciones de fijación sugeridas por los modelos con las posiciones de objetos previamente segmentados manualmente -saliencia extendida- [AL06]. De este modo se puede analizar la cantidad de fijaciones necesarias para cubrir un cierto porcentaje de los objetos marcados .

En general el objetivo de los modelos basados en mapas de saliencia es

localizar una función que minimice el error de la predicción de las fijaciones oculares asumiendo la salida del modelo como un mapa de densidad de probabilidad de fijación y compararlo con el mapa obtenido a partir de las fijaciones de los sujetos. Además hay que tener en cuenta que, durante la observación de vídeos, la probabilidad de que dos observadores fijen su atención en el mismo instante exactamente en un mismo punto es baja. Pero tal como muestran Berg et al. aún siendo baja, la coherencia entre las fijaciones de diferentes observadores es mayor para los humanos que para los primates subhumanos [BBM+09].

A continuación se describirán varias de las métricas de uso más extendido para realizar la evaluación cuantitativa de la calidad de un algoritmo frente a los ya existentes en la bibliografía.

3.3.1. Coeficiente de correlación

El coeficiente de correlación es un índice que mide la relación lineal entre dos variables. Presenta varias ventajas interesantes como son la capacidad de comparar dos variables mediante un solo valor y el hecho de que dicho valor esté comprendido entre -1 y 1.

Para obtener el coeficiente de correlación, para la compararación de mapas de saliencia, se emplea el mapa de atención de los sujetos $H_M(x)$ y el mapa de saliencia del modelo computacional $S_M(x)$. A partir de estos dos mapas mediante la ecuación 3.2 se define el coeficiente de correlación ρ

$$\rho = \frac{\sum_x [(H_M(x) - \mu_h) \cdot (S_M(x) - \mu_s)]}{\sqrt{\sum_x (H_M(x) - \mu_h)^2 \cdot \sum_x (S_M(x) - \mu_s)^2}} \tag{3.2}$$

siendo μ_h y μ_s los valores medios de los dos mapas de saliencia $H_M(x)$ y $S_M(x)$ respectivamente. Cuando ρ se aproxima a ± 1 indica la existencia de una relación lineal casi perfecta entre las dos variables.

3.3.2. Distancia relativa Fijación-Azar

La distancia fijación-azar compara los valores muestreados del mapa de saliencia del modelo $S_M(x)$ en las posiciones de las fijaciones oculares de los sujetos con valores obtenidos en posiciones aleatorias.

La distancia fijación-azar s_{ftc} emplea los datos crudos de las posiciones XY de las fijaciones oculares de los sujetos, por lo que no es necesaria la creación del mapa de atención de los sujetos $H_M(x)$.

Para obtener esta medida se calcula el valor promedio de las muestras del mapa computacional de saliencia en las posiciones de fijación de los sujetos

$\bar{s}_{fix}$, a continuación se calcula el valor promedio de las muestras del mismo mapa computacional de saliencia en posiciones XY generadas al azar $\bar{s}_r$ y empleando la ecuación 3.3 se obtiene la distancia fijación azar.

$$s_{ftc} = \frac{\bar{s}_{fix} - \bar{s}_r}{\bar{s}_r} \tag{3.3}$$

Los valores de la saliencia obtenidos del mapa $S_M(x)$ serán más altos que los valores obtenidos en posiciones aleatorias cuando exista una buena correspondencia entre los datos humanos y los datos del modelo.

3.3.3. Saliencia Normalizada

La saliencia normalizada de la ruta de exploración, más conocida como NSS, es una técnica muy empleada en tareas de búsqueda visual para determinar la diferencia entre la trayectoria seguida por las fijaciones oculares del sujeto y la predicha por el modelo [PLN02, PIIK05].

El cálculo de la métrica NSS descrito por Peters et al. [PIIK05] emplea un rango dinámico variable basado en la variabilidad de los mapas de saliencia. Para ello en primer lugar se normalizan los mapas de saliencia $S_M(x)$ de tal modo que tengan media cero y desviación estándar unidad, eq. 3.4. A continuación se extraen los valores de la saliencia normalizada para las localizaciones en las que se ha producido una fijación a lo largo de toda la trayectoria de exploración. La media de esos valores es lo que se denomina la saliencia normalizada de la ruta de exploración.

$$S^*_M(x) = \frac{S_M(x) - \overline{S_M(x)}}{Std(S_M(x))} \tag{3.4}$$

Valores de NSS mayores que cero indicarán una gran correspondencia entre las posiciones de las fijaciones y las posiciones predichas por el modelo evaluado. Por el contrario un valor cero indicará que no hay correspondencia. Valores menores que cero indicarán una correspondencia opuesta entre las posiciones de las fijaciones y las predichas por el modelo.

Esta misma métrica ha sido aplicada sobre fotogramas de vídeos cortos por Marat et al. [MHPG+09], para ello toma los valores de la saliencia sobre el mapa $S^*_M(x)$ en las posiciones de fijación de diferentes sujetos para cada instante y promediando dichos valores obtiene la NSS global de cada fotograma. Dorr et al. hacen una extensión al dominio espacio-temporal de tal modo que pequeños desplazamientos temporales de las fijaciones resulten en una contribución positiva [DMGB10]. Para ello se aplica un filtro Gaussiano espacio-temporal a las posiciones (x,y,t) de las fijaciones y a continuación aplica la ecuación 3.4 de modo semejante a Peters et al. [PIIK05], y Marat et al. [MHPG+09].

3.3.4. Distancia de movimiento de tierras

La distancia de movimiento de tierras -*earth moving distance*-, conocida como EMD, permite determinar la distancia entre dos distribuciones de probabilidad. Tal como indica su nombre, suponiendo las distribuciones como pequeños montículos de tierra, esta distancia representa el coste mínimo que conlleva convertir una distribución en la otra transportando para ello la tierra de un montículo al otro. En el contexto de este trabajo, sea S_M el mapa de saliencia del modelo, entonces obtenemos R como la función de densidad de probabilidad para $S_M(x)$ en posiciones aleatorias y H la función de densidad de probabilidad para las posiciones de fijación. La principal aportación de esta métrica está en el hecho de tener en cuenta, además de la ordenación de los datos, la diferencia espacial de los mismos. Formalmente basándose en la descripción de Pele et al. [PW09], (implementación usada en este trabajo):

$$EMD(R,H) = (\min_{\{f_{ij}\}} \sum_{i,j} f_{ij} d_{ij}) / \sum_{i,j} f_{ij} \quad , f_{ij} \geq 0 \tag{3.5}$$

siendo $\sum_{i,j} f_{ij} = \min(\sum_i R_i, \sum_j H_j)$, cada f_{ij} representa la cantidad transportada desde el i-ésimo origen hasta el j-ésimo destino. Siendo d_{ij} la distancia base entre el bin i el bin j de los histogramas.

Un valor grande de EMD implica una gran diferencia entre ambas distribuciones, mientras que un valor igual a cero indica que ambas distribuciones son idénticas.

3.3.5. Métrica de similaridad

La métrica de Mannan, Ruddock y Wooding, más conocida como métrica de similaridad, introducida por Mannan et al. en 1995, compara la proximidad espacial entre dos conjuntos de fijaciones [MRW95]. El conjunto generado por el modelo $S(x)$ y el conjunto generado por los observadores $H(x)$. Esta métrica compara las distancias lineales entre un conjunto de localizaciones y las fijaciones más próximas del otro conjunto y viceversa.

Un valor alto de este parámetro indica una gran similaridad, mientras que un valor bajo indica lo contrario. Como valor de contraste se calcula la misma métrica sobre todos los pares de sujetos. Si el modelo es capaz de predecir las fijaciones de los sujetos su valor de similaridad será comparable a la similaridad existente entre un humano y otro. Basándose en las distancias correspondientes a dos patrones D_m y D_{mr}, mediante la ecuación 3.6 se define el índice de similaridad I_s:

$$I_s = 100 \cdot [1 - \frac{D_m}{D_{mr}}] \tag{3.6}$$

siendo D_m

$$D_m{}^2 = \frac{n_1 \cdot \sum_{j=1}^{n_2} d_{2j}^2 + n_2 \cdot \sum_{i=1}^{n_1} d_{1i}^2}{2n_1 n_2 \cdot (w^2 + h^2)} \tag{3.7}$$

y siendo D_{mr} definido del mismo modo que D_m. No obstante el conjunto de fijaciones D_{mr} ha sido generado mediante patrones aleatorios del mismo tamaño de la muestra que se compara.

Una de las deficiencias de esta métrica es la gran contribución de los puntos extremos al valor de la suma en la ecuación 3.7. Otra de las principales críticas a esta métrica se puede explicar mediante un ejemplo. Supongamos que dos observadores se encuentran mirando dos objetos. Uno de ellos emplea el 90 % del tiempo observando el objeto A y el resto del tiempo el objeto B. Por el contrario el otro sujeto está en la situación opuesta, o sea destina el 90 % del tiempo a observar el objeto B y el resto lo destina al objeto A. En esta situación la métrica descrita determinará que ambas distribuciones son idénticas [TBG05].

3.3.6. Divergencia de Kullback-Leibler

La divergencia de Kullback-Leibler es una medida que permite discriminar entre dos hipótesis, y está directamente relacionada con la figura de mérito de un detector óptimo. Esta medida tiene su origen en la medida de la información de Fisher, realizando a mayores una serie de suposiciones y aproximaciones de la ecuación 3.8 descritas en [Kul59], pp.26 y extendidas en [SM06].

$$J(\delta) \approx \delta^2 I(0) \tag{3.8}$$

En el contexto de este trabajo será empleada como medida de disimilitud entre dos funciones de densidad de probabilidad. Estas dos funciones se obtienen a partir del mapa de saliencia del modelo $S_M(x)$ usando la ecuación 3.9 para la divergencia de Kullback Leibler empleando los valores de saliencia en las posiciones de fijación de los humanos y en posiciones aleatorias.

$$KL(r \mid h) = \sum r(x) \cdot log\left(\frac{r(x)}{h(x)}\right) \tag{3.9}$$

siendo r la función de densidad de probabilidad para $S_M(x)$ en posiciones aleatorias y h la función de densidad de probabilidad para las posiciones de fijación.

Esta no es una función de distancia ($KL(r \mid h) \neq KL(h \mid r)$) puesto que no es simétrica, y cuando se emplea la base dos para el logaritmo de la ecuación 3.11 esta medida representa la cantidad de información adicional

necesaria para predecir la distribución r a partir de la información de la distribución h. Para evitar la asimetría descrita, Itti et al. proponen el empleo de una extensión simétrica descrita mediante la ecuación 3.10.

$$KL^* = 0.5 \cdot (KL(r \mid h) + KL(h \mid r)) \tag{3.10}$$

Del mismo modo tambien puede ser empleada esta función para diferenciar las distribuciones obtenidas a partir del mapa de atención humano $H_M(x)$ y del mapa de saliencia del modelo $S_M(x)$ conforme a la ecuación 3.11.

$$KL^{**}(p \mid h) = \sum p(x) \cdot log\left(\frac{p(x)}{h(x)}\right) \tag{3.11}$$

siendo p la función de densidad de probabilidad para las fijaciones predichas por el modelo de prueba $S_M(x)$ y h la función de densidad de probabilidad para $H_M(x)$. En ese caso los valores óptimos esperados serán próximos a cero puesto que ambas distribuciones deberán ser lo más parecidas.

3.3.7. Análisis de Curvas ROC

El análisis ROC (Receiving Operator Characteristic) originalmente publicado por Green y Sweets en 1966 [GS66], ha sido posteriormente empleado en múltiples trabajos de muy diversos campos, como por ejemplo el descrito por Wallis et al. en en el campo de la la Neurociencia [WM03], Tatler et al. [TBG05], D. Berg et al. [BBM+09], y recientemente también ha sido publicada una versión especialmente adaptada por el autor de este trabajo a datos de respuestas neuronales individuales [PVLA08]. Para los datos de atención visual se han realizado ligeras modificaciones para adaptar los análisis a este nuevo campo en el que han demostrado ofrecer una información mucho más completa que la media global de la ROC, ya que este análisis permite determinar la selectividad de los datos a lo largo del tiempo.

El análisis ROC da una medida del grado de solape entre dos distribuciones de datos. Para cada fotograma de un vídeo dado se han comparado dos condiciones, generando dos distribuciones de valores de saliencia, F y R. Estas condiciones comparadas serán los valores de la saliencia para las posiciones de fijación (F), verdaderos positivos, frente a los valores de saliencia para las posiciones de no-fijación (R), falsos positivos. En ambos casos las condiciones F y R representan los niveles de gris obtenidos del mapa de saliencia para cada una de las condiciones elegidas. En la figura 3.24 se muestran dos ejemplos de cálculo de la ROC para dos distribuciones con distinto grado de solape.

El índice ROC se obtiene tomando cada nivel de saliencia (criterio) para cada condición, y representando la proporción de F que excede el valor de la

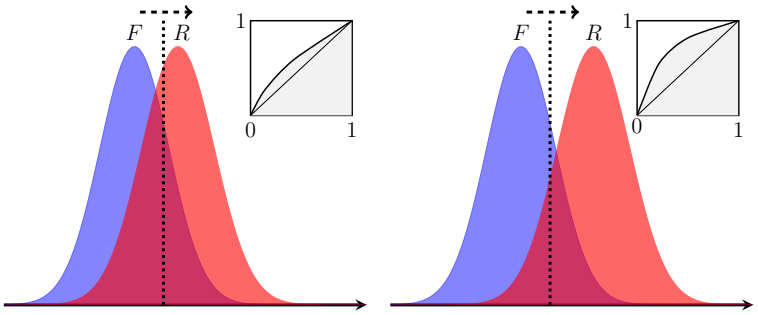


Figura 3.24: Dos distribuciones de datos de ejemplo y representación de la curva ROC instantánea para dos casos con diferente grado de solape.

observación frente a la proporción de R que excede el valor de esa observación. A partir de ahí se obtiene el área bajo la curva ROC. Un valor de 0.5 indicará que las dos distribuciones se solapan. Por ello la saliencia obtenida con ese método sería prácticamente equivalente al azar. Mientras que un valor de 1.0 indicaría que F y R son completamente separables, por lo que los valores de saliencia serían buenos candidatos para representar la saliencia verdadera mostrada por los sujetos.

Para obtener los límites para la selectividad del índice ROC se ha realizado además un test de permutación de los datos, para garantizar que los valores son significativamente diferentes del azar. Para cada fotograma los valores de saliencia fueron asignados de modo arbitrario a las diferentes condiciones (verdaderos y falsos positivos) y los valores de la ROC fueron calculados para cada permutación (N=2000 iteraciones). A continuación se construyó un histograma con todos esos valores y fueron obtenidos los valores del máximo y mínimo para la ROC. Estos valores mínimo y máximo se obtienen como los valores del histograma ROC, para los cuales la probabilidad de encontrar valores mayores/menores es menor que 0.01 ($p \leq 0.01$).

Este procedimiento produce unos márgenes temporales para los valores ROC que varían a medida que cambiamos de fotograma. Por todo lo anterior se establece como criterio para la selectividad de la ROC aquellos puntos en los cuales el índice ROC excede los límites obtenidos mediante el test de permutación.

En la figura 3.25 se muestra un ejemplo de ROC temporal en el que podemos apreciar como a partir de los primeros 5-10 fotogramas se produce una buena correspondencia entre las posiciones de las fijaciones realizadas por los sujetos y las fijaciones predichas por el algoritmo. La línea continua negra representa el valor de la ROC variando entre 0 y 1. Las líneas discontinuas casi horizontales representan los márgenes de significación obtenidos mediante el

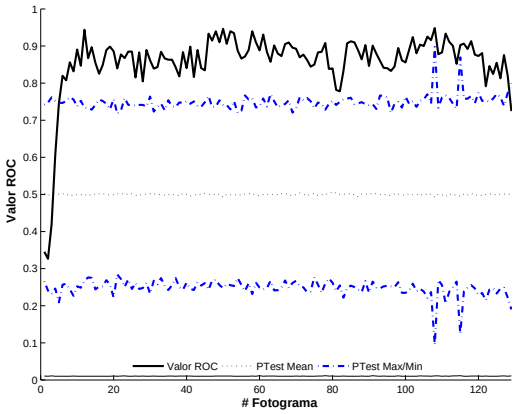


Figura 3.25: Evolución del índice ROC para el modelo AWSD como función del tiempo para el vídeo *RCETraffic1*. La línea negra continua gruesa representa el valor de la ROC variando entre 0 y 1. Las líneas discontinuas casi horizontales representan los márgenes de significación obtenidos mediante el test de permutación.

test de permutación. Los instantes en los que estas líneas se separan de los valores 0.25 y 0.75 son instantes en los que hay pocos datos de fijaciones de los sujetos.

3.3.8. Métricas seleccionadas en este trabajo: Razones objetivas

Una métrica útil, para la descripción del tipo de datos analizados en este trabajo, debería ser sensible a diferencias en las distribuciones de los datos pero no debería verse afectada por la variabilidad de los mismos. Además la medida devuelta debería ser lo más constante posible independientemente de la cantidad de datos disponible, aumentando tan solo en la precisión del valor. En particular con los experimentos de atención visual el número de sujetos y el de fijaciones sobre un estímulo son limitados, por lo que la robustez de la métrica es un aspecto importante. Habitualmente se emplean varias métricas para analizar los resultados de los modelos para garantizar que las conclusiones extraídas no dependan de la elección de la métrica de evaluación.

El empleo de la métrica KLD se recomienda en situaciones en las que se valora la capacidad de un modelo para aprovechar las similitudes del comportamiento entre múltiples sujetos. La principal deficiencia que se le achaca a este modelo es la no linealidad, que conlleva que los resultados sean dependientes del número de sujetos. Otra característica negativa de esta métrica es

la dependencia ante transformaciones monótonas no lineales al contrario que sucede con la métrica ROC. Por ejemplo, si se aplica una función monótona continua, tal como un logaritmo, la valoración de la ROC no se ve afectada [IB09]. La divergencia de Kullback-Leibler se ha incluido principalmente como método de evaluación comparativo, puesto que parte de los trabajos referenciados en esta tesis emplean esta medida.

Las curvas ROC se han incluido por la completitud de la información que aportan. Mediante la ROC y el test de permutación, disponemos de una medida global que permite valorar el posicionamiento de nuestro algoritmo frente a otras propuestas. Pero además se accede a una información más detallada, que permite, de un modo sencillo, comprobar en qué instantes de qué vídeos es más eficiente[3] cada uno de los modelos evaluados.

En este trabajo la elección de las curvas ROC combinadas con el Test de permutación se ha realizado basándose en las características de la información que devuelve dichas métrica y en el hecho de que cumple con los requisitos básicos esperados para una buena métrica; ausencia de parámetros configurables y presencia de una escala intuitiva sin la necesidad de comparar con un modelo patrón [WBKK11]. Las demás opciones descritas en esta sección también han sido evaluadas observando gran variabilidad en los resultados y por ello un comportamiento poco robusto para la tipología de los datos empleados.

Tanto para la métrica ROC, la KLD y la NSS, además de los márgenes obtenidos mediante el test de permutación se ha empleado la técnica no paramétrica del *bootstrap* sobre cada fotograma para calcular los intervalos de confianza (CI) sobre cada uno de los valores instantáneos de las métricas. Esto ha dado lugar a las nuevas métricas que en este trabajo hemos denominado $s\text{-}AUC$, $s\text{-}KLD$ y $s\text{-}NSS$. Para ello además de tomar la distribución R de posiciones de fijación aleatorias sobre cada fotograma, se ha repetido este proceso múltiples veces dando lugar a n distribuciones R_n de posiciones aleatorias. Esta forma de cálculo de los CI evita las engorrosas suposiciones de normalidad de los datos necesarias para el empleo de las desviaciones estandar [ZOM11].

3.4. Metodología de evaluación

Los datos de partida que se han empleado en este trabajo están sujetos a ciertas limitaciones, como puede ser el hecho de presentar los vídeos a los sujetos en lugar de grabar al sujeto directamente frente a las situaciones

[3]Eficiencia en este caso se refiere a la comparación entre la realidad objetiva de los resultados de los sujetos y la respuesta de la metodología propuesta.

reales que representan los vídeos. Esto introduce un factor humano, esto es, el responsable de grabar las situaciones reales, que introduce sesgos en los propios vídeos de forma consciente o inconsciente. Para la obtención de los resultados mostrados en el capítulo 4 de este trabajo se han intentado corregir estos fenómenos que ya han sido observados durante la comparación de modelos previos y que se detallan a continuación.

3.4.1. Sesgo central y bordes

El fenómeno del sesgo central y su influencia en los resultados de los algoritmos no es un resultado reciente [Bus35], sino que su influencia ya ha sido ampliamente descrita en la literatura tanto para atención visual aplicada sobre imágenes estáticas [LMLCBT06, Tat07, ZK11], como en el caso de secuencias dinámicas [ZTW09, BSI12]. Este efecto ha sido achacado tanto a motivos fisiológicos como a motivos de carácter *top-down*. En el primer caso, la no uniformidad de la distribución de fotoreceptores en la retina descrita en la sección 1.1.1 origina una menor sensibilidad a las altas frecuencias en las regiones periféricas. Este efecto ha sido observado tanto en tareas de simple inspección visual [PLN02, TBG05], como en tareas de búsqueda guiada [Wan95, PLN02]. Además también se ha observado una ligera tendencia a la realización de sacadas de menor distancia [BAS75, TBV06].

En cuanto a los motivos de carácter *top-down*, en muchos casos este sesgo es generado por los propios creadores de los datos [Tat07]. Otro factor que puede interferir es el hecho de que habitualmente los experimentos de visión emplean estímulos de fijación previos a la tarea que suelen estar centrados en la pantalla.

En las figuras 3.2, 3.3 y 3.9 se han presentado varios fotogramas de ejemplo con diferentes BD de prueba. En todos ellos se puede apreciar cómo el cámara ha centrado los objetos relevantes de la escena. Es habitual que la persona que captura un vídeo dirija la cámara hacia las zonas más relevantes, con ello centra las zonas más salientes en la escena. Esto hace que en muchos casos los observadores, de modo inconsciente, tiendan a seleccionar las localizaciones centrales en lugar de las potencialmente más importantes.

Otro de los casos en los que se aprecia este efecto son los videojuegos. La perspectiva isométrica del juego fuerza al avatar del jugador a estar siempre en una posición próxima al centro, ligeramente desplazado hacia abajo. Se puede ver un ejemplo en la figura 3.4 en uno de los vídeos de muestra de la BD CRCNS. Del mismo modo hay otros muchos ejemplos en las bases de datos revisadas en los que se ve reflejado este efecto.

En cuanto al efecto de los bordes, en la bibliografía también se remarca el hecho de que los resultados de los algoritmos pueden ver su rendimiento

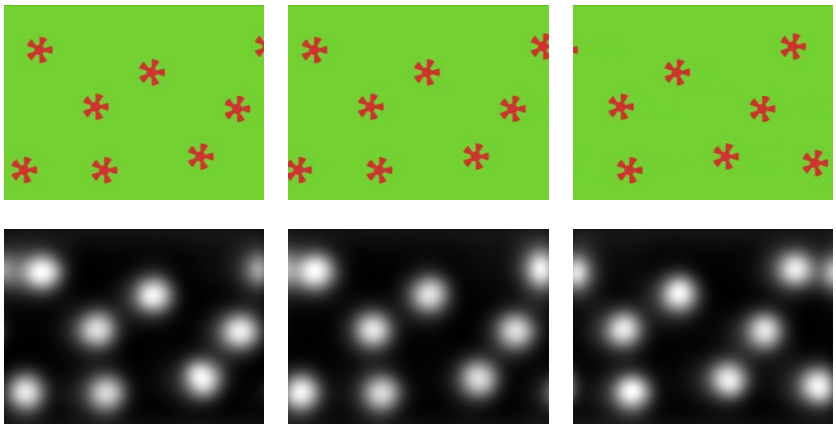


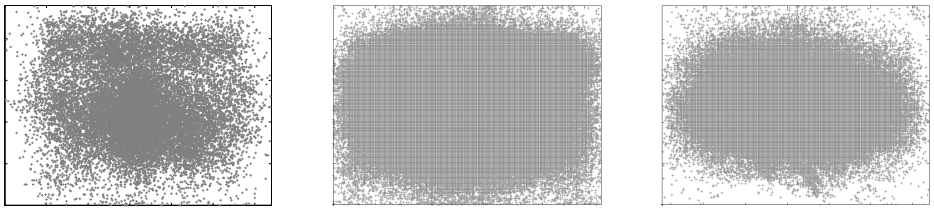
Figura 3.26: Imagen de una secuencia sintética de la BD de CITIUS, en la fila superior se muestran los fotogramas originales y en la inferior los correspondientes mapas de saliencia del modelo AWSD.

claramente afectado por la presencia de artefactos en los bordes del mapa de saliencia. En la figura 3.26 se muestra un ejemplo de este efecto al emplear el algoritmo AWSD para varios fotogramas de una secuencia sintética en la que los objetos se están desplazando hacia la izquierda. En la parte superior izquierda un objeto va a desaparecer, pero antes de eso surge una sombra en la zona en la que no hay nada; algo semejante se puede apreciar en la parte superior derecha en la que tras haber aparecido la figura surge una sombra en la zona donde tampoco hay nada.

Para el caso de los vídeos podemos definir dos tipos de bordes. Los espaciales se refieren a la región próxima al borde de los fotogramas y darán lugar a efectos de borde debido a la actuación de los diferentes operadores espaciales de los algoritmos empleados. Los temporales vendrán dados por los instantes en los que los bloques de imágenes procesadas comienzan y terminan dando igualmente lugar a efectos de borde (ya sea para un solo bloque, en cuyo caso el efecto se traslada al inicio y fin de la secuencia o para diferentes bloques si los algoritmos realizan procesamiento por bloques de imágenes).

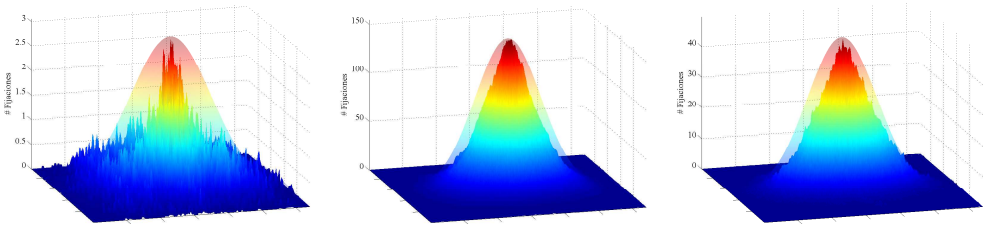
3.4.2. Compensación del sesgo central

Tal como se describe en la sección previa, se ha observado que la distribución de las fijaciones de los sujetos no es en absoluto uniforme sobre todo el espacio de la imagen. Si representamos todas las fijaciones de cada uno de los sujetos para todos los vídeos de varias de las BD de prueba empleadas,



(a) CRCNS (b) DIEM (c) Holly2Train

Figura 3.27: Fijaciones de todos los sujetos para todos los vídeos de la BD CRCNS (20.326), DIEM (1.050.020) y Holly2Train (345.455).



(a) BD CRCNS (b) BD DIEM (c) BD Holly2Train

Figura 3.28: Histogramas de las posiciones XY y el modelo de Gaussiana ajustado para la BD (a) CRCNS, (a) DIEM y (c) Holly2Train.

figura 3.27, podemos apreciar que existe un claro sesgo hacia la región central. El potencial predictivo de dicho sesgo puede ofrecer un umbral inferior de referencia para los modelos computacionales, ya que es independiente de la escena y del sujeto. En algún caso un modelo simple consistente en un mapa de saliencia construido como una gausiana centrada en la imagen, puede llegar a ofrecer resultados mejores que alguno de los algoritmos existentes en la bibliografía. En este trabajo se ha probado esta hipótesis mediante un modelo de gaussiana ajustado a los histogramas de posiciones XY de las fijaciones oculares, como el que se muestra en la figura 3.28, para tres BD de prueba (función envolvente transparente).

Ya que es evidente la presencia de este efecto en nuestros datos, para resolver este problema, durante la validación de los resultados, una de las soluciones propuestas es emplear las propias posiciones de las fijaciones de los sujetos como puntos de muestreo de la saliencia [DE03, Tat07, ZTW09]. De este modo al realizar las comparaciones de los diferentes modelos con lás métricas descritas en la sección 3.3 se emplean como falsos positivos en lugar de posiciones aleatorias, posiciones de fijación de otros fotogramas diferentes al actual. Mediante esta técnica se consigue que las dos distribuciones comparadas presenten el mismo sesgo y por ello no existan diferencias significativas

asociadas a dicho efecto.

3.4.3. Evaluación de la fiabilidad temporal

Una de las aportaciones de este trabajo es la evaluación de la fiabilidad temporal de la *s-AUC* al aplicarla sobre las BD de prueba. La combinación de esta medida con el test de permutación aporta una gran información acerca de la cantidad de tiempo durante la cual el algoritmo probado ofrece unos resultados significativos. Esta información, además de añadir una medida global para todo el conjunto de vídeos, permite detectar de modo muy rápido y por simple inspección de las funciones *s-AUC* los instantes en los que las metodologías fallan. Por ello resulta muy práctico y permite inferir las situaciones en las que se pueden detectar problemas con cada uno de los algoritmos. En la figura 4.1 del capítulo de resultados se muestra un ejemplo.

Otra característica que afecta a la fiabilidad temporal, y que está presente en los modelos dinámicos, es el hecho del procesamiento de los vídeos en bloques. Desde el momento en el que se agrupan fotogramas se produce una difusión lateral de la información de los fotogramas vecinos, por lo que la selección del tamaño de estos bloques determinará la cantidad de tiempo que se está anticipando, o revisando según sea el caso, dicha información.

Un efecto perceptible de la elección del tamaño de bloque es un ligero aumento y disminución del nivel de contraste en periodos del tamaño del bloque, dando lugar a un efecto de parpadeo con un frecuencia relacionada con el propio tamaño del bloque. Este efecto se pueder minimizar tomando tamaños de bloque más cortos o mediante el empleo de análisis con ventanas deslizantes que contemplen el solape de fotogramas entre bloques. Este efecto de parpadeo se puede apreciar en otros modelos evaluados como el GBVS.

Por último, otro factor que condiciona los resultados es el hecho de emplear la duración de las fijaciones para crear las funciones gaussianas asociadas a las posiciones de fijación de los sujetos. Al hacerlo se está anticipando una información que no está disponible al inicio del periodo de fijación. La selección del parámetro α, descrito en la sección 3.2.2, determina la influencia de la duración de las fijaciones en el mapa de atención de los sujetos. Puesto que en este trabajo tanto la ROC como NSS emplean las distribuciones de los valores de saliencia del mapa $S_M(x)$ del modelo en las posiciones de fijación y en posiciones de no fijación. No se comparan los mapas $S_M(x)$ y $H_M(x)$ directamente sino que se emplean las posiciones XY de fijación. Por ello, esta contribución influye a nivel visual en las representaciones en las que se muestra el mapa de atención obtenido a partir de las fijaciones humanas.

Capítulo 4

Resultados experimentales

Este capítulo se centra en la evaluación de la capacidad de predicción de fijaciones oculares del modelo de saliencia AWSD, descrito en el capítulo 2, frente a otros modelos dinámicos del estado del arte, descritos en la sección 1.2.2. Asumiendo la *hipótesis de saliencia visual*, se realiza una comparación del rendimiento de los modelos frente a los resultados experimentales obtenidos por los sujetos, durante la realización de tareas de observación libre de vídeos, con diferentes bases de datos y haciendo uso de las métricas descritas en la sección . Para validar los resultados se han empleado las bases de datos externas descritas en la sección 3.1 y nuestra propuesta, la BD CITIUS descrita en la sección 3.

El procedimiento seguido se puede resumir en los siguientes pasos: *recuperación* de la información de los movimientos oculares para todos los vídeos de la BD, *creación* del mapa de atención a partir de las fijaciones humanas para cada fotograma de cada vídeo, *creación* del mapa de saliencia del modelo comparado asociado a cada fotograma de cada vídeo de la BD, *comparación* de ambos mapas mediante las metodologías adecuadas.

Organización del capítulo

Inicialmente, en la sección 4.1.1, se indicarán los modelos finalmente seleccionados para comparar el modelo AWSD con modelos del estado del arte de otros autores. A continuación, en la seccción 4.1.2, se indicarán las metodologías seleccionadas, mostrando resultados que permiten apreciar los problemas de las metodologías previas y los beneficios de la aplicación de la metodología sugerida. En la sección 4.1.3 se empleará dicha metodología para evaluar la capacidad de los humanos de predecir al colectivo y a continuación en la sección 4.1.4 se analizará globalmente la calidad de los modelos de saliencia probados frente a todas las BD que han sido evaluadas. Por último,

en la sección 4.2.1, se mostrará el comportamiento sobre vídeos individuales en los que se muestran efectos relevantes que permiten apreciar el funcionamiento del modelo tanto sobre escenas naturales como vídeos sintéticos.

4.1. Predicción de fijaciones oculares.

Las bases de datos elegidas, en conjunto, presentan escenas suficientemente heterogéneas, lo cual permite comparar el funcionamiento de los modelos de modo global y que los resultados obtenidos, aun no siendo totalmente independientes de la categoría de los vídeos, puedan ser extrapolables.

4.1.1. Modelos comparados

Todos los modelos que han sido utilizados para realizar la comparación, descritos en la sección 1.2.2, transforman las imágenes de entrada en una distribución bidimensional de valores, que representa el concepto original de Koch y Ulman de mapa de saliencia [CS85]. Estos modelos actualmente se encuentran disponibles para descargar a través de las webs de los autores. Los modelos se han elegido en base a tres criterios: 1) que incorporen características dinámicas en su diseño, 2) que sean representativos de las distintas opciones que marcan el estado del arte actual y 3) accesibilidad al código original de los autores. La configuración empleada durante la evaluación de los distintos modelos es la aportada por los autores y no se varía para ningún experimento.

Modelos de saliencia computacional elegidos. Se ha comparado nuestro modelo frente a diez modelos de saliencia dinámicos: GBVSm [HKP07], PQFT [GZ10], SEOD [SM09a], SUNDAY [ZTW09], SURP [IB05], ESA-D [RKSH10], DCOF [ZLR+13] y dos variantes del modelo seminal de Itti [IKN98] denotados por CIOFM y CIORFM.

Modelos de referencia. En la comparativa se introducen además otros tres modelos de *referencia* [JDT12]:

- H50 performance: El mapa de saliencia de este modelo, en cada fotograma, se conforma a partir del 50 % de las fijaciones totales del conjunto de humanos que observó el vídeo. Es una cota superior muy optimista que permite relativizar las puntuaciones alcanzadas por los modelos en cada una de las medidas.

- CENTER: El mapa de saliencia se conforma mediante el streching de una gaussiana simétrica para lograr la relación de aspecto de cada fotograma. Predice que las posiciones centrales de la imagen son más relevantes que las más alejadas del centro. Si el valor alcanzado por este modelo es similar al obtenido por CHANCE nos indicará que las medidas tienen compensado el sesgo espacial.

- CHANCE: Este modelo selecciona alatoriamente los píxeles salientes en cada fotograma (ruido blanco de la misma duración que los vídeos) y marcará la cota inferior en las medidas.

Otro parámetro que manejaremos, es el valor de la congruencia inter observador (IOC) promedio [MB13] que cuantifica la capacidad que tiene cada observador en predecir las fijaciones del resto de sujetos. Dado que se emplea la *s-AUC* como medida, su valor está acotado entre 0 y 1. Un valor de 1 indica que el conjunto de observadores fija en las mismas áreas y un valor nulo o bajo, indica que las secuencias de exploración de los sujetos no está correladas, lo que se traduce en una fuerte variabilidad entre los observadores y, por ende, en bajos rendimientos de los modelos computacionales.

Para la creación de los modelos, los autores han empleado diferentes plataformas y lenguajes de programación, por lo que no se ofrece una comparativa de tiempos de ejecución de los diferentes modelos. El modelo AWSD, GBVSm, PQFT, CIORFM, ESA-D, DCOF y SEOD, han sido desarrollados sobre MATLAB. El modelo CIOFM y SURP, han sido desarrollados en C++ y se ejecutan sobre una máquina virtual Linux descargable desde la web del ILAB. Por último, el modelo SUNDAY también se ha implementado en C++, usando las librerías de OpenCV y se ha ejecutado sobre una máquina virtual Linux.

4.1.2. Metodologías empleadas

En este trabajo se han validado los resultados empleando varias metodologías diferentes que siguen la evolución temporal de las publicaciones.

- `Análisis original de ITTI`. Empleado en sus publicaciones con las dos bases de datos de vídeos denominadas ORG y MTV, correspondientes a dos experimentos publicados en [Itt04, Itt05] y [CI06a, CI06b] respectivamente. Emplea las métricas ROC y KLD para extraer las conclusiones.

- `Análisis de Zhang y Cotrel`. Descrito originalmente por Zhang et al. [ZTM+08] y posteriormente empleado por H.J. Seo et al. [SM09b]

entre otros autores. También emplea la ROC y KLD con la corrección del efecto del sesgo central.

- `Análisis ROC propio`. Adaptación de los métodos publicados por otros autores [TBG05, BT06, HKP07, KWSF06] y por el autor de este trabajo en [PVLA08, PVLA09], para la obtención de un análisis *s-AUC* y *s-NSS* dependientes del tiempo [LAGDFVP13].

Los análisis que aquí se detallan se fundamentan en el uso de la ROC, KLD o NSS como medidas de la calidad de los resultados, pero todos ellos difieren en el nivel de detalle al que se aplica las métricas.

En el primer caso, con el análisis original de ITTI, en primer lugar se seleccionan las fijaciones de todos los sujetos sobre todos los vídeos del experimento elegido. A continuación se localiza el fotograma correspondiente al instante en el que se ha iniciado una fijación. Sobre ese mismo fotograma, pero en el vídeo de saliencia del modelo, se identifican los valores de saliencia en las posiciones de fijación y se almacenan además otros N valores para puntos de coordenadas aleatorias. El resultado de este procedimiento es un conjunto de valores de saliencia, correspondiente a los puntos en los que los sujetos han fijado realmente, y otro con múltiples puntos aleatorios en los que los sujetos no han realizado fijaciones. Empleando los datos normalizados a los valores máximos de saliencia se crean dos distribuciones y se aplica la ROC para comprobar que ambas distribuciones no son compatibles. Si esto sucede entonces se puede afirmar que los resultados del modelo mejoran los resultados del azar.

Empleando los vídeos de la base de datos CITIUS, al aplicar este mismo análisis (usando los scripts del autor replicando el procedimiento que se describe en las publicaciones relacionadas para los modelos CIOFM [IKN98, IK00, IK01] y SURP [IB05, IB09]) con los diez modelos elegidos y los tres de referencia, se ha observado que los valores de la ROC/KLD sugieren un mejor comportamiento por parte del modelo GBVS (0.800/0.695), el AWSD (0.796/0.685) y el CIORFM (0.793/0.665). Estando los demás modelos claramente separados en la clasificación. Se puede apreciar el fallo de esta metodología de evaluación, ya descrito por otros autores [ZTW09], al comparar los valores de los mejores modelos con el propio modelo CENTER que alcanza un valor de 0.784/0.606. CENTER no dista demasiado de los primeros modelos y además es muy superior varios de los modelos evaluados. El hecho de que el modelo GBVSm salga bien valorado en esta comparación es razonable puesto que el propio modelo implementa una compensación del sesgo central [HKP07]. El modelo CHANCE como es de esperar presenta un valor ROC próximo a 0.5, y un valor de KLD que aun siendo bajo tambien supera a varios de los modelos comparados.

El segundo análisis, propuesto por Zhang y Cotrel [ZTM+08], aumenta ligeramente el nivel de detalle. En él se seleccionan las fijaciones de todos los sujetos que están fijando durante la duración de un determinado fotograma. Este proceso se repite para todos los fotogramas de cada vídeo. Se comparan las distribuciones de aciertos (valores del mapa de saliencia del modelo, correspondientes a posiciones en las que los sujetos fijan) frente a falsos positivos (valores del mapa de saliencia en posiciones XY aleatorias, que se corresponden con posiciones de fijación de los sujetos en otros instantes). La selección de posiciones XY de fijaciones de fotogramas diferentes al actual compensa el efecto del sesgo central.

Este análisis añade una gran cantidad de información, ya que ahora no sólo se dispone de una medida global del modelo con la BD, sino también de una medida de calidad individual del modelo asociada a cada uno de los vídeos. No se muestran resultados con este segundo análisis ya que la siguiente metodología que se describe a continuación incluye a ésta.

El tercer análisis, propuesto para este trabajo, añade una medida instantánea de la fiabilidad de los resultados mediante el test de permutación. Esta medida indica la calidad individual de cada uno de los fotogramas y ofrece una medida de fiabilidad de la medida de calidad (ROC/NSS), e informa de la cantidad de datos disponible en cada momento, teniendo así la certeza de que el valor obtenido sea incompatible con una medida aleatoria de saliencia [LAGDFVP13]. En este último análisis también se hace la compensación del efecto de sesgo central mediante la elección de falsos positivos entre los aciertos de fotogramas de otros vídeos diferentes al actual tal como se describe en la sección 3.4.2.

Este análisis emplea métodos híbridos [MB13], [BSI12] que involucran mapas de saliencia y un conjunto de fijaciones producidas por los observadores humanos. Por la naturaleza dinámica de los datos, el número de fijaciones durante cada fotograma del vídeo es muy bajo. Por ello, es necesario utilizar medidas capaces de producir resultados lo más fiables posible con pocas fijaciones por fotograma. En Wilming et al. [WBKK11] se analiza el rendimiento de varias medidas de evaluación y concluyen que el aerea bajo la curva ROC (AUC) parece ser la mejor opción para la evaluación de los modelos de predicción de fijaciones. Como segunda opción, se decanta por la NSS [PIIK05] dada su baja demanda de datos y facilidad para compensar el center-bias. Descartamos la KLD, con bases teóricas muy consistentes pero con requerimiento altos de datos para estimar las densidades de probabilidad y su sensibilidad frente a valores erráticos.

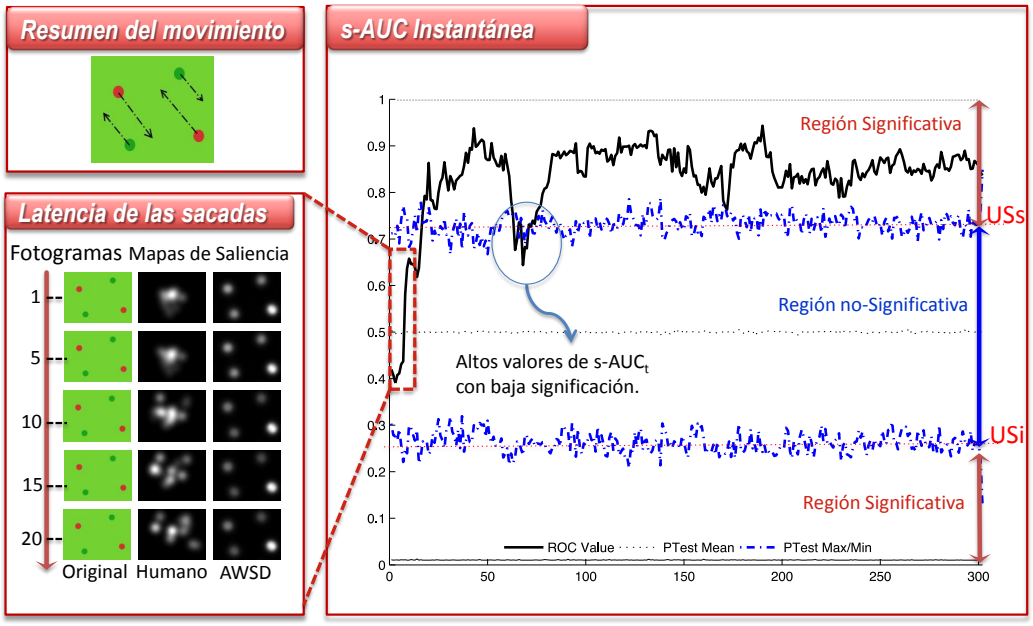


Figura 4.1: Representación de $s\text{-}AUC_t$ para el modelo AWSD. La línea negra continua gruesa representa el valor de la $s\text{-}AUC$ variando entre 0 y 1. Las líneas discontinuas casi horizontales representan los márgenes de significación obtenidos mediante el test de permutación para $p \leq 0.01$

Significación estadística y temporal

Para comparar el comportamiento de un conjunto de modelos de saliencia es necesario realizar una clasificación basada en una medida global sobre la base de datos. Esta ordenación debe ser complementada con mecanismos estadísticos que nos aporten información para responder a las siguientes cuestiones: 1) ¿las ordenaciones producidas son estadísticamente significativas? y 2) ¿podría el azar, a la hora de elegir la muestra, explicar los valores alcanzados por las medidas ($s\text{-}AUC$ o $s\text{-}NSS$) en cada instante del vídeo? La primera cuestión se contesta verificando si los intervalos de confianza de la medida global están superpuestos. Para la segunda, utilizaremos un test de permutaciones con un valor de $p \leq 0.01$ que evalúa la fuerza de la evidencia numérica para desechar el azar como una probable explicación suficiente. Tanto para la obtención de los intervalos de confianza BCa (*Bias-Corrected-and-Accelerated* [XHZ11]) como para el test de permutaciones [PVLA09, PVLA08], utilizamos el bootstrap por su enfoque menos restrictivo y más robusto que el estadístico clásico, puesto que nos evita asumir suposiciones de normalidad de la distribución del estadístico analizado (métodos no paramétricos).

En la figura 4.1 se muestra un ejemplo del test de permutación para la

s-AUC. La línea continua negra representa el valor instantáneo de la medida variando entre $[0, 1]$ y las líneas discontinuas azules representan los márgenes de significación obtenidos mediante el test de permutación, tal que rechazamos la hipótesis nula (H_o), con el valor de $p \leq 0.01$. El recuento de todos los instantes durante los que los valores de $s\text{-}AUC_t$ están por encima del límite señalado (USs), o por debajo de (USi), nos indicará el porcentaje temporal durante el cual valores altos o bajos de la medida, que produce el modelo, son significativos (denotados por $\%USs$ y $\%USi$, respectivamente). Cuanto más tiempo el modelo produzca valores superiores a USs mayor será la concordancia entre los valores de las fijaciones y los puntos de saliencia producidos por el modelo. Debemos resaltar que estos umbrales están directamente relacionados con la cantidad de datos disponibles (n_{fix}) en cada instante. Si $n_{fix} \to 0$, entonces $USs \to 1$ y $USi \to 0$, ensanchando la zona no significativa y provocando que valores elevados de la medida sean poco significativos. Los instantes en los que los valores de la medida son inferiores a USi existe una alta discordancia fiable entre las fijaciones de los humanos y el modelo (casi todas las fijaciones no son puntos de saliencia) y esto *puede* ocurrir en los 10-15 fotogramas iniciales con independencia del contenido del vídeo. Este efecto inicial que revela el test de permutación en los fotogramas iniciales (ampliado en el marco de la izquierda de la figura 4.1), se explica porque las fijaciones de los humanos al iniciar la observación de un nuevo vídeo, se van al centro del fotograma y la mirada de cada observador requiere un tiempo para evolucionar hasta las regiones reales de interés que está visionando (latencia de las sacadas $\sim$ 200 ms).

Otro ejemplo en el que el test de permutación muestra un comportamiento muy interesante es el de la figura 4.2. Para el vídeo *mtvclip28* de la base CRCNS, el modelo AWSD tiene una $s\text{-}AUC$=0.72 elevada pero una baja significación temporal $\%USs = 0$. Esto sucede como consecuencia del ensanchamiento de los umbrales de significación ($\%USs$, $\%USi$) debido al escaso número de fijaciones existente en cada fotograma (3.6 fijaciones de promedio).

Además de estos tests, al igual que Riche et al., en este trabajo utilizaremos el factor de Kendal (FK) para cuantificar el grado de concordancia entre las ordenaciones producidas por las medidas utilizadas. Dicho factor, para nuestro caso se obtiene como el cociente entre la varianza de la suma de las puntuaciones obtenidas por los modelos y la varianza máxima (la obtenida cuando todas las puntuaciones son idénticas) [RDM+13]. Valores inferiores a 0.5 indican un desacuerdo y valores superiores a 0.7 indican fuerte acuerdo. Superado el umbral de fuerte acuerdo, valores de 0.9 indican acuerdo

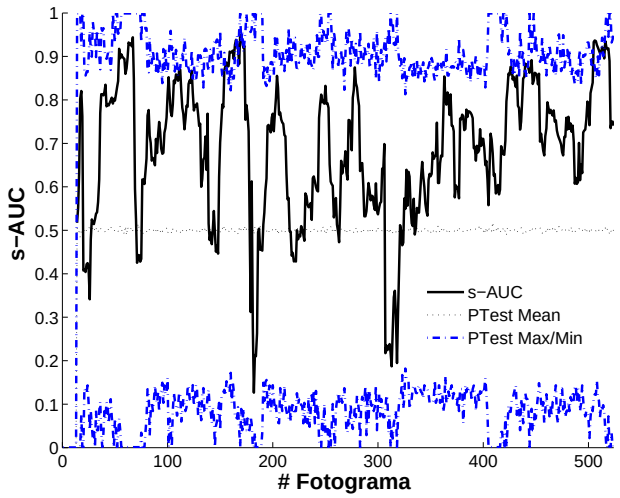


Figura 4.2: Para el vídeo mtvclip28 de CRCNS, el modelo AWSD tiene una $s\text{-}AUC$=0.72 elevada pero una baja significación temporal $\%USs = 0$.

inusualmente alto y 1 total acuerdo.

4.1.3. Resultados indivíduos vs. población

Además de determinar la bondad de los resultados obtenidos con los diferentes modelos, se ha realizado la comparación haciendo uso del humano como referencia de clasificación. Para ello se ha tomado un subconjunto de fijaciones asociadas a un subconjunto de los humanos que han sido empleados como grupo de predicción (técnica descrita por Tilke Judd et al. [JDT12] entre otros). De este modo se valora en qué medida los resultados de una parte de los humanos son suficientes para predecir los de toda la población. Esta técnica de comparación entre observadores, semejante al modelo MEP (posición promedio de los ojos) descrito por Borji et al. [BSI11], al modelo IO (inter observador) [BSI12], es válida cuando es posible poner en correspondencia los fotogramas de los diferentes vídeos observados por los sujetos. En tareas interactivas, más allá de la simple observación libre, esta comparación pasa a no ser una cota superior, ya que no existe esa correspondencia directa entre fotogramas para diferentes observadores.

Los resultados de la $s\text{-}AUC$ para este análisis, sobre la BD CITIUS, se muestran en la figura 4.3 (serie ⊣□⊢). El valor de la $s\text{-}AUC$ aumenta a medida que el porcentaje de fijaciones se aproxima al 100 %. Tal como se puede apreciar en esta gráfica, con porcentajes de fijaciones bajos, ya se alcanzan resultados de $s\text{-}AUC$ próximos a 0.8. Se ha tomado como sujeto de referencia

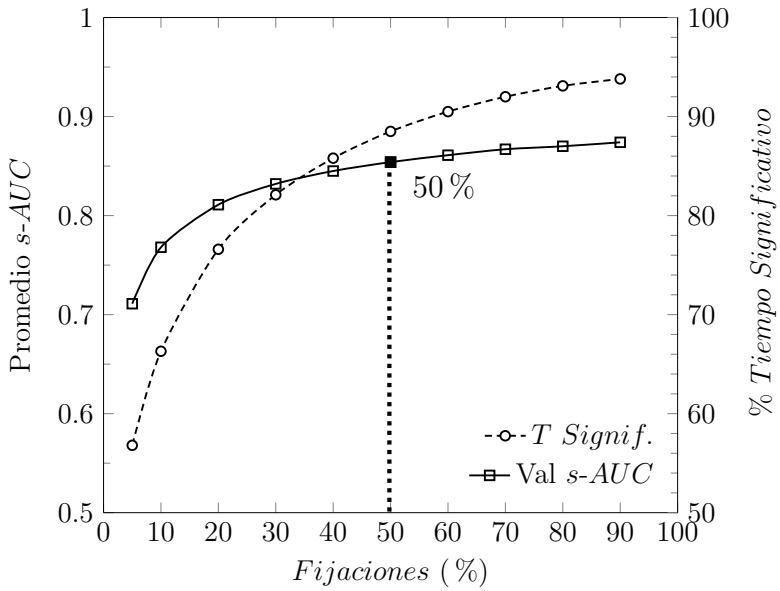


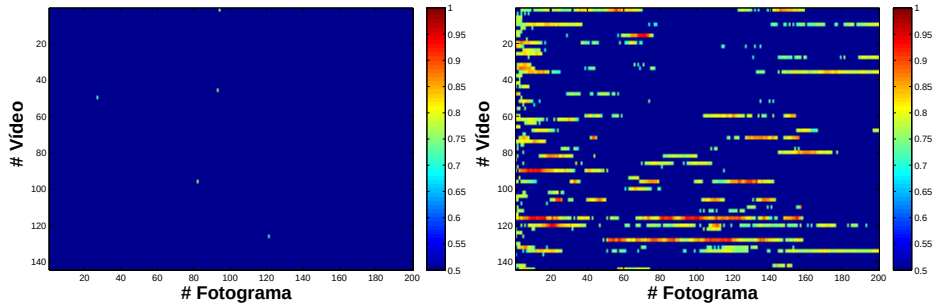
Figura 4.3: Capacidad de predicción de los sujetos del comportamiento global de la población. Se muestra (—□—) en trazo continuo el valor de la *s-AUC* y (-○-) en trazo discontinuo el porcentaje de tiempo durante el cual la *s-AUC* es significativa respecto a los umbrales del test de permutación.

o *SuperHumano* al que se corresponde al empleo del 50 % del número total de fijaciones, este valor resaltado en la figura 4.3 (marcador —■—) presenta un valor de *s-AUC* de 0.85 (modelo H50). Este sujeto tiene una capacidad predictiva mucho mayor que la de ninguno de los sujetos de la BD de modo aislado y da información acerca del umbral superior al cual se aproximan los modelos evaluados.

Del mismo modo, tal como se aprecia en la figura 4.3 (serie -○-), si representamos el porcentaje de tiempo durante el cual los mapas de atención dan resultados de ROC significativa, vemos que aumentan rápidamente. Estos valores temporales también han sido generados variando el porcentaje de las fijaciones de los sujetos empleadas.

4.1.4. Resultados globales con *s-AUC/s-NSS*

El objetivo del siguiente experimento ha sido la comparación de los diez modelos evaluados, entre los cuales se encuentra el modelo AWSD y los tres modelos de referencia usando la base de datos CITIUS. De este modo hemos podido comprobar cuales son los que mejor predicen las fijaciones de los sujetos con las diferentes categorías de esta BD.



(a) RANDOM (b) CENTER

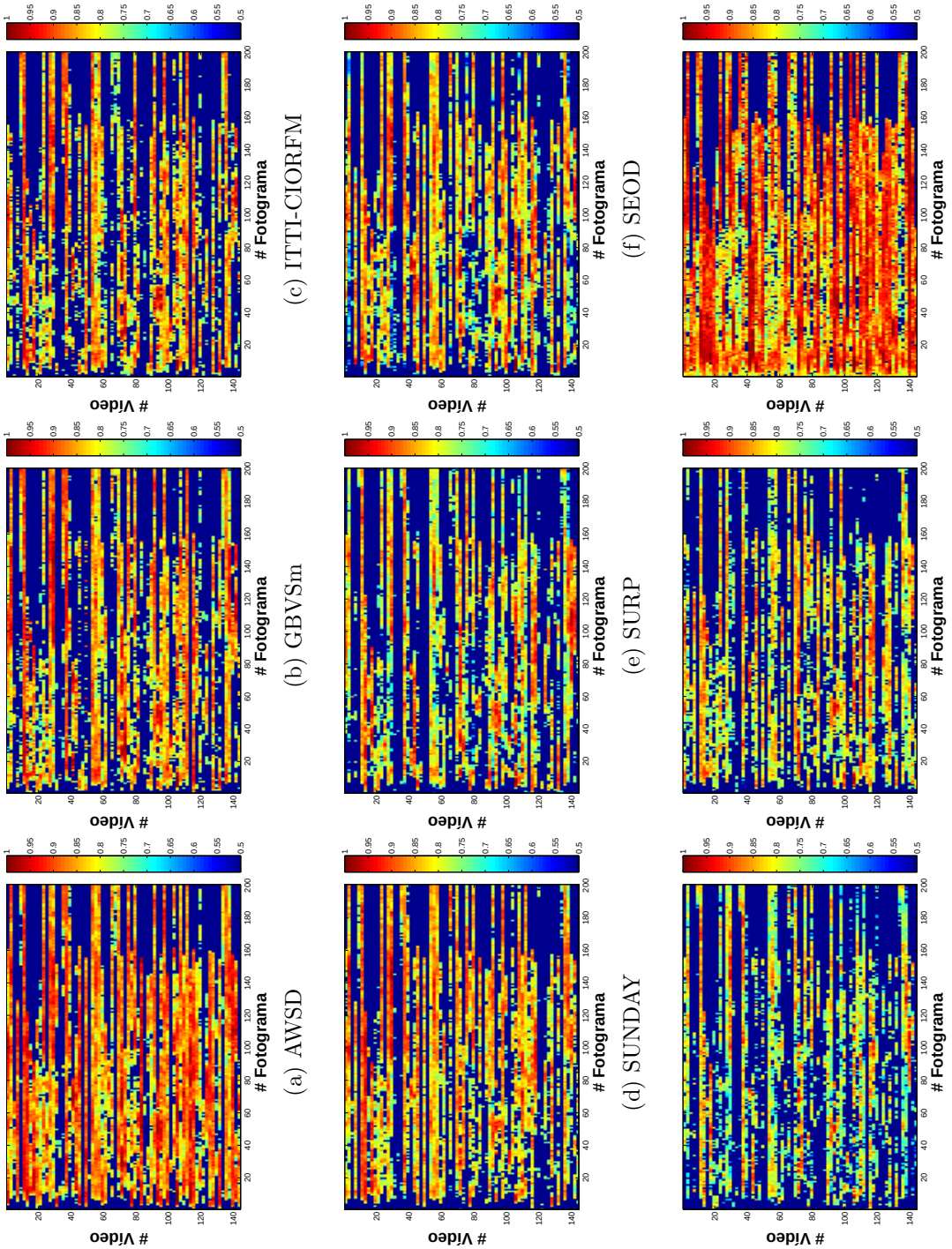
Figura 4.4: $s\text{-}AUC_t$ para los modelos RANDOM(a) y CENTER(b) con todos los vídeos de la BD CITIUS. La codificación de color de azul a rojo representa el valor de $s\text{-}AUC$ variando de 0.5 a 1.0.

Análisis cualitativo- En las figuras 4.4 y 4.5 se puede apreciar la evolución temporal de los valores de las $s\text{-}AUC$ para todos los vídeos de la BD CITIUS utilizando los mejores ocho de los diez modelos aquí descritos y el modelo de referencia H50. El aspecto de los modelos CHANCE y GAUSS es semejante para todas las BD por lo que se muestra en la figura 4.4 y no se volverá a presentar para las demás BD. Los modelos de la figura 4.5 están en orden descendente de la (a) a la (h) en función del valor global de la $s\text{-}AUC$. Los dos modelos peor valorados con cada BD tampoco se muestran para facilitar la visualización de los resultados a página completa. La codificación de color se ha realizado mostrando tan sólo los valores de la $s\text{-}AUC$ comprendidos entre 0.5 y 1, ya que los valores significativos inferiores indicarían que la distribución aleatoria ofrece un mejor resultado que los valores de saliencia generados por el método. En la figura 4.6 se muestra la información equivalente pero haciendo uso de la métrica $s\text{-}NSS$.

En las figuras 4.5a (AWSD) y 4.5i (H50) se aprecia claramente un mayor contenido de rojo (valores próximos a 1 para la $s\text{-}AUC$). Lo mismo sucede con las figuras 4.6a y 4.6i para la métrica $s\text{-}NSS$. Esto sugiere valores $s\text{-}AUC$ y $s\text{-}NSS$ altos durante un mayor porcentaje de la duración de los vídeos. Este resultado se interpreta como un mejor comportamiento global del modelo AWSD ante los vídeos de la BD CITIUS visualmente comparable al obtenido por el modelo humano H50.

Experimento I.Vídeos Sintéticos

Análisis cuantitavo global- En la tabla 4.1 se muestra la clasificación de los distintos modelos para las dos métricas seleccionadas ($s\text{-}AUC$ y $s\text{-}NSS$)

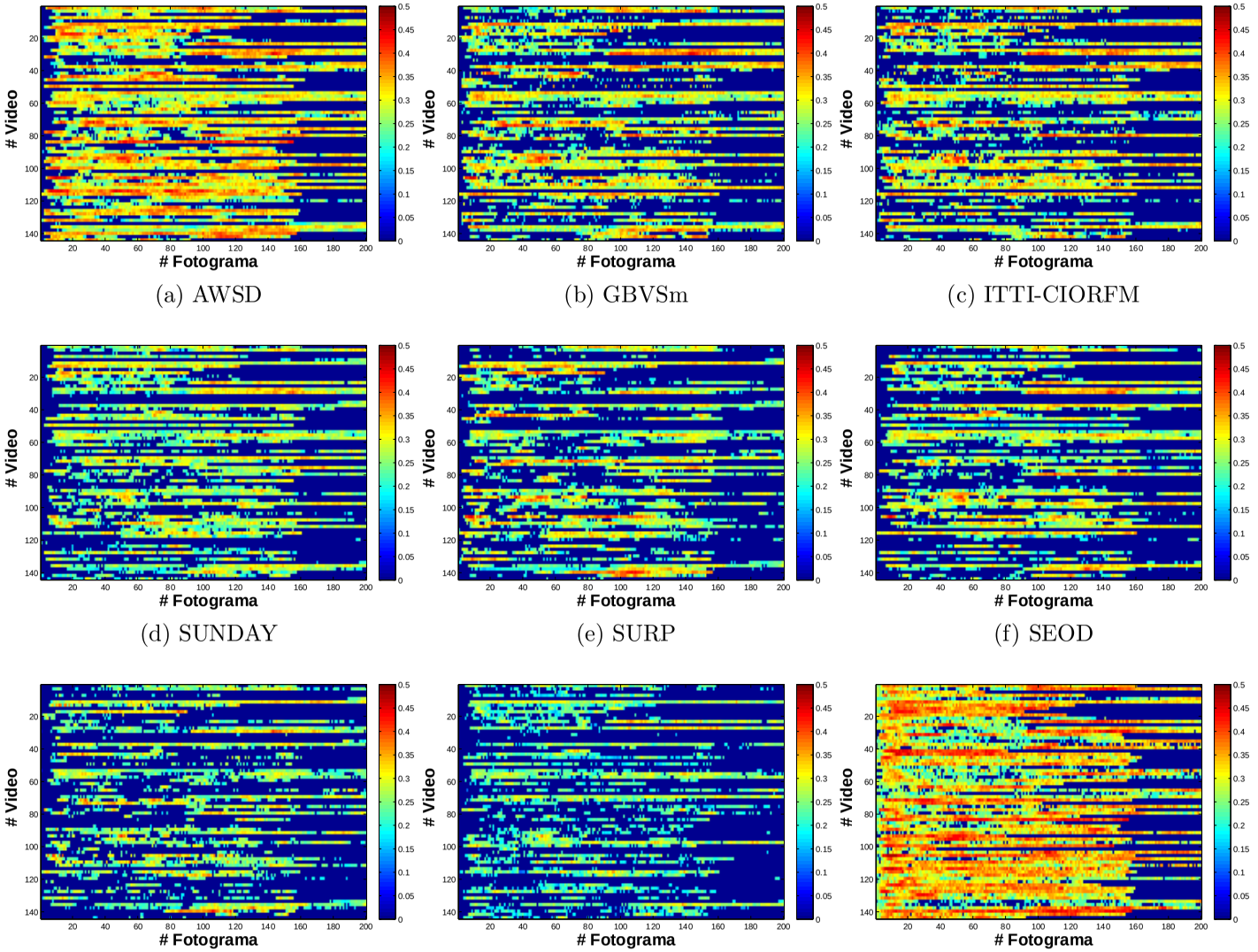


(a) AWSD (b) GBVSm (c) ITTI-CIORFM

(d) SUNDAY (e) SURP (f) SEOD

(g) CIOFM (h) PQFT (i) H50

Figura 4.5: $s\text{-}AUC_t$ para los modelos elegidos (a)-(h) y el modelo de referencia (i) con todos los vídeos de la BD CITIUS. La codificación de color de azul a rojo representa el valor de $s\text{-}AUC$ variando de 0.5 a 1.0.



(a) AWSD (b) GBVSm (c) ITTI-CIORFM

(d) SUNDAY (e) SURP (f) SEOD

(g) CIOFM (h) PQFT (i) H50

Figura 4.6: $s\text{-}NSS_t$ para los modelos elegidos (a)-(h) y el modelo de referencia (i) con todos los vídeos de la BD CITIUS. La codificación de color de azul a rojo representa el valor de $s\text{-}NSS$ variando de 0 a 0.5.

	CITIUS-S dataset			
	s-AUC		*s-NSS*	
Modelo	mean (BCa C.I.)^{Orden}	%USs	mean (BCa C.I.)^{Orden}	%USs
AWSD	**0.813(0.803,0.823)**^{1}	83.7	**0.266(0.258,0.274)**^{1}	83.8
GBVSm	0.806 (0.795,0.816) ^{2}	80.3	0.251 (0.242,0.259) ^{2}	77.9
CIORFM	0.784 (0.774,0.795) ^{3}	73.8	0.247 (0.238,0.256) ^{3}	74.6
SUNDAY	0.782 (0.773,0.791) ^{4}	76.1	0.214 (0.207,0.221) ^{4}	69.2
SURP	0.734 (0.725,0.742) ^{5}	69.4	0.192 (0.184,0.199) ^{7}	58.6
SEOD	0.728 (0.719,0.735) ^{6}	69.3	0.198 (0.191,0.204) ^{5}	62.4
CIOFM	0.722 (0.714,0.730) ^{7}	69.3	0.195 (0.188,0.202) ^{6}	59.2
PQFT	0.719 (0.710,0.727) ^{8}	54.5	0.160 (0.153,0.167) ^{9}	41.9
ESA-D	0.690 (0.681,0.699) ^{9}	52.2	0.153 (0.146,0.159)^{10}	40.5
DCOF	0.681 (0.671,0.691)^{10}	49.7	0.162 (0.154,0.169) ^{8}	52.2
H50	0.831 (0.821,0.841)	87.4	0.280 (0.272,0.288)	85.4
CENTER	0.479 (0.463,0.495)	15.9	-0.015(-0.029,-0.002)	17.0
CHANCE	0.499 (0.494,0.504)	0.0	-0.001(-0.005,0.004)	0.1

Tabla 4.1: Rendimiento de los modelos de saliencia para la predicción de las fijaciones oculares con los vídeos de la BD CITIUS-S.

y la cota máxima que puede alcanzar cada una de ellas (H50). La concordancia entre las dos ordenaciones es elevada $FK = 0.96$. Independientemente de la métrica, el AWSD encabeza el ordenamiento de los modelos comparados en cuanto a los valores medios de las métricas y al porcentaje de tiempo que el modelo produce valores de las métricas compatibles con los mapas de saliencia humanos. No obstante, si atendemos a la superposición de los CI de ambas métricas, la superioridad del AWSD frente al GBVSm no es estadísticamente significativa pero si con respecto al resto de los modelos. Es interesante resaltar que los resultados del modelo H50 y el AWSD son compatibles estadísticamente como indican sus CI; sin embargo, esta afirmación no se puede mantener para el segundo clasificado (GBVSm) con ninguna de las métricas mostradas.

Las discrepancias de ordenamiento entre la *s-AUC* y la *s-NSS* se producen en los modelos SURP, SEOD, CIOFM y PQFT cuyos resultados no son estadísticamente distinguibles. Finalmente, destacar que los valores alcanzados por modelo CENTER son compatibles con el CHANCE, revelando una compensación eficaz del C-B en ambas medidas.

Análisis cuantitativo temporal- La fiabilidad de los resultados de estas métricas quedan avalados por el test de permutación. Para el AWSD (tabla 4.1), el porcentaje de significación es del $83,7/83,8\,\%$ del tiempo to-

tal que conforma la base de vídeos. Excepto para el segundo clasificado (80, 3/77, 9 %), este porcentaje cae rápidamente a partir del tercer clasificado, llegando a porcentajes, para el último clasificado DCOF, de tan solo el 49.7/52.2 %. Es decir, para este modelo el azar puede explicar los valores alcanzados por la métrica *s-AUC* durante el $\sim$ 51 % del tiempo y con la métrica *s-NSS* el $\sim$ 48 % !.

En la tabla 4.2 mostramos las pequeñas gráficas de la *s-AUC* temporal (*s-AUC*$_t$), para el modelo AWSD, indicando el nivel superior de significación con una línea continua roja, además se incluye en la tabla el valor promedio de la *s-AUC* y el valor promedio de *%USs* del test de permutación para cada vídeo de la base CITIUS-S. En los valores iniciales de las gráficas de *s-AUC*$_t$ se aprecia el efecto de la latencia de las sacadas durante el cual los humanos inician la exploración de los primeros fotogramas en el centro al contrario que el modelo AWSD, produciéndose una marcada discordancia para los valores de *s-AUC*. Para los efectos *pop-out* utilizados, existe una elevada concordancia del modelo AWSD con la respuesta de los observadores humanos, donde el 67 % de los vídeos posee un valor de la *s-AUC* superior a 0.8 y la significación de este valor ocurre durante más del $USs \geq 85\,\%$ de los fotogramas de cada vídeo.

Análisis cualitativo- En la figura 4.7, comparamos visualmente los mapas de saliencia de los seis mejores modelos de saliencia respecto del mapa de atención generado por los humanos. Se muestran seis vídeos (los cuatro en los que hay mayor concordancia entre los humanos y los modelos, y los dos más difíciles teniendo en cuenta la tabla 4.2). El *cluster* más difícil está constituido por dos vídeos que representan una serie de puntos móviles, que a partir de un determinado instante, dan lugar a un agrupamiento que forma una estructura pentagonal que atrapa a la atención. Como vemos en la gráficas de *s-AUC*$_t$, y a pesar del bajo valor de *s-AUC*, vemos que los dos primeros modelos consiguen captar ese agrupamiento.

Experimento II.Vídeos Naturales

Análisis cuantitativo global- En la tabla 4.3 se muestra la clasificación de los modelos comparados que denota una fuerte coincidencia en las ordenaciones sugeridas por ambas medidas ($FK = 0.91$). Las medidas toman valores en los intervalos $[0.499, 0.797]$ para la *s-AUC* y para la s-NSS $[-0.003, 0.252]$. El sesgo central aparece compensado en ambas medidas tomado valores similares al modelo CHANCE. Respecto a los modelos computacionales destacamos, que con independencia de la medida, el AWSD encabeza el ranking y es estadísticamente superior al segundo modelo clasificado. Las principales diferencias respecto a la base CITIUS-S radican en la variación del rendimiento

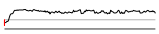
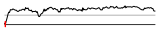
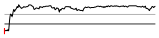
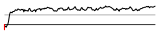
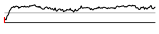
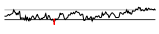
Id. Vídeo	$sAUC_t$	$sAUC$	$\%USs.$
SC01		0.91 ± 0.04	98.0
SC02		0.87 ± 0.05	97.0
SC03		0.86 ± 0.09	96.4
SC04		0.86 ± 0.06	97.4
SC05		0.85 ± 0.07	91.4
SC06		0.85 ± 0.06	97.4
SC07		0.85 ± 0.06	92.6
SC08		0.85 ± 0.06	92.7
SC09		0.85 ± 0.05	98.0
SC10		0.85 ± 0.05	97.7
SC11		0.85 ± 0.05	97.7
SC12		0.84 ± 0.09	84.3
SC13		0.84 ± 0.08	95.7
SC14		0.84 ± 0.06	92.1
SC15		0.83 ± 0.09	91.4
SC16		0.83 ± 0.08	89.4
SC17		0.83 ± 0.07	95.4
SC18		0.83 ± 0.06	93.7
SC19		0.82 ± 0.07	86.1
SC20		0.77 ± 0.12	66.2
SC21		0.76 ± 0.07	71.3
SC22		0.75 ± 0.14	62.6
SC23		0.75 ± 0.12	64.9
SC24		0.75 ± 0.10	55.0
SC25		0.73 ± 0.13	57.7
SC26		0.71 ± 0.15	49.0
SC27		0.69 ± 0.15	54.7

Tabla 4.2: Valor instantáneo de la *s-AUC* con CITIUS-S

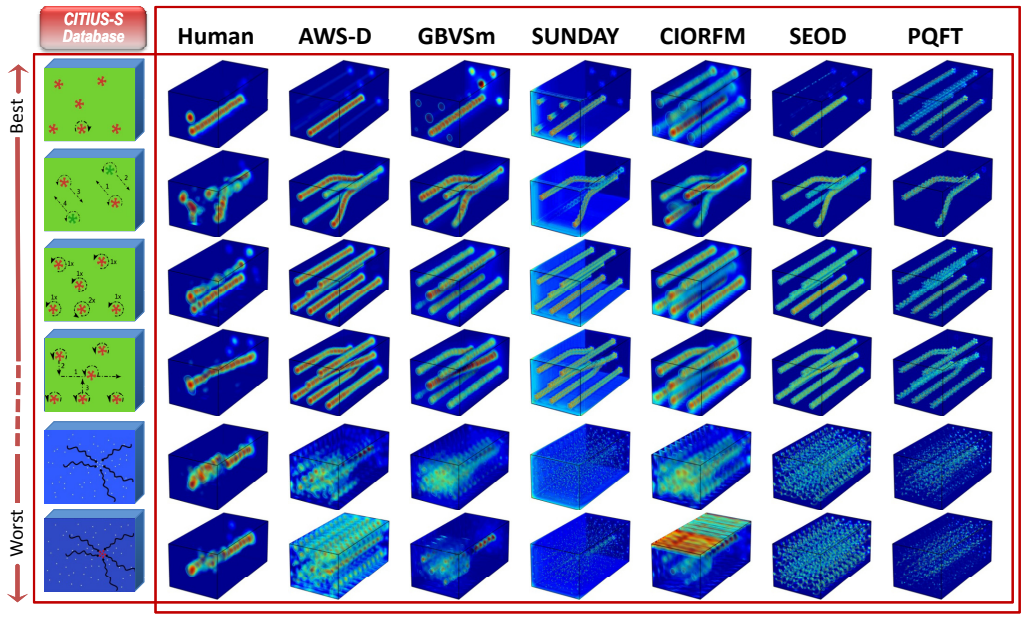


Figura 4.7: Mapas de atención de los humanos frente al los mapas de atención de los seis mejores modelos para seis ejemplos representativos de la BD CITIUS-S.

de los modelos GBVSm (segundo clasificado en la anterior prueba) que ahora retrocede hasta el quinto lugar (según *s-AUC*) o tercer lugar (según s-NSS). El modelo PQFT (antepenúltimo lugar en la anterior prueba) alcanza el segundo puesto (*s-AUC*) o el quinto (*s-NSS*). También se produce un mayor alejamiento entre el rendimiento de H50 respecto del primer clasificado siendo ahora diferencias significativas. Según *s-AUC* los modelos PQFT, SEOD y SUNDAY conforman el segundo grupo de modelos estadísticamente equivalentes, seguido por los modelos GBVSm, CIORFM y SURP y cerrando el ordenamiento CIOFM, ESA-D y DCOF. Salvo las excepciones señaladas, la *s-NSS* produce similares resultados.

Análisis cuantitativo temporal- Con los vídeos reales se acentúan las diferencias en el comportamiento temporal de los modelos. El porcentaje de tiempo que el AWSD (tabla 4.3) produce valores significativos con ambas medidas es del $USs = 75\,\%$ del tiempo total de la base de datos. Este porcentaje desciende en más de 15 puntos porcentuales para los modelos que se incluyen en el segundo cluster y en casi 60 puntos respecto del último. No obstante, el modelos H50 supera en 20 puntos al AWSD, relegándolo a medio camino entre el observador humano ideal y el resto de los modelos computacionales evaluados. En la parte superior de la figura 3.18 de la sec-

	CITIUS-R dataset			
	s-AUC		*s-NSS*	
Modelo	mean (BCa C.I.)^{Orden}	%USs	mean (BCa C.I.)^{Orden}	%USs
AWSD	**0.797(0.786,0.807)**[1]	74.7	**0.252(0.243,0.260)**[1]	74.9
PQFT	0.728 (0.718,0.737) [2]	58.8	0.153 (0.146,0.160) [5]	45.4
SEOD	0.727 (0.716,0.738) [3]	58.8	0.152 (0.143,0.160) [6]	46.3
SUNDAY	0.724 (0.713,0.735) [4]	56.1	0.174 (0.165,0.182) [2]	52.9
GBVSm	0.697 (0.683,0.710) [5]	49.4	0.160 (0.149,0.171) [3]	48.2
CIORFM	0.686 (0.672,0.700) [6]	47.8	0.158 (0.146,0.169) [4]	48.1
SURP	0.674 (0.662,0.686) [7]	44.3	0.147 (0.137,0.157) [7]	43.7
CIOFM	0.635 (0.625,0.643) [8]	37.7	0.116 (0.108,0.124) [8]	32.5
ESA-D	0.632 (0.619,0.644) [9]	27.3	0.116 (0.107,0.125) [9]	27.1
DCOF	0.499 (0.486,0.512)[10]	16.6	-0.003(-0.013,0.007)[10]	15.9
H50	0.884 (0.875,0.891)	96.3	0.326 (0.319,0.332)	95.7
CENTER	0.521 (0.503,0.538)	16.5	0.016 (0.002,0.031)	17.7
CHANCE	0.500 (0.494,0.505)	0.1	-0.000(-0.005,0.004)	0.1

Tabla 4.3: Rendimiento de los modelos de saliencia para la predicción de las fijaciones oculares durante la observación de los vídeos de la BD CITIUS-R.

ción 3.2, presentamos los fotogramas resumen que componen la base de datos CITIUS-R

En la tablas 4.5 y 4.6 mostramos las gráficas de $s\text{-}AUC_t$, el valor promedio de la *s-AUC* y el promedio de $\%USs$ del test de permutación para cada vídeo de la base CITIUS-R. En la primera tabla se muestran los vídeos con fondo estático (etiqueta RCE) y en la segunda, con fondo móvil (etiqueta RCD). Para todos ellos, de nuevo, es visible en los instantes iniciales el efecto del fenómeno de la latencia de las sacadas. Los valores de *s-AUC* y los porcentajes temporales son muy elevados para el modelo AWSD en donde el 53 % de los vídeos de la base CITIUS-R tiene un valor de la $s\text{-}AUC \geq 0.8$ y solo para el $\approx 9\,\%$ el valor de $s\text{-}AUC \leq 0.6$. El 51 % de vídeos de CITIUS-R posee alto índice de fiabilidad temporal ($\%USs \geq 80\,\%$) y, aunque inferior al obtenido en CITIUS-S, es elevado para vídeos reales. Otro hecho relevante, reflejado en la tabla 4.4, es que el comportamiento de AWSD se ve afectado ligeramente por el hecho de que los vídeos tengan o no fondo móvil, tal y como refleja el valor medio y el tiempo de significación para ambas categorías: para fondo móvil $s\text{-}AUC_{RCD} = 0.79$, $\%USs = 72.6$ y para fondo estático $s\text{-}AUC_{RCE} = 0.80$; $\%USs = 78.1$. No obstante, el resto de modelos acusa negativamente este efecto como se muestra en la tabla 4.4.

Análisis cualitativo- En la figura 4.8, mostramos una comparativa entre

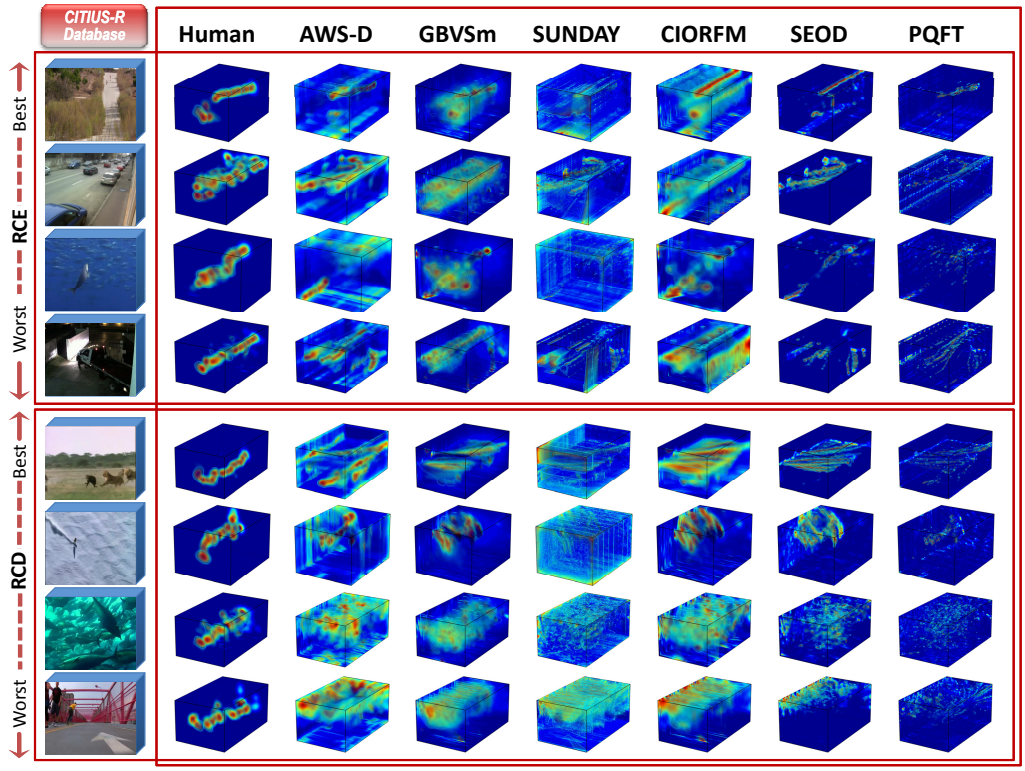


Figura 4.8: Mapas humano frente a los mapas de saliencia de los seis mejores modelos, para ocho ejemplos representativos de la BD CITIUS-R.

	CITIUS-RCE		CITIUS-RCD	
Id. Modelo	$sAUC$(BCa C.I.)	$\%USs$	$sAUC$(BCa C.I.)	$\%USs$
AWSD	0.807(0.796,0.818)	78.1	0.790(0.779,0.800)	72.6
SUNDAY	0.782(0.772,0.793)	73.9	0.685(0.674,0.696)	44.3
SEOD	0.779(0.769,0.788)	75.2	0.693(0.681,0.705)	47.9
PQFT	0.756(0.746,0.765)	67.4	0.709(0.700,0.719)	53.0
GBVSm	0.714(0.701,0.727)	52.8	0.685(0.670,0.699)	47.1
CIORFM	0.708(0.695,0.721)	53.2	0.672(0.657,0.686)	44.2
SURP	0.702(0.691,0.713)	52.1	0.655(0.642,0.667)	39.1
CIOFM	0.672(0.663,0.680)	46.8	0.610(0.600,0.619)	31.6
ESA-D	0.636(0.624,0.649)	27.0	0.629(0.617,0.640)	27.5
DCOF	0.602(0.591,0.613)	29.5	0.431(0.417,0.445)	8.0

Tabla 4.4: Comparativa de los modelos dinámicos con CITIUS RCE y RCD.

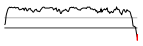
Id. Vídeo	$sAUC_t$	$sAUC$	$\%USs.$
RCD01		0.92 ± 0.06	98.1
RCD02		0.92 ± 0.04	98.7
RCD03		0.91 ± 0.09	95.0
RCD04		0.90 ± 0.11	94.2
RCD05		0.89 ± 0.05	98.1
RCD06		0.88 ± 0.07	94.9
RCD07		0.87 ± 0.07	97.3
RCD08		0.87 ± 0.05	98.7
RCD09		0.86 ± 0.08	98.1
RCD10		0.85 ± 0.14	84.3
RCD11		0.83 ± 0.07	92.9
RCD12		0.83 ± 0.06	90.6
RCD13		0.82 ± 0.11	86.0
RCD14		0.81 ± 0.09	81.5
RCD15		0.80 ± 0.07	85.1
RCD16		0.79 ± 0.15	67.3
RCD17		0.79 ± 0.10	69.8
RCD18		0.78 ± 0.12	72.6
RCD19		0.77 ± 0.13	70.7
RCD20		0.76 ± 0.14	56.8
RCD21		0.76 ± 0.12	60.1
RCD22		0.74 ± 0.12	52.8
RCD23		0.72 ± 0.15	50.3
RCD24		0.72 ± 0.13	56.7
RCD25		0.66 ± 0.14	35.4
RCD26		0.60 ± 0.12	21.2
RCD27		0.52 ± 0.11	2.5
RCD28		0.48 ± 0.11	0.0

Tabla 4.5: Gráfica del valor instantáneo de la *s-AUC* y valores promedio, máximo y mínimo para los vídeos naturales de la BD de CITIUS-RCD.

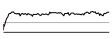
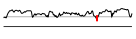
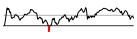
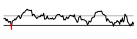
Id. Vídeo	$sAUC_t$	$sAUC$	$\%USs.$
RCE01		0.92 ± 0.05	98.4
RCE02		0.90 ± 0.03	100.0
RCE03		0.88 ± 0.10	90.1
RCE04		0.88 ± 0.09	96.1
RCE05		0.88 ± 0.08	97.5
RCE06		0.88 ± 0.06	98.7
RCE07		0.86 ± 0.09	96.2
RCE08		0.83 ± 0.08	87.7
RCE09		0.83 ± 0.06	93.9
RCE10		0.81 ± 0.05	91.9
RCE11		0.79 ± 0.13	69.2
RCE12		0.79 ± 0.09	78.4
RCE13		0.77 ± 0.06	75.9
RCE14		0.74 ± 0.11	50.3
RCE15		0.67 ± 0.09	27.5
RCE16		0.62 ± 0.15	24.3
RCE17		0.56 ± 0.12	10.5

Tabla 4.6: Gráfica del valor instantáneo de la *s-AUC* y valores promedio, máximo y mínimo para los vídeos naturales de la BD de CITIUS-RCE.

los mapas de saliencia de los mejores modelos y el mapa de atención humano. En esta figura mostramos los dos vídeos más fáciles y más difíciles (según tablas 4.6 y 4.5) para las dos categorías en las que hemos dividido la base de datos CITIUS-R: RCE sin *egomotion* y RCD con *egomotion*).

Experimento III. Bases de datos públicas externas

El objetivo de este experimento ha sido la comparación de los diez modelos evaluados, entre los cuales se encuentra el modelo AWSD y los tres modelos de referencia pero con bases de datos externas, figura 4.9 y tablas 4.9-4.13. Así podemos descartar que nuestro modelo de algún modo esté sesgado debido a su validación inicial con la BD CITIUS o por el contrario es totalmente aplicable a diferentes escenarios.

Como denominador común, y con independencia de la medida, el modelo AWSD se sitúa en cabeza del ranking en todas las BD. Además en dos de ellas, los resultados del modelo AWSD presentan una diferencia estadísticamente significativa respecto al segundo clasificado.

Los valores tomados por los factores de Kendal para todas las BD evaluadas están en el intervalo $[0.95, 0.99]$ revelando fuertes concordancias entre las medidas. Las posiciones por detrás del AWSD no son estables y los modelos se alternan, con bastante superposición entre sus CI. Como cabría esperar, al aumentar la variedad y amplitud de los estímulos, los modelos computacionales separan sus rendimientos de la cota superior marcada por H50 hasta un 20 %.

Respecto a los porcentajes de tiempo de significación de la métrica *s-AUC*, se aprecia que para las BD CITIUS (-S y -R), DIEM y GC dichos valores están en torno al $60-70$ % para los modelos de mejor comportamiento, llegando a un $25-30$ % para los peores modelos. Para las AE-UCFSA, ASCMN y la CRCNS los porcentajes de tiempo son claramente inferiores, llegando a valores tan bajos como los mostrados con la BD de CRCNS. En la BD de CRCNS, debemos resaltar que los porcentajes de significación son inferiores al 13 % debido al escaso número de fijaciones en cada fotograma que aporta esta BD (aprox. un promedio de 3.6 fijaciones en cada fotograma). Esto hace que los intervalos de significación se ensanchen, y aunque los valores de *s-AUC* sean elevados, no pueden superar los valores instantáneos de significación (ver Fig. 4.2). Por ello, los resultados obtenidos para esta base de datos no son significativos estadísticamente. El conjunto de fijaciones por fotograma de la CRCNS debería ser ampliado, en la linea sugerida por [JDT12].

Análisis cualitativo- En la figura 4.10, mostramos una comparativa visual entre los mapas de saliencia de los mejores modelos con la humana para dos

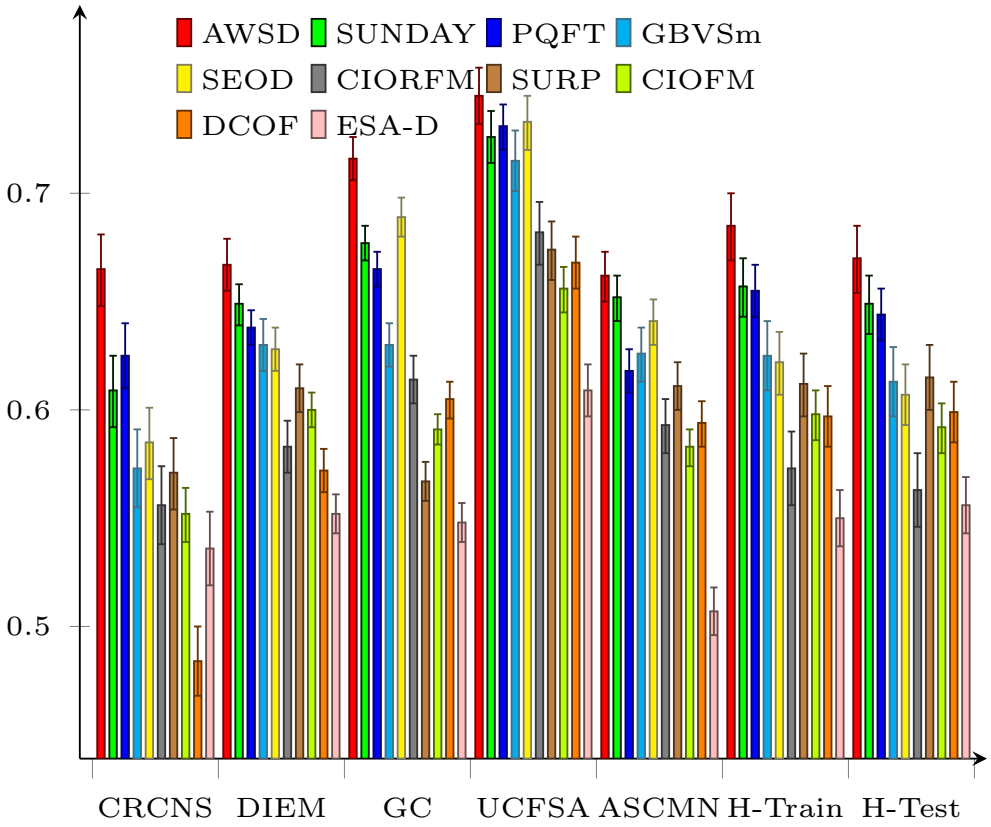


Figura 4.9: Clasificación de los modelos de saliencia mediante la *s-AUC* con las BD públicas de prueba.

vídeos de cada una de las restantes bases de datos utilizadas en el artículo.

Los resultados de *s-AUC* y la *s-NSS* que se presentan en la tabla 4.8 muestran globalmente valores menores de ambas medidas respecto a la BD CITIUS, la disminución de los porcentajes de tiempo de significación para todos los modelos, se pueden justificar debido a que durante este experimento se ha incluido la información auditiva. Dicha información sesga la conducta de los usuarios dirigiéndola hacia el origen del sonido, algo que los modelos no contemplan.

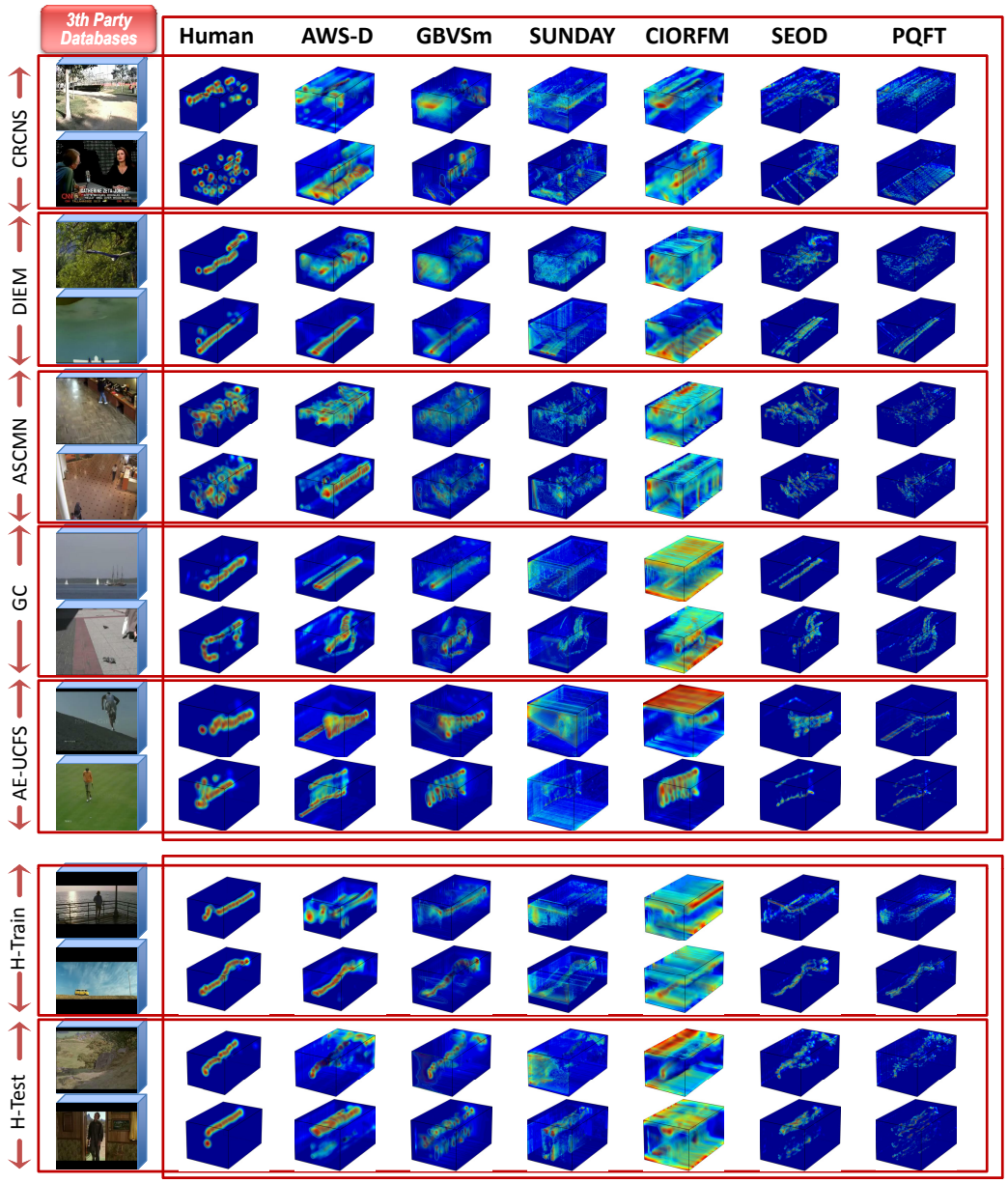


Figura 4.10: Mapas de saliencia con dos vídeos de ejemplo de las BD públicas utilizadas (CRNNS, DIEM, ASCMN, CG, AE-UCFS y Holly2).

	Base de datos CRCNS					
	s-AUC			s-NSS		
Modelo	mean (BCa C.I.)^{Orden}	%USs	%USi	mean (BCa C.I.)^{Orden}	%USs	%USi
AWSD	**0.665(0.648,0.682)**[1]	2.1	0.0	**0.141(0.127,0.155)**[1]	16.0	0.6
PQFT	0.625 (0.610,0.640) [2]	4.7	0.6	0.098 (0.085,0.111) [2]	10.0	1.8
SUNDAY	0.609 (0.592,0.625) [3]	1.7	0.1	0.088 (0.075,0.102) [3]	10.2	0.9
SEOD	0.585 (0.568,0.601) [4]	4.3	1.3	0.063 (0.049,0.077) [4]	8.6	2.5
GBVSm	0.573 (0.554,0.591) [5]	1.5	0.1	0.063 (0.048,0.078) [5]	10.0	1.4
SURP	0.571 (0.554,0.587) [6]	4.7	1.5	0.061 (0.046,0.074) [6]	10.5	3.3
CIORFM	0.556 (0.538,0.574) [7]	1.0	0.1	0.048 (0.033,0.063) [7]	9.2	1.3
CIOFM	0.552 (0.540,0.565) [8]	13.1	11.3	0.030 (0.018,0.041) [8]	9.0	10.2
ESA-D	0.525 (0.509,0.541) [9]	2.3	0.5	0.025 (0.012,0.039) [9]	7.6	2.1
DCOF	0.483 (0.467,0.499)[10]	1.7	1.0	-0.009 (-0.023,0.004)[10]	6.4	3.9
H50	0.872 (0.865,0.879)	36.2	0.0	0.330 (0.323,0.336)	44.4	0.0
CENTER	0.512 (0.492,0.532)	0.6	0.1	0.010 (-0.006,0.027)	9.7	2.5
CHANCE	0.500 (0.485,0.514)	0.1	0.0	-0.000 (-0.012,0.011)	4.9	1.5

Tabla 4.7: Medidas *s-AUC* y *s-NSS* de los diez modelos de saliencia y los tres de referencia con la BD ASCMN (KF=0.99). Intevalos de confianza BCa (95 %) y los pct. de tiempo en los que las medidas exceden los umbrales del test de permutación ($\%USs$ y $\%USi$). Para CRCNS el valor medio de $IOC = 0.6$ ($IOC_{CRCNS-ORG} = 0.578$ e $IOC_{CRCNS-MTV} = 0.622$)

	Base de datos DIEM					
	s-AUC			s-NSS		
Modelo	mean (BCa C.I.)^{Orden}	%USs	%USi	mean (BCa C.I.)^{Orden}	%USs	%USi
AWSD	**0.667(0.655,0.679)**[1]	59.5	4.5	**0.135(0.126,0.145)**[1]	58.5	4.6
SUNDAY	0.649 (0.639,0.658) [2]	53.4	2.3	0.110 (0.103,0.117) [2]	49.5	2.3
PQFT	0.638 (0.630,0.646) [3]	50.4	1.6	0.085 (0.079,0.091) [5]	38.1	1.6
GBVSm	0.630 (0.617,0.641) [4]	50.3	6.9	0.096 (0.087,0.105) [3]	46.0	6.5
SEOD	0.628 (0.618,0.638) [5]	48.6	3.9	0.080 (0.072,0.087) [7]	37.6	4.3
SURP	0.610 (0.598,0.621) [6]	45.3	6.5	0.088 (0.079,0.097) [4]	43.3	6.7
CIOFM	0.600 (0.591,0.607) [7]	43.0	6.4	0.083 (0.076,0.090) [6]	38.8	5.2
CIORFM	0.583 (0.571,0.595) [8]	39.0	10.7	0.069 (0.059,0.078) [8]	38.9	10.7
DCOF	0.568 (0.558,0.578) [9]	35.0	10.9	0.053 (0.046,0.061) [9]	32.1	10.4
ESA-D	0.539 (0.531,0.548)[10]	19.9	7.1	0.041 (0.035,0.047)[10]	22.2	5.8
H50	0.807 (0.797,0.816)	92.9	0.0	0.257 (0.248,0.265)	92.4	0.0
CENTER	0.492 (0.478,0.507)	24.9	26.0	-0.006 (-0.018,0.006)	26.3	26.6
CHANCE	0.500 (0.497,0.503)	0.1	0.1	0.000 (-0.003,0.003)	0.4	0.4

Tabla 4.8: Medidas *s-AUC* y *s-NSS* de los diez modelos de saliencia y los tres de referencia con la BD DIEM (KF=0.96). Intevalos de confianza BCa (95 %) y los pct. de tiempo en los que las medidas exceden los umbrales del test de permutación ($\%USs$ y $\%USi$).

	Base de datos AE-UCFS					
	s-AUC			s-NSS		
Modelo	mean (BCa C.I.)^{Orden}	%USs	%USi	mean (BCa C.I.)^{Orden}	%USs	%USi
AWSD	**0.745(0.732,0.757)**[1]	46.9	0.1	**0.206(0.195,0.216)**[1]	47.7	0.2
SEOD	0.733 (0.720,0.745) [2]	46.5	0.5	0.168 (0.159,0.178) [4]	35.6	0.4
PQFT	0.731 (0.721,0.741) [3]	43.4	0.1	0.161 (0.153,0.168) [5]	30.9	0.2
SUNDAY	0.726 (0.714,0.738) [4]	41.2	0.1	0.182 (0.172,0.191) [2]	38.6	0.1
GBVSm	0.715 (0.700,0.728) [5]	42.4	0.6	0.175 (0.164,0.186) [3]	40.3	0.6
CIORFM	0.682 (0.667,0.697) [6]	32.2	0.8	0.157 (0.144,0.169) [6]	33.1	0.7
SURP	0.674 (0.660,0.687) [7]	33.4	0.9	0.147 (0.135,0.158) [7]	30.8	0.8
DCOF	0.661 (0.649,0.673) [8]	35.9	3.3	0.126 (0.117,0.136) [9]	31.2	3.6
CIOFM	0.656 (0.645,0.666) [9]	34.4	1.8	0.133 (0.124,0.142) [8]	25.3	1.5
ESA-D	0.580 (0.570,0.591)[10]	8.2	1.3	0.078 (0.070,0.087)[10]	7.3	1.1
H50	0.851 (0.840,0.860)	81.3	0.0	0.299 (0.291,0.307)	80.6	0.0
CENTER	0.502 (0.483,0.520)	10.7	10.1	0.002(-0.014,0.017)	12.6	11.6
CHANCE	0.499 (0.493,0.506)	0.0	0.0	-0.000(-0.006,0.005)	0.1	0.0

Tabla 4.9: Medidas *s-AUC* y *s-NSS* de los diez modelos de saliencia y los tres de referencia con la BD AE-UCFSA (KF=0.95). Intevalos de confianza BCa (95 %) y los pct. de tiempo en los que las medidas exceden los umbrales del test de permutación ($\%USs$ y $\%USi$).

	Base de datos ASCMN					
	s-AUC			s-NSS		
Modelo	mean (BCa C.I.)^{Orden}	%USs	%USi	mean (BCa C.I.)^{Orden}	%USs	%USi
AWSD	**0.662(0.650,0.673)**[1]	18.0	0.0	**0.136(0.126,0.145)**[1]	16.8	0.1
SUNDAY	0.652 (0.641,0.662) [2]	13.7	0.0	0.119 (0.111,0.128) [2]	9.9	0.0
SEOD	0.641 (0.630,0.651) [3]	14.2	0.2	0.102 (0.093,0.110) [3]	6.8	0.2
GBVSm	0.626 (0.614,0.639) [4]	11.2	0.2	0.099 (0.089,0.109) [4]	8.4	0.2
PQFT	0.618 (0.608,0.627) [5]	8.1	0.0	0.088 (0.080,0.096) [6]	3.3	0.1
SURP	0.611 (0.600,0.622) [6]	9.5	0.2	0.098 (0.088,0.107) [5]	8.5	0.3
CIORFM	0.593 (0.580,0.605) [7]	7.4	0.1	0.080 (0.070,0.091) [7]	7.8	0.1
DCOF	0.589 (0.579,0.599) [8]	5.6	0.9	0.072 (0.063,0.080) [9]	3.4	0.8
CIOFM	0.583 (0.575,0.591) [9]	10.6	1.8	0.077 (0.069,0.085) [8]	5.1	2.0
ESA-D	0.500 (0.491,0.510)[10]	0.4	0.5	0.008(-0.001,0.016)[10]	0.5	0.8
H50	0.835 (0.827,0.842)	57.3	0.0	0.296 (0.290,0.303)	59.6	0.0
CENTER	0.478 (0.463,0.492)	1.2	3.2	-0.019(-0.032,-0.007)	1.8	3.8
CHANCE	0.501 (0.493,0.509)	0.0	0.0	0.001(-0.006,0.008)	0.1	0.0

Tabla 4.10: Medidas *s-AUC* y *s-NSS* de los diez modelos de saliencia y los tres de referencia con la BD ASCMN (KF=0.98). Intevalos de confianza BCa (95 %) y los pct. de tiempo en los que las medidas exceden los umbrales del test de permutación ($\%USs$ y $\%USi$).

	Base de datos GC					
	s-AUC			s-NSS		
Modelo	mean (BCa C.I.)^{Orden}	%USs	%USi	mean (BCa C.I.)^{Orden}	%USs	%USi
AWSD	**0.716(0.706,0.726)**[1]	61.1	0.1	**0.177(0.169,0.185)**[1]	61.1	0.1
SEOD	0.689 (0.680,0.698) [2]	55.6	0.3	0.119 (0.112,0.126) [3]	41.3	0.2
SUNDAY	0.677 (0.668,0.685) [3]	51.1	0.2	0.136 (0.129,0.142) [2]	48.8	0.2
PQFT	0.665 (0.657,0.673) [4]	44.1	0.2	0.110 (0.104,0.115) [4]	31.1	0.2
GBVSm	0.630 (0.619,0.639) [5]	40.4	2.7	0.108 (0.100,0.116) [5]	40.2	2.3
CIORFM	0.614 (0.603,0.624) [6]	36.0	3.4	0.096 (0.087,0.105) [6]	36.4	3.0
DCOF	0.598 (0.590,0.607) [7]	29.3	1.9	0.076 (0.069,0.082) [8]	24.4	2.5
CIOFM	0.591 (0.584,0.598) [8]	29.3	2.8	0.089 (0.083,0.095) [7]	28.1	2.7
SURP	0.567 (0.558,0.576) [9]	25.7	4.6	0.064 (0.056,0.071) [9]	26.7	3.9
ESA-D	0.530 (0.521,0.538)[10]	7.3	2.8	0.026 (0.020,0.032)[10]	8.0	3.1
H50	0.829 (0.821,0.836)	95.4	0.0	0.269 (0.263,0.274)	94.7	0.0
CENTER	0.496 (0.484,0.508)	10.9	9.5	-0.004 (-0.014,0.006)	11.9	10.2
CHANCE	0.500 (0.496,0.505)	0.1	0.0	0.000 (-0.003,0.004)	0.1	0.0

Tabla 4.11: Medidas *s-AUC* y *s-NSS* de los diez modelos de saliencia y los tres de referencia con la BD GC (KF=0.99). Intevalos de confianza BCa (95 %) y los pct. de tiempo en los que las medidas exceden los umbrales del test de permutación ($\%USs$ y $\%USi$).

	Base de datos Holly2Train					
	s-AUC			s-NSS		
Test Model	mean (BCa C.I.)^{Orden}	%USs	%USi	mean (BCa C.I.)^{Orden}	%USs	%USi
AWSD	**0.685(0.670,0.700)**[1]	35.5	0.7	**0.155(0.143,0.168)**[1]	35.8	0.9
SUNDAY	0.657 (0.644,0.671) [2]	26.3	0.6	0.123 (0.112,0.133) [2]	23.5	0.7
PQFT	0.655 (0.642,0.666) [3]	25.1	0.4	0.102 (0.093,0.111) [3]	14.9	0.6
GBVSm	0.625 (0.609,0.641) [4]	24.8	2.2	0.100 (0.088,0.113) [4]	22.9	2.4
SEOD	0.622 (0.608,0.637) [5]	22.2	2.0	0.082 (0.070,0.093) [6]	15.0	2.0
SURP	0.612 (0.597,0.626) [6]	22.9	3.2	0.089 (0.077,0.101) [5]	19.9	3.1
CIOFM	0.598 (0.586,0.609) [7]	23.6	5.5	0.082 (0.071,0.092) [7]	17.6	4.5
DCOF	0.597 (0.583,0.611) [8]	21.3	3.7	0.078 (0.066,0.089) [8]	19.9	4.3
CIORFM	0.573 (0.556,0.590) [9]	17.3	4.0	0.063 (0.049,0.076) [9]	17.9	4.5
ESA-D	0.550 (0.537,0.563)[10]	9.6	2.5	0.044 (0.033,0.055)[10]	9.2	2.7
H50	0.862 (0.851,0.871)	84.2	0.0	0.315 (0.306,0.323)	84.5	0.0
CENTER	0.508 (0.490,0.527)	10.8	9.2	0.007 (-0.009,0.022)	12.7	10.8
CHANCE	0.500 (0.494,0.507)	0.0	0.0	0.000 (-0.005,0.006)	0.1	0.1

Tabla 4.12: Medidas *s-AUC* y *s-NSS* de los diez modelos de saliencia y los tres de referencia con la BD Holly2 Train (KF=0.99). Intevalos de confianza BCa (95 %) y los pct. de tiempo en los que las medidas exceden los umbrales del test de permutación ($\%USs$ y $\%USi$).

	Base de datos Holly2Test					
	s-AUC			s-NSS		
Test Model	mean (BCa C.I.)^{Orden}	%USs	%USi	mean (BCa C.I.)^{Orden}	%USs	%USi
AWSD	**0.670(0.655,0.685)**[1]	33.0	1.1	**0.143(0.131,0.155)**[1]	33.3	1.4
SUNDAY	0.649 (0.635,0.662) [2]	24.6	0.7	0.116 (0.105,0.126) [2]	21.8	0.8
PQFT	0.644 (0.632,0.656) [3]	22.7	0.5	0.093 (0.084,0.102) [3]	12.4	0.7
SURP	0.615 (0.599,0.629) [4]	23.1	2.7	0.092 (0.079,0.104) [4]	20.7	2.8
GBVSm	0.613 (0.597,0.629) [5]	23.3	2.6	0.092 (0.079,0.104) [5]	21.6	2.7
SEOD	0.607 (0.592,0.621) [6]	19.9	2.6	0.071 (0.060,0.082) [8]	13.7	2.5
DCOF	0.599 (0.585,0.613) [7]	21.7	3.4	0.080 (0.068,0.091) [6]	20.3	4.1
CIOFM	0.592 (0.580,0.603) [8]	22.7	5.6	0.075 (0.065,0.085) [7]	16.8	4.9
CIORFM	0.563 (0.546,0.580) [9]	16.3	4.5	0.054 (0.040,0.068) [9]	17.0	5.1
ESA-D	0.556 (0.542,0.569)[10]	10.5	2.1	0.049 (0.038,0.059)[10]	10.2	2.3
H50	0.865 (0.855,0.875)	85.6	0.0	0.318 (0.309,0.326)	85.4	0.0
CENTER	0.508 (0.490,0.527)	11.8	10.3	0.006 (-0.009,0.022)	13.8	11.9
CHANCE	0.500 (0.494,0.507)	0.0	0.0	0.000 (-0.005,0.006)	0.1	0.1

Tabla 4.13: Medidas *s-AUC* y *s-NSS* de los diez modelos de saliencia y los tres de referencia con la BD Holly2 Test (KF=0.9). Intevalos de confianza BCa (95 %) y los pct. de tiempo en los que las medidas exceden los umbrales del test de permutación ($\%USs$ y $\%USi$).

4.2. Reproducción de efectos *pop-out*

Así como en la anterior sección se analizan los resultados de modo global cuantitativa y cualitativamente, sin entrar a valorar el comportamiento ante situaciones puntuales, en esta sección se intentará ver a los modelos durante su funcionamiento con ejemplos reales y se intentará dar explicación a ciertos efectos de *pop-out* relacionados con el movimiento [Ros99]. En imagen estática, *pop-out* se refiere a una región del espacio que presenta una característica diferencial, que lo hace fácilmente identificable dentro de la escena. En vídeo el efecto *pop-out* se produce cuando la característica diferencial es el movimiento o el parpadeo; esto no excluye la presencia de fenómenos de *pop-out* asociados a características estáticas.

4.2.1. Efectos *pop-out* en escenas naturales

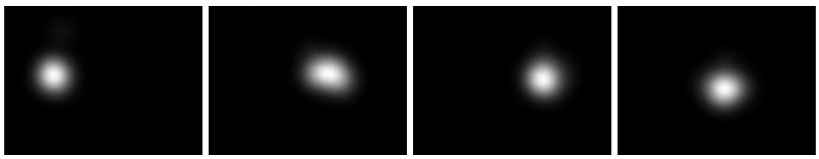
Para mostrar los vídeos de saliencia de los modelos evaluados se ha empleado una representación tridimensional como la mostrada en la figura 3.23 de la sección 3.2.2. Se ha añadido además falsocolor a las imágenes 3D, puesto que permite resaltar más los detalles. A continuación se muestran varios ejemplos de los modelos evaluados aplicados sobre escenas de las categorías RCE y RCD, escenas naturales con cámara estática y dinámica.

Vídeo RCD-Eagle-1

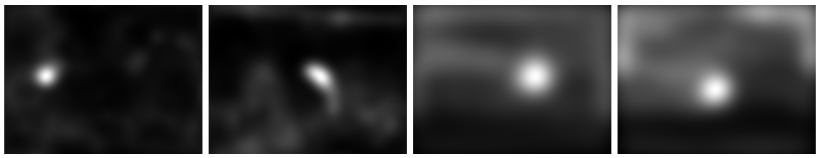
La figura 4.11 muestra el comportamiento del modelo AWSD en una situación en la que se producen movimientos simultáneos de la cámara y del objeto perseguido. El mapa de saliencia del modelo AWSD, figura 4.11c, y el mapa de atención de los sujetos, figura 4.11b, son semejantes, debido a que la atención es atraída por el objeto perseguido de modo inmediato. El fondo en este caso actúa como distractor, pero no evita el correcto funcionamiento del AWSD, incluso en instantes en los cuales el objeto perseguido llega a estar casi totalmente ocluido por los objetos del fondo (figura 4.11, fotograma 128).

En la figura 4.12 se muestra la composición espacio-temporal para el mapa de atención obtenido directamente de los datos de las fijaciones de los sujetos, figura 4.12a, comparada con el resultado de los dos modelos que mejor valoración alcanzan (según la $s\text{-}AUC$); el modelo AWSD, figura 4.12b y el GBVSm 4.12c. El nivel de ruido que presentan los mapas de los modelos de saliencia AWSD y GBVSm aumentan de modo significativo en los instantes en los que el objeto perseguido se entremezcla con los árboles, fotogramas 20 a 40 y al final del vídeo a partir del fotograma 130; siendo menor en los periodos de tiempo en los que el objeto está claramente visible.

(a) Fotogramas 10, 26, 128 y 140.

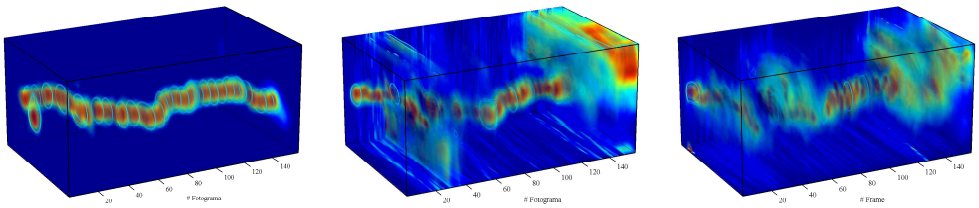


(b) Mapa de atención de los sujetos.



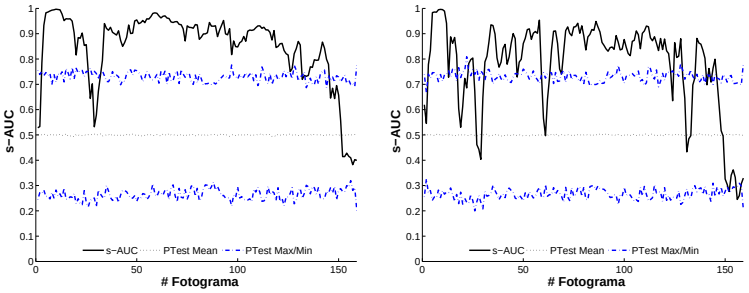
(c) Mapa de saliencia del modelo AWSD.

Figura 4.11: Fotogramas del vídeo RCD-Eagle-1, (a) vídeo original, (b) el mapa de atención de los sujetos y (c) el mapa del modelo AWSD.



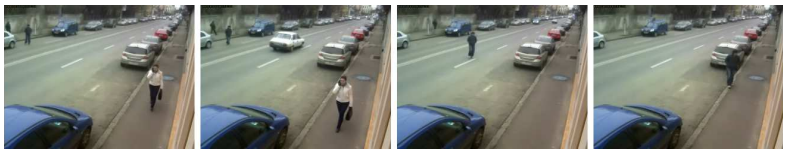
(a) Mapa de atención (b) Modelo AWSD (c) Modelo GBVSm

Figura 4.12: Composición tridimensional del vídeo RCD-Eagle-1. (a) mapa de atención y mapa de saliencia del modelo (b) AWSD y (c) GBVSm.

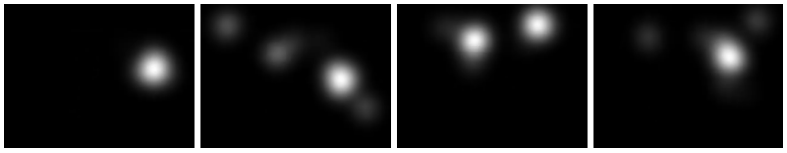


(a) AWSD (b) GBVSm

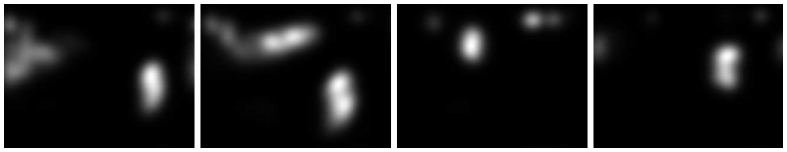
Figura 4.13: Evolución temporal de la *s-AUC* para el modelo (a) AWSD y (b) GBVSm para el vídeo RCD-Eagle-1.



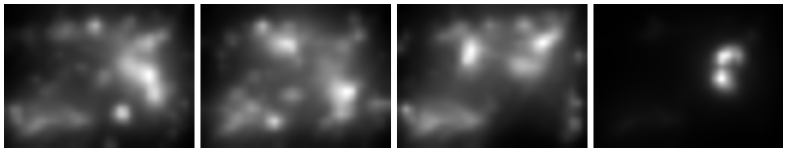
(a) Fotogramas 6, 23, 85, 151.



(b) Mapa de atención de los sujetos.



(c) Mapa de saliencia del AWSD.



(d) Mapa de saliencia del GBVSm.

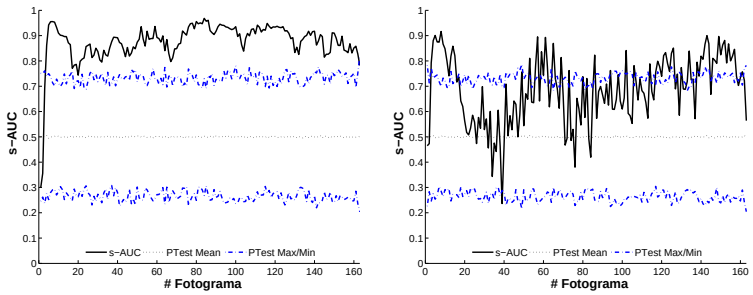
Figura 4.14: Fotogramas del vídeo RCE-Traffic-2, (a) vídeo original, (b) el mapa de atención y los mapas del (c) AWSD y (d) GBVSm.

Este mismo efecto se puede apreciar directamente a partir del seguimiento de los valores de la métrica *s-AUC* mostrada en las figuras 4.13a y 4.13b, dando lugar a unos valores promedio *s-AUC* de 0.85 y 0.78 respectivamente.

Con este sencillo ejemplo se pretende mostrar la gran importancia de crear una métrica ($s\text{-}AUC_t$), que devuelva valores para cada instante de tiempo, y no solo una valoración global. Mediante la simple inspección del comportamiento instantáneo de la métrica podemos discernir los intervalos temporales de mayor dificultad para cualquier modelo evaluado.

Vídeo RCE-Traffic-2

Con este ejemplo de la categoría de vídeos con fondo estático se pretende comparar el comportamiento del modelo AWSD ($s\text{-}AUC$=0.88), frente a uno de sus principales competidores, el GBVSm ($s\text{-}AUC$=0.69). En la figura 4.14 se muestran varios fotogramas en los que el modelo AWSD muestra un comportamiento acorde a los humanos. En este caso son varios los objetos que compiten por la saliencia y a diferentes escalas; la persona y el coche



(a) AWSD (b) GBVSm

Figura 4.15: Evolución temporal de la *s-AUC* para el modelo (a) AWSD y (b) GBVSm para el vídeo RCE-Traffic-2.

que están cerca, frente a una persona distante. Para este caso, tanto la componente dinámica, como el análisis multiescala ayudan al modelo AWSD a alcanzar estos buenos resultados.

4.2.2. Efectos *pop-out* con vídeos sintéticos

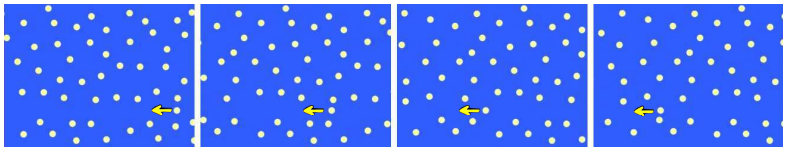
En los fotogramas de los vídeos que se mostrarán a continuación, pertenecientes a vídeos sintéticos de las categorías SCD y SCE, se puede observar cómo diferentes características pre-atencionales como el color, la forma o el movimiento dan lugar a diversos efectos de *pop-out* [Not93, WH04].

Vídeo SCD-circulos-1sentido-contrario

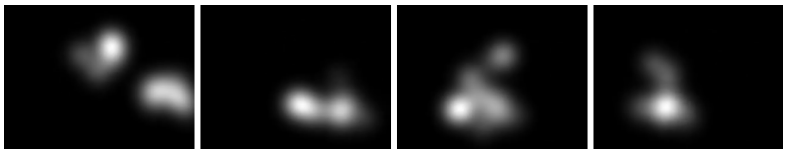
En este vídeo se muestra una serie de objetos de forma circular desplazándose hacia la derecha y uno desplazándose en sentido contrario. En la figura 4.16b se aprecia como las fijaciones de los humanos persiguen a este objeto a lo largo de la pantalla. El elevado número de distractores interfiere en la identificación del objeto saliente y desde los primeros instantes el nivel de ruido generado por el modelo AWSD es mayor que en otros casos.

El valor de *s-AUC* para el modelo AWSD es el mayor de los modelos comparados, $s\text{-}AUC = 0.714$ con un %USs del 49 %, frente a valores de *s-AUC* que oscilan entre 0,61 y 0,32 y porcentajes con valores entre el 0 % y el 18.6 % para el resto de los modelos. El H50 alcanza un 0.89 y un 94 % de %USs.

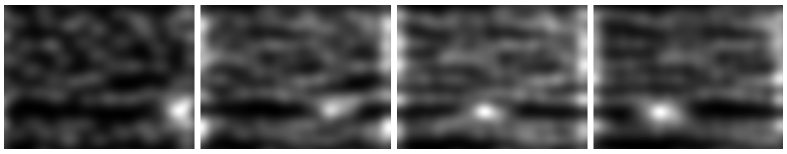
Examinando los mapas de atención de la figura 4.16b se puede ver que, mientras algunos sujetos mantienen la mirada en el centro de la escena hasta el fotograma 25, el modelo AWSD, ya en los primeros fotogramas, detecta la presencia de un objeto extraño tan pronto como éste aparece por la parte derecha de la imagen desplazándose en sentido contrario, figura 4.16c. Esta



(a) Fotogramas 21, 38, 52 y 59.



(b) Mapa de atención de los sujetos.



(c) Mapa de saliencia del AWSD.

Figura 4.16: Fotogramas del vídeo SCD-circulos-1sentido-contrario, (a) un objeto se mueve en sentido contrario al resto de los objetos, (b) el mapa de atención de los sujetos y (c) el mapa de saliencia del modelo AWSD.

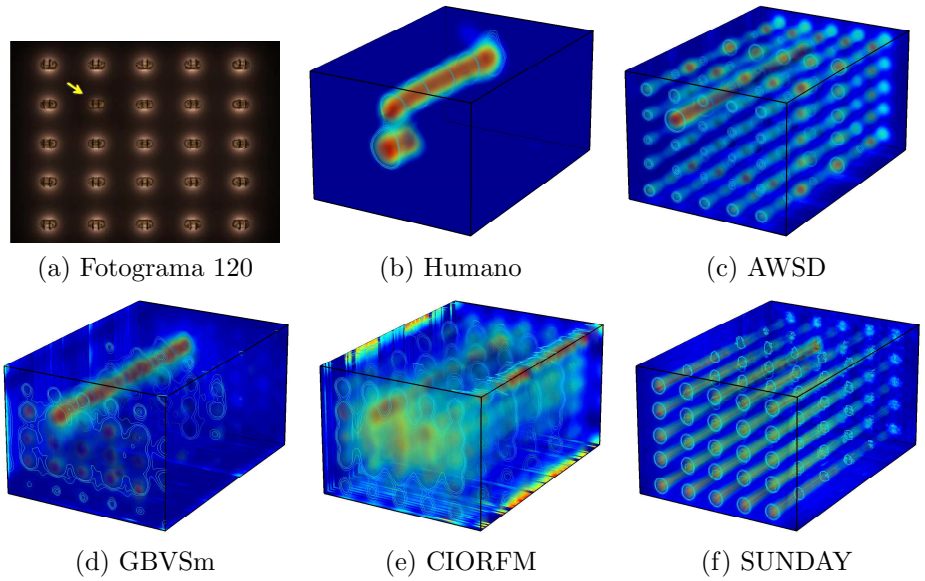
rápida respuesta es debida a la capacidad del modelo de identificar movimiento de modo eficiente.

Vídeo SCE-Lámparas

Este ejemplo muestra el efecto de la presencia de una sola característica diferencial (el parpadeo) en el fenómeno de *pop-out*, cuando los objetos replicados son elementos complejos. Se presenta un conjunto de lámparas, figura 4.17 y se produce un cambio de luminosidad en una de ellas.

Transcurren escasos segundos hasta que los sujetos dejan de mirar al centro e identifican el objeto con la característica diferencial, figura 4.17b. A partir de ese instante, los sujetos mantienen la atención en el objeto con la característica diferencial.

El modelo AWSD, figura 4.17c reproduce el comportamiento anterior con un valor de $s\text{-}AUC$ de 0.84, y el porcentaje de tiempo que se superan el USs es del 83,3 %. En este caso la contribución de la componente estática se aprecia como unos pequeños circulos salientes situados sobre las posiciones de las demás lámparas. En las figuras (d), (e) y (f) se presentan los tres modelos mejor posicionados de los demás evaluados, el modelo GBVS muestra claramente una buena respuesta en este caso, en los otros dos la identificación del objeto saliente no es evidente.



(a) Fotograma 120 (b) Humano (c) AWSD

(d) GBVSm (e) CIORFM (f) SUNDAY

Figura 4.17: (a) Fotograma 20 del vídeo SCE-Lámparas, (b) mapa Humano y mapas de los modelos (c) AWSD, (d) GBVSm, (e) CIORFM y (f) SUNDAY.

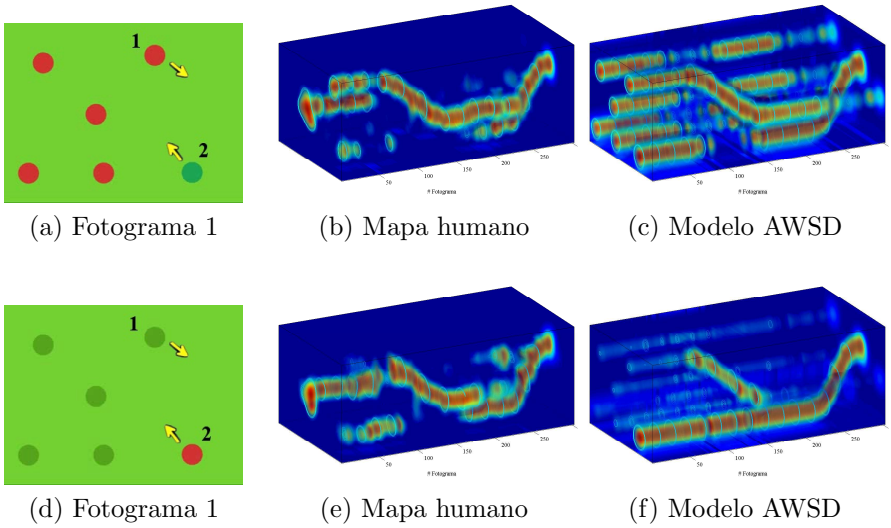
Competición de características, color y movimiento

Tal como describen Joseph et al. [JCN97], la presencia de múltiples características puede afectar al rendimiento del sistema visual durante la identificación de los objetos de mayor saliencia de una escena. Para mostrar la interferencia del movimiento sobre el color se presentan los siguientes vídeos en los que se combina saliencia debida al color y al movimiento.

Vídeos SCE-Bur1DistractoresAltos/Bajos

Los fotogramas de los dos vídeos de ejemplo que se muestran en la figura 4.18 presentan el efecto de asimetría de color y la influencia del movimiento sobre este efecto. El vídeo de la figura 4.18a muestra una serie de círculos rojos (distractores) frente a un círculo verde que atrae la atención, pero justo al inicio del vídeo el circulo superior rojo comienza a moverse, compitiendo por la atención. En la figura 4.18d sucede lo mismo pero ahora es el circulo verde el que inicia la marcha. Los humanos persiguen claramente a ambos objetos cuando se inicia el movimiento, como se puede apreciar en las figuras 4.18b y 4.18e.

De los resultados mostrados en las figura 4.19 se puede intuir que el movimiento de los objetos presenta mayor peso que el color en este vídeo de ejemplo. La posible explicación de este efecto es que el movimiento en este



(a) Fotograma 1 (b) Mapa humano (c) Modelo AWSD

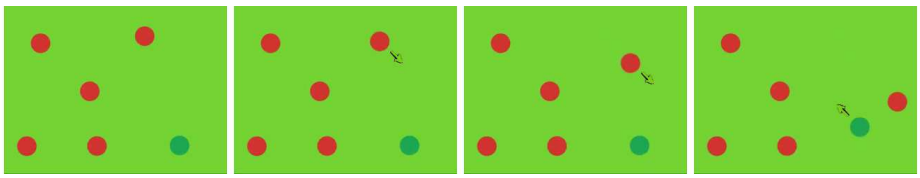
(d) Fotograma 1 (e) Mapa humano (f) Modelo AWSD

Figura 4.18: Fotogramas de los vídeos SCE-Bur1Distractores (a) -Altos y (d) -Bajos. (b) y (e), mapas de atención humanos. (c) y (f), mapas del AWSD.

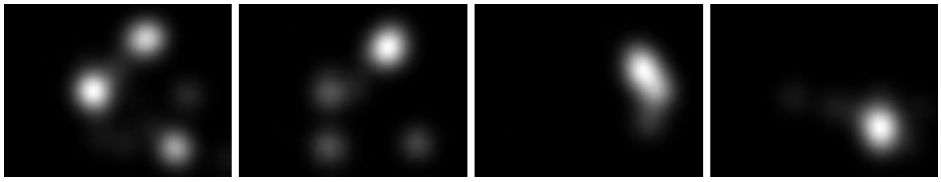
caso es una característica que surge a partir de un determinado instante. Este cambio provoca que la atención del sujeto que ya ha revisado la escena se dirija ahora hacia lo "novedoso"de la escena, que es el movimiento. Ese proceso de revisión se aprecia en la parte inferior izquierda de las figuras 4.18b y 4.18e, en las que se dproducen varias fijaciones sobre el objeto diferente, antes de que comience a moverse.

Respecto al efecto inicial de asimetría de color, igual que sucedía con el modelo estático, se puede apreciar que el mapa de saliencia del modelo AWSD es diferente en el caso del círculo rojo sobre círculos verdes, figuras 4.18c, que en el caso opuesto, figura 4.18f. Esto mismo se representa en las figuras 4.19c y 4.19f para varios fotogramas elegidos manualmente durante los períodos de movimiento de los objetos, región en la que el modelo AWSD y el mapa humano son similares (fotogramas 80 a 140 para el círculo rojo y 230 hasta el final para el círculo verde).

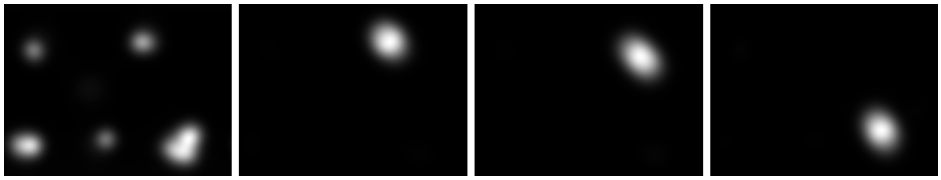
La saliencia debida al color compite con el movimiento durante toda la duración del vídeo. Desde el punto de vista de la implementación del modelo AWSD, el comportamiento es fácilmente explicable ya que la elección de un esquema de integración prioritario lleva, en el caso extremo de imágenes estáticas, al resultado devuelto por la componente estática [GDLFVP12].



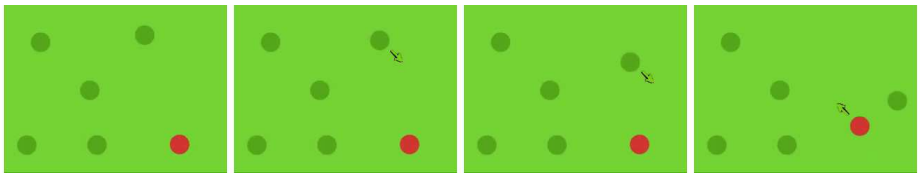
(a) Fotogramas 10,90,110,250 del vídeo SCE-Bur1DistractoresAltos.



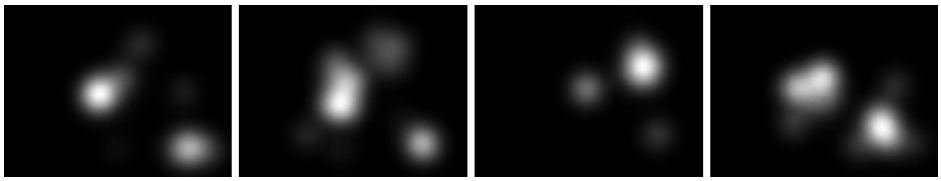
(b) Mapa de atención de los sujetos (SCE-Bur1DistractoresAltos).



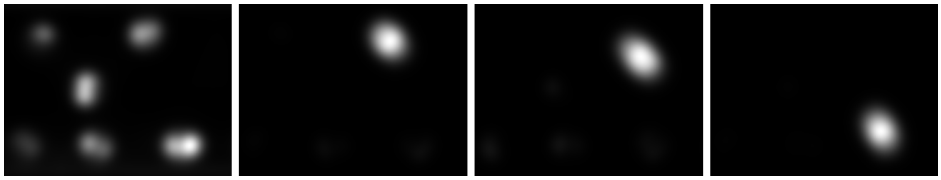
(c) Mapa del modelo AWSD (SCE-Bur1DistractoresAltos).



(d) Fotogramas 10,90,110,250 del vídeo SCE-Bur1DistractoresBajos.

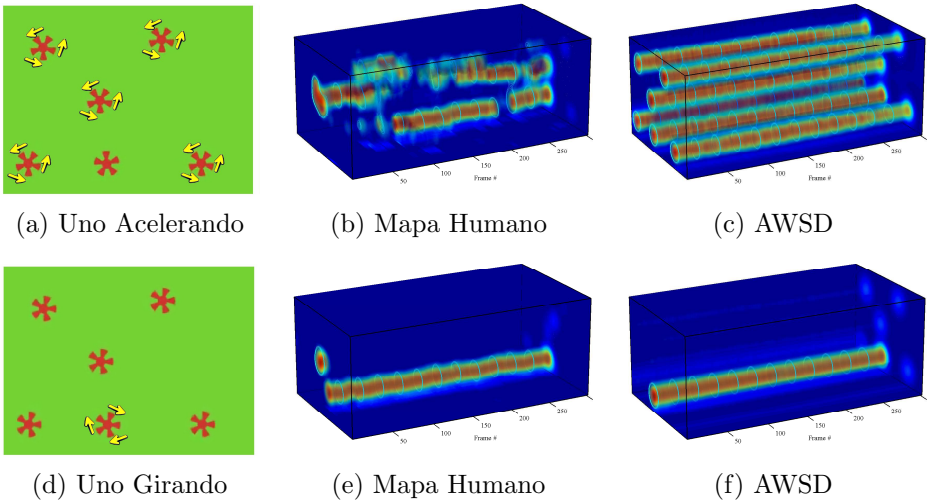


(e) Mapa de atención de los sujetos (SCE-Bur1DistractoresBajos).



(f) Mapa del modelo AWSD (SCE-Bur1DistractoresBajos).

Figura 4.19: Fotogramas de los vídeos (a) SCE-Bur1DistractoresAltos y (d) SCE-Bur1DistractoresBajos, varios objetos de diferentes colores se desplazan por la pantalla (a),(d) y su correspondiente mapa de atención (b),(e) obtenido con las fijaciones de los humanos y (c),(f) con el modelo AWSD.



(a) Uno Acelerando (b) Mapa Humano (c) AWSD

(d) Uno Girando (e) Mapa Humano (f) AWSD

Figura 4.20: Fotogramas explicativos de los vídeos (a) SCE-UnoAcelerando y (d) SCE-UnoGirando. (b) y (e) mapas de atención humano. (c) y (f) respuesta del modelo AWSD.

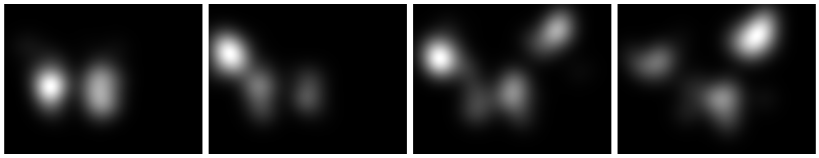
4.2.3. Presencia y ausencia de movimiento.

Los fotogramas de la figura 4.20 se corresponden con un grupo de vídeos en los que se muestran varios objetos del mismo color realizando diferentes movimientos de rotación sobre si mismos. Al inicio de los vídeos SCE-UnoAcelerando y SCE-UnoGirando los sujetos miran al centro (sesgo central), primeros fotogramas de la figura 4.20b y 4.20e. A continuación la asimetría de las respuestas de los sujetos se ve reflejada en la diferencia entre la figura 4.20c en la que la atención se divide entre los distintos objetos en movimiento mientras que la figura 4.20e muestra como desde el inicio los sujetos fijan la atención en el objeto en movimiento y la mantienen hasta el final.

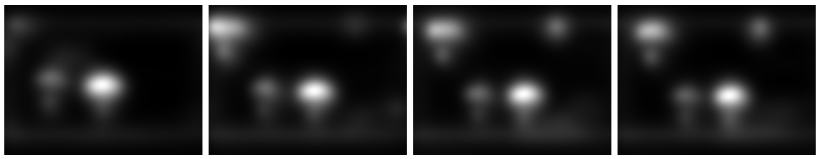
Mediante estos vídeos se puede evaluar el comportamiento del modelo AWSD en presencia/ausencia de movimiento [Wol01]. Comparando el vídeo en el que todos los objetos están quietos (ausencia de movimiento) menos uno de ellos, figura 4.20d, frente al caso opuesto en el que todos giran (presencia de movimiento) salvo uno que está quieto, figura 4.20a, se puede ver que el mapa de atención humano no es simétrico ni tampoco lo es el mapa de saliencia del modelo AWSD [Ros99].

El modelo AWSD replica este comportamiento asimétrico con la diferencia de que para el caso en el que todos los objetos giran, figura 4.20c, el modelo no se decanta por ninguno en particular, dando igual peso a todos los elementos

(a) Fotogramas 3220, 3240, 3255 y 3270.



(b) Mapa de atención de los sujetos.



(c) Mapa de saliencia del modelo AWSD.

Figura 4.21: (a) Fotogramas de una secuencia dinámica de la BD DIEM, en la que un cámara en movimiento persigue varias personas haciendo parapente, (b) el mapa obtenido con las fijaciones y (c) con el modelo AWSD.

presentes en el vídeo, incluso al objeto que está quieto.

Basándose en la teoría de integración de características de Treisman [TG88] y de acuerdo con la afirmación de Wolfe de que el movimiento es una de las características preatencionales [WH04], la interpretación de este resultado podría ser la siguiente; en presencia de movimiento, se activa el mapa correspondiente a esa ubicación sin necesidad de realizar por ello un proceso de revisión secuencial. En el caso contrario, la ausencia de movimiento no activa ningún mapa, por lo que se hace necesario detectar esa ausencia mediante un proceso de revisión secuencial mucho más lento y dependiente del número de distractores presentes en la escena.

Conclusiones

En este trabajo se propone un modelo de atención visual selectiva, puramente *bottom-up*, capaz de detectar saliencia tanto en imágenes estáticas como en vídeo. La idea básica sobre la que se sustenta el modelo AWSD es que la saliencia se produce en aquellos puntos donde la energía local espacio-temporal posee la máxima variabilidad respecto de la distribución media de esta característica en un espacio multiescala.

El mecanismo base subyacente es el blanqueado adaptativo que reduce las redundancias de segundo orden de la cadena de procesado y permite una adaptación a las estadísticas de las imágenes entrantes. Los buenos resultados obtenidos por el AWSD nos permiten afirmar que los mecanismos subyacentes en el modelo son susceptibles de convertirse en criterios estándar para el diseño de sistemas de atención *bottom-up*.

La capacidad de predicción de fijaciones y la respuesta ante experimentos *pop-out* dinámicos del AWS-D demuestran la superioridad con respecto resto de modelos evaluados. Las medidas empleadas han sido combinadas con un test de fiabilidad que nos permite analizar la significación temporal de los modelos ($\%USs$) y la calidad del conjunto de fijaciones de las bases de datos (ensanchamiento de los umbrales USs y USi en la base CNRCS). La aplicación de la técnica del shuffling a la medida NSS consigue compensar el sesgo central y eliminar así uno de sus principales handicaps. Se ha comprobado que el uso conjunto de las medidas $s\text{-}AUC$ y $s\text{-}NSS$ permite identificar de modo rápido situaciones ente las cuales los modelos no se comportan del modo esperado.

Además se ha construido la base de datos CITIUS, que incluye estímulos sintéticos *pop-out* dinámicos, un número elevado número de fijaciones en los fotogramas de cada vídeo y los vídeos se han elegido para minimizar sesgos *top-down* de los observadores dejando fuera entornos de interiores, deportes de masas, etc. Contiene gran cantidad de información que permite valorar el comportamiento de los modelos ante la presencia de múltiples características espacio-temporales simultáneas: color, forma, movimiento, variación de la iluminación, cambios en las características del fondo, incluyendo tanto escenas

naturales como sintéticas. Estas propiedades complementan a la mayoría de las bases de datos públicas existentes para la comparación eficiente en vídeo.

Los esfuerzos futuros se concentrarán en tres ámbitos, por un lado, mejorar la eficiencia computacional en base al rediseño y a la paralelización de las etapas especialmente exigentes con los recursos de cálculo y memoria (p.e. descomposición multiescala 2D y 3D), por otro, lograr una atención dirigida a proto-objetos mediante la inclusión de etapas competitivas que potencien los mecanismos del base-grouping que el AWSD contiene implícitos en su concepción (atribuidos al proceso adaptativo del blanqueado [RMvdH+14, GDLFVP12]), y el tercero sería la incorporación de información *top-down*, ya sea mediante detectores de caras o de personas.

Apéndice A

Blanqueado: Definición formal

A continuación se detalla la estrategia de blanqueado que se aplicará sobre las diferentes características espacio-temporales a lo largo de las diferentes etapas del modelo descrito en esta tesis.
Definición: Dado un vector aleatorio y centrado se dice que sus componentes están blanqueadas (*whitened*) si sus elementos están decorrelacionados y poseen varianza unidad.
Metodología: El blanqueado de los datos se realiza mediante la estrategia habitual que consiste en la decorrelación mediante PCA, seguida de un escalado que iguala las magnitudes de los autovalores [Bis06].
Formalización: Sea $\mathbf{X} = \{\mathbf{x}_1 \cdots \mathbf{x}_j \cdots \mathbf{x}_n\}$ la matriz de datos original, constituida por n vectores columna centrados ($\bar{\mathbf{x}}_j = 0$) de m elementos, que son los valores de los píxeles de la imagen, ($\mathbf{x}_j^T = (x_1 \cdots x_i \cdots x_m)$).

Partiendo de la siguiente ecuación se define la transformación de blanqueado sobre la matriz $\mathbf{X}$:

$$\mathbf{Z} = \mathbf{X}\Phi\Lambda^{-1/2} \tag{A.1}$$

donde $\mathbf{Z}$ es la representación correspondiente a la matriz de datos blanqueada, de tamaño $m \times n$. Las matrices Φ y Λ son las matrices de los autovectores y autovalores, respectivamente, de la matriz de covarianza $\Sigma = E\{\mathbf{X}^T\mathbf{X}\}$. Las matrices anteriores tienen rango $(n \times n)$ y cumplen la condición:

$$\Sigma\Phi = \Phi\Lambda \tag{A.2}$$

Por lo tanto, la matriz Φ diagonaliza a la matriz de covarianza Σ. Y la matriz Λ se comporta como un factor de escala que da lugar a una matriz de covarianza que es la matriz identidad.

Bibliografía

[ADF10] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 73–80, 2010.

[AHES09] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009, IEEE Conference on*, pages 1597–1604, 2009.

[Ahm92] Subutai Ahmad. *Visit: an efficient computational model of human visual attention.* PhD thesis, Champaign, IL, USA, 1992.

[AL06] Tamar Avraham and Michael Lindenbaum. Esaliency-a stochastic attention model incorporating similarity information and knowledge-based preferences. In *Proc. Int'l Workshop Representation and Use of Prior Knowledge in Vision, with European Conf. Computer Vision*, 2006.

[AM09] Muhammad Aziz and Bärbel Mertsching. Towards standardization of evaluation metrics and methods for visual attention models. In Lucas Paletta and John Tsotsos, editors, *Attention in Cognitive Systems*, volume 5395 of *Lecture Notes in Computer Science*, pages 227–241. Springer Berlin / Heidelberg, 2009.

[AMK89] Richard A. Abrams, David E. Meyer, and Sylvan Kornblum. Speed and accuracy of saccadic eye movements: Characteristics of impulse variability in the oculomotor system. *Journal of Experimental Psychology: Human Perception and Performance*, 15(3):529–543, 1989.

[AO91] Subutai Ahmad and Stephen Omohundro. Efficient visual search: A connectionist solution. In *Proceedings of the 13th*

Annual Conference of the Cognitive Science Society. Hillsdale, NJ: Erlbaum, 1991.

[AR92] Joseph J. Atick and A. Norman Redlich. What does the retina know about natural scenes? *Neural Computation*, 4(2):196–210, 1992.

[Att54] Fred Attneave. Some informational aspects of visual perception. *Psychological review*, 61(3):183, 1954.

[Bar61] H. B. Barlow. *Possible principles underlying the transformations of sensory messages*. MIT Press, Cambridge, MA, 1961.

[BAS75] A T Bahill, D Adler, and L Stark. Most naturally occurring human saccades have magnitudes of 15 degrees or less. *Investigative Ophthalmology and Visual Science*, 14(6):468–9, 1975.

[BBM+09] David J. Berg, Susan E. Boehnke, Robert A. Marino, Douglas P. Munoz, and Laurent Itti. Free viewing of dynamic stimuli by humans and monkeys. *Journal of Vision*, 9(5):1–15, 2009.

[BCDH10] A. Benoit, A. Caplier, B. Durette, and J. Herault. Using human visual system modeling for bio-inspired low level image processing. *Computer Vision and Image Understanding*, 114(7):758– 773, 2010.

[BEU96] Kjell Brunnström, Jan-Olof Eklundh, and Tomas Uhlin. Active fixation for scene exploration. *International Journal of Computer Vision*, 17:137–162, 1996.

[BFGP12] Sebastiano Battiato, Giovanni Maria Farinella, Nicolò Grippaldi, and Giovanni Puglisi. Content-based image resizing on mobile devices. In *VISAPP (2)*, pages 87–90, 2012.

[BHBH92] M. Barth, D. Hirayama, G. Beni, and S. Hackwood. A color vision inspection system for integrated circuit manufacturing. *Semiconductor Manufacturing, IEEE Transactions on*, 5(4):290–301, 1992.

[BI13] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):185–207, 2013.

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[BJM98] Maik Bollmann, Christoph Justkowski, and Bärbel Mertsching. Utilizing color information for the gaze control of an active vision system, 1998.

[BMB01] G. Backer, B. Mertsching, and M. Bollmann. Data- and model-driven gaze control for an active-vision system. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(12):1415–1429, 2001.

[BMHC08] Luke Barrington, T.K. Marks, J.H. Hsiao, and G.W. Cottrell. Nimble: A kernel density model of saccade-based visual memory. *Journal of Vision*, 8(14):1–14, 2008.

[BPE+01] M.S. Beauchamp, L. Petit, T.M. Ellmore, J. Ingeholm, and J.V. Haxby. A parametric fmri study of overt and covert shifts of visuospatial attention. *Neuroimage*, 14(2):310–321, 2001.

[BPH93] William Beaudot, Patricia Palagi, and Jeanny Hérault. Realistic simulation tool for early visual processing including space, time and colour data. In José Mira, Joan Cabestany, and Alberto Prieto, editors, *New Trends in Neural Computation*, volume 686 of *Lecture Notes in Computer Science*, pages 370–375. Springer Berlin / Heidelberg, 1993.

[BSI11] Ali Borji, Dicky Sihite, and Laurent Itti. Computational modeling of top-down visual attention in interactive environments. In *Proceedings of the British Machine Vision Conference*, pages 85.1–85.12. BMVA Press, 2011.

[BSI12] A. Borji, D.N. Sihite, and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *Image Processing, IEEE Transactions on*, PP(99):1, 2012.

[BT06] Neil D. B. Bruce and John K. Tsotsos. Saliency based on information maximization. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 155–162, MIT Press, 2006. MIT Press.

[BT09] Neil D. B. Bruce and John K. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3):1–24, 2009.

[Bur09] Alexandre Bur. *Computer models of dynamic visual attention*. PhD thesis, Universitè de Neuchâtel, 2009.

[Bus35] G. T. Buswell. *How People Look at Pictures*. University of Chicago Press, Chicago, 1935.

[BWMH07a] A. Bur, P. Wurtz, RM Miiri, and H. Hugli. Dynamic visual attention: competitive versus motion priority scheme. In *Proc. Int'l Conf. Computer Vision Systems*, 2007.

[BWMH07b] Alexandre Bur, Pascal Wurtz, Rene M. Muri, and Heinz Hügli. Motion integration in visual attention models for predicting simple dynamic scenes. In *Proceedings of the IS&T/SPIE 19th Annual Symposium on Electronic Imaging SPIE*, volume 6492-47, 2007.

[CAJT07] Enric Celaya, Jose-Luis Albarral, Pablo Jiménez, and Carme Torras. Natural landmark detection for visually-guided robot navigation. In Roberto Basili and Maria Pazienza, editors, *AI*IA 2007: Artificial Intelligence and Human-Oriented Computing*, volume 4733 of *Lecture Notes in Computer Science*, pages 555–566. Springer Berlin / Heidelberg, 2007.

[Cav99] Kyle R. Cave. The featuregate model of visual selection. *Psychological Research*, 62:182–194, 1999.

[CB99] Kyle Cave and Narcisse Bichot. Visuospatial attention: Beyond a spotlight model. *Psychonomic Bulletin and Review*, 6:204–223, 1999.

[CBN04] Martin Clauss, Pierre Bayerl, and Heiko Neumann. A statistical measure for evaluating regions-of-interest based attention algorithms. In CarlEdward Rasmussen, HeinrichH. Bülthoff, Bernhard Schölkopf, and MartinA. Giese, editors, *Pattern Recognition*, volume 3175 of *Lecture Notes in Computer Science*, pages 383–390. Springer Berlin Heidelberg, 2004.

[CF88] J. J. Clark and N. J. Ferrier. Modal Control Of An Attentive Vision System. In *Proceedings of the Second International Conference on Computer Vision*, pages 514–523, 1988.

[CGTB11] Marvin M. Chun, Julie D. Golomb, and Nicholas B. Turk-Browne. A taxonomy of external and internal attention. *Annual Review of Psychology*, 62(1):73–101, 2011.

[Cha91] D Chapman. *Vision, Instruction, and Action*. PhD thesis, 1991.

[CI06a] R. Carmi and L. Itti. The role of memory in guiding attention during natural vision. *Journal of Vision*, 6(9):898–914, 2006.

[CI06b] R. Carmi and L. Itti. Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research*, 46(26):4333–4345, 2006.

[Cri84] F Crick. Function of the thalamic reticular complex: the searchlight hypothesis. *Proceedings of the National Academy of Sciences*, 81(14):4586–4590, 1984.

[CS85] Koch C and Ullman S. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4(4):219–27, 1985.

[CS01] Christos Constantinidis and Michael A. Steinmetz. Neuronal responses in area 7a to multiple-stimulus displays: I. neurons encode the location of the salient stimulus. *Cerebral Cortex*, 11(7):581–591, 2001.

[CWS+07] Colin W.G. Clifford, Michael A. Webster, Garrett B. Stanley, Alan A. Stocker, Adam Kohn, Tatyana O. Sharpee, and Odelia Schwartz. Visual adaptation: Neural, psychological and computational aspects. *Vision Research*, 47(25):3125–3131, 2007.

[DE03] Parkhurst D.J. and Niebur E. Scene content selected by active vision. *Spatial Vision*, 16(2):125–154, 2003.

[DFVP08] Raquel Dosil, Xosé R. Fdez-Vidal, and Xosé M. Pardo. Motion representation using composite energy features. *Pattern Recognition*, 41(3):1110 – 1123, 2008.

[DL14] Eizaburo Doi and Michael S. Lewicki. A simple model of optimal population coding for sensory systems. *PLoS Comput Biol*, 10(8):e1003761, 2014.

[DLF+14] Lu Dong, Weisi Lin, Yuming Fang, Shiqian Wu, and Hock Soon Seah. Saliency detection in computer rendered images based on object-level contrast. *Journal of Visual Communication and Image Representation*, 25(3):525 – 533, 2014.

[DMGB10] Michael Dorr, Thomas Martinetz, Karl R. Gegenfurtner, and Erhardt Barth. Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision*, 10(10), 2010.

[Dor10] Michael Dorr. *Computational models and systems for gaze guidance*. Dr.-ing. thesis, 2010.

[DPFV06] Raquel Dosil, Xosé M. Pardo, and Xosé R. Fdez-Vidal. Data-driven synthesis of composite-feature detectors for 3d image analysis. *Image and Vision Computing*, 24(3):225 – 238, 2006.

[DPZ02] Gustavo Deco, Olga Pollatos, and Josef Zihl. The time course of selective visual attention: theory and experiments. *Vision Research*, 42(27):2925–2945, 2002.

[DS02] Doug DeCarlo and Anthony Santella. Stylization and abstraction of photographs. *ACM Trans. Graph.*, 21(3):769–776, 2002.

[DWTS90] R. Desimone, M. Wessinger, L. Thomas, and W. Schneider. Attentional control of visual perception: Cortical and subcortical mechanisms. *Cold Spring Harbor Symposia on Quantitative Biology*, 55:963–971, 1990.

[DZ01] Gustavo Deco and Josef Zihl. A neurodynamical model of visual attention: Feedback enhancement of spatial resolution in a hierarchical system. *Journal of Computational Neuroscience*, 10:231–253, 2001.

[EI08] Lior Elazary and Laurent Itti. Interesting objects are visually salient. *Journal of Vision*, 8(3), 2008.

[EY85] C. W. Eriksen and Y. Y. Yeh. Allocation of attention in the visual field. *Journal of experimental psychology. Human perception and performance*, 11(5):583–597, 1985.

[FBR05] Simone Frintrop, Gerriet Backer, and Erich Rome. Goal-directed search with a top-down modulated computational attention system. In WalterG. Kropatsch, Robert Sablatnig, and Allan Hanbury, editors, *Pattern Recognition*, volume 3663 of *Lecture Notes in Computer Science*, pages 117–124. Springer Berlin Heidelberg, 2005.

[FLL+12] Yuming Fang, Weisi Lin, Bu-Sung Lee, Chiew-Tong Lau, Zhenzhong Chen, and Chia-Wen Lin. Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum. *Multimedia, IEEE Transactions on*, 14(1):187–198, 2012.

[FM06] J.H. Fecteau and D.P. Munoz. Salience, relevance, and firing: a priority map for target selection. *Trends in cognitive sciences*, 10(8):382–390, 2006.

[FRC10] Simone Frintrop, Erich Rome, and Henrik I. Christensen. Computational visual attention systems and their cognitive foundations: A survey. *ACM Trans. Appl. Percept.*, 7(1):6:1–6:39, 2010.

[Fuk86] Kunihiko Fukushima. A neural network model for selective attention in visual pattern recognition. *Biological Cybernetics*, 55:5–15, 1986.

[FVE91] Daniel J. Felleman and David C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47, 1991.

[GCF06] Daniel J Graham, Damon M Chandler, and David J Field. Can the theory of "whitening."explain the center-surround properties of retinal ganglion cell receptive fields? *Vision research*, 46(18):2901–13, 2006.

[GDFVPD12] Antón Garcia-Diaz, Xosé R. Fdez-Vidal, Xosé M. Pardo, and Raquel Dosil. Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing*, 30(1):51 – 64, 2012.

[GDLFVP12] Antón García-Díaz, Víctor Leborán, Xosé R. Fdez-Vidal, and Xosé M. Pardo. On the relationship between optical variability, visual saliency, and eye fixations: A computational approach. *Journal of Vision*, 12(6), 2012.

[GHV09] Dashan Gao, Sunhyoung Han, and N. Vasconcelos. Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(6):989–1005, 2009.

[GJM92] G.-J. Giefing, H. Janssen, and H. Mallot. Saccadic object recognition with an active vision system. In *Pattern Recognition, 1992. Vol.I. Conference A: Computer Vision and Applications, Proceedings., 11th IAPR International Conference on*, pages 664–667, 1992.

[GKG98] J. P. Gottlieb, M. Kusunoki, and M. E. Goldberg. the representation of visual salience in monkey parietal cortex. *Nature*, 391:481, 1998.

[GKS05] Moshe Gur, Igor Kagan, and D. Max Snodderly. Orientation and direction selectivity of neurons in v1 of alert monkeys: Functional relationships and laminar distributions. *Cerebral Cortex*, 15(8):1207–1221, 2005.

[GM92] Melvyn A. Goodale and A.David Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20– 25, 1992.

[GMT89] S. Grossberg, E. Mingolla, and D. Todorovic. A neural network architecture for preattentive vision. *Biomedical Engineering, IEEE Transactions on*, 36(1):65–84, 1989.

[GS66] David M. Green and John A. Swets. *Signal detection theory and psychophysics*, volume 1974. Wiley New York, 1966.

[GZ10] Chenlei Guo and Liming Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *Image Processing, IEEE Transactions on*, 19(1):185 –198, 2010.

[HC08] Daw-Sen Hwang and Shao-Yi Chien. Content-aware image resizing using perceptual seam carving with human attention

model. In *Multimedia and Expo, 2008 IEEE International Conference on*, pages 1029–1032, 2008.

[HD07] Jeanny Hérault and Barthélémy Durette. Modeling visual perception for image processing. In Francisco Sandoval, Alberto Prieto, Joan Cabestany, and Manuel Graña, editors, *Computational and Ambient Intelligence*, volume 4507 of *Lecture Notes in Computer Science*, pages 662–675. Springer Berlin / Heidelberg, 2007.

[HE09] Milton Heinen and Paulo Engel. Nlook: a computational attention model for robot vision. *Journal of the Brazilian Computer Society*, 15:3–17, 2009.

[Hee88] David J. Heeger. Optical flow using spatiotemporal filters. *International Journal of Computer Vision*, 1:279–302, 1988.

[Hen03] John M. Henderson. Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, 7(11):498 – 504, 2003.

[HF04] J.M. Henderson and F. Ferreira. Scene perception for psycholinguists. In *The interface of language, vision, and action: Eye movements and the visual world*, pages 1–58. Psychology Press, 2004.

[HH05a] Dietmar Heinke and Glyn W Humphreys. Computational models of visual selective attention : A review. *Psychology*, 1(Part 4):1–34, 2005.

[HH05b] D.E. Huber and Christopher G. Healey. Visualizing data with motion. In *Visualization, 2005. VIS 05. IEEE*, pages 527–534, 2005.

[HHH09] Aapo Hyvärinen, Jarmo Hurri, and Patrick O. Hoyer. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision.* Computational Imaging and Vision, Vol. 39. Springer, 2009.

[HHK12] Xiaodi Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(1):194 –201, 2012.

[HK09] Jonathan Harel and Christof Koch. Attention in cognitive systems. In Lucas Paletta and John K. Tsotsos, editors, *Lecture Notes in Artificial Intelligence*, chapter On the Optimality of Spatial Attention for Object Detection, pages 1–14. Springer-Verlag, Berlin, Heidelberg, 2009.

[HKP07] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 545–552. MIT Press, Cambridge, MA, 2007.

[HM93] G.W. Humphreys and H.J. Muller. Search via recursive rejection (serr): A connectionist model of visual search. *Cognitive Psychology*, 25(1):43 – 110, 1993.

[HP07] Liqiang Huang and Harold Pashler. A boolean map theory of visual attention. *Psychological review*, 114(3):599, 2007.

[HR00] Stewart H. C. Hendry and R. Clay Reid. The koniocellular pathway in primate vision. *Annual Review of Neuroscience*, 23(1):127–153, 2000.

[HTP07] Liqiang Huang, Anne Treisman, and Harold Pashler. Characterizing the limits of human visual awareness. *Science*, 317(5839):823–825, 2007.

[HV10] Sunhyoung Han and Nuno Vasconcelos. Biologically plausible saliency mechanisms improve feedforward object recognition. *Vision Research*, 50(22):2295–2307, 2010.

[HYZF10] Jun Huang, Xiaokang Yang, Rui Zhang, and Xiangzhong Fang. Re-ranking image search results by multiscale visual saliency model. In *Broadband Multimedia Systems and Broadcasting (BMSB), 2010 IEEE International Symposium on*, pages 1–4, 2010.

[HZ07] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, 2007.

[HZ08] Xiaodi Hou and Liqing Zhang. Thumbnail generation based on global saliency. In Rubin Wang, Enhua Shen, and Fanji Gu, editors, *Advances in Cognitive Neurodynamics ICCN 2007*, pages 999–1003. Springer Netherlands, 2008.

[HZL+10] Gang Hua, Cha Zhang, Zicheng Liu, Zhengyou Zhang, and Ying Shan. Efficient scale-space spatiotemporal saliency tracking for distortion-free video retargeting. In Hongbin Zha, Rin-ichiro Taniguchi, and Stephen Maybank, editors, *Computer Vision – ACCV 2009*, volume 5995 of *Lecture Notes in Computer Science*, pages 182–192. Springer Berlin Heidelberg, 2010.

[IB05] L. Itti and P. F. Baldi. *A Principled Approach to Detecting Surprising Events in Video.* San Siego, CA, 2005.

[IB09] L. Itti and P. F. Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306, 2009.

[IK99] Laurent Itti and Christof Koch. Comparison of feature combination strategies for saliency-based visual attention systems. volume 3644, pages 473–482, 1999.

[IK00] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12):1489–1506, 2000.

[IK01] L. Itti and C. Koch. Computational modelling of visual attention. *Nature reviews. Neuroscience*, 2(3):194–203, 2001.

[IKN98] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(11):1254–1259, 1998.

[Itt04] L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *Image Processing, IEEE Transactions on*, 13(10):1304–1318, 2004.

[Itt05] L. Itti. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, 12(6):1093–1123, 2005.

[Itt06] Laurent Itti. Quantitative modelling of perceptual salience at human eye position. *Visual Cognition*, 14(4-8):959–984, 2006.

[Jam50] William James. *The Principles of Psychology, Vol. 1.* Dover Publications, 1950.

[JB85] Curtis L. Baker Jr and Oliver J. Braddick. Eccentricity-dependent scaling of the limits for short-range apparent motion perception. *Vision Research*, 25(6):803 – 812, 1985.

[JCN97] J S Joseph, M M Chun, and K Nakayama. Attentional requirements in a 'preattentive' feature search task. *Nature*, 387(6635):805–7, 1997.

[JDT12] Tilke Judd, Fredo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations, 2012.

[JEDT09] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2106–2113, 2009.

[JGS+73] B. Julész, EN Gilbert, LA Shepp, HL Frisch, et al. Inability of humans to discriminate between visual textures that agree in second-order statistics-revisited. *Perception*, 2(4):391–405, 1973.

[KAA03] Paul L Kaufman, Albert Alm, and Francis Heed Adler. *Adler's physiology of the eye : clinical application.* Mosby, St. Louis, 10th ed edition, 2003.

[Kah73] D. Kahneman. *Attention and effort.* Englewood Cliffs, NJ: Prentice-Hall., 1973.

[KC10] C. Kanan and G. Cottrell. Robust classification of objects, faces, and flowers using natural image statistics. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2472–2479, 2010.

[KJK11] Wonjun Kim, Chanho Jung, and Changick Kim. Spatiotemporal saliency detection and its applications in static and dynamic scenes. *Circuits and Systems for Video Technology, IEEE Transactions on*, 21(4):446 –456, 2011.

[Koh07] Adam Kohn. Visual adaptation: Physiology, mechanisms, and functional benefits. *Journal of Neurophysiology*, 97(5):3155–3164, 2007.

[KOS11] Elena Kokkinara, Oyewole Oyekoya, and Anthony Steed. Modelling selective visual attention for autonomous virtual characters. *Computer Animation and Virtual Worlds*, 22(4):361–369, 2011.

[Kow11] Eileen Kowler. Eye movements: The past 25 years. *Vision Research*, 51(13):1457 – 1483, 2011.

[KSJ+00] E.R. Kandel, J.H. Schwartz, T.M. Jessell, et al. *Principles of neural science*, volume 4. McGraw-Hill New York, 2000.

[KTZC09] Christopher Kanan, Mathew H. Tong, Lingyun Zhang, and Garrison W. Cottrell. Sun: Top-down saliency using natural statistics. *Visual Cognition*, 17(6-7):979–1003, 2009.

[Kul59] Solomon Kullback. *Information Theory and Statistics*. Wiley, New York, 1959.

[KWSF06] Wolf Kienzle, Felix A. Wichmann, Bernhard Schölkopf, and Matthias O. Franz. A nonparametric approach to bottom-up visual saliency. In *NIPS'06*, pages 689–696, 2006.

[KZ07] Ansgar R. Koene and Li Zhaoping. Feature-specific interactions in salience from combined feature contrasts: Evidence for a bottom–up saliency map in v1. *Journal of Vision*, 7(7), 2007.

[Laa99] J. Laarni. Allocating attention in the visual field: the effects of cue type and target–distractor confusability. *Acta Psychologica*, 103(3):281 – 294, 1999.

[LaB97] David LaBerge. Attention, awareness, and the triangular circuit. *Consciousness and Cognition*, 6(2–3):149 – 181, 1997.

[LAGDFVP13] Víctor Leborán Alvarez, Antón García-Díaz, XoséR. Fdez-Vidal, and XoséM. Pardo. Dynamic saliency from adaptative whitening. In JoséManuel Ferrández Vicente, JoséRamón Álvarez Sánchez, Félix Paz López, and Fco.Javier Toledo Moreo, editors, *Natural and Artificial Computation in Engineering and Medical Applications*, volume 7931 of *Lecture*

Notes in Computer Science, pages 345–354. Springer Berlin Heidelberg, 2013.

[LBF05] KangWoo Lee, H. Buxton, and Jianfeng Feng. Cue-guided search: a computational model of selective attention. *Neural Networks, IEEE Transactions on*, 16(4):910–924, 2005.

[LDP04] V. Leborán, R. Dosil, and X.M. Pardo. Anisotropic difusión applied to surface reconstruction of implicit surfaces. In *Actas del XIV Congreso Español de Informática Gráfica (CEIG'2004)*, pages 257–270, 2004.

[LFD15] Pablo Lanillos, Joao Filipe Ferreira, and Jorge Dias. Multisensory 3d saliency for artificial attention systems. In *3rd Workshop on Recognition and Action for Scene Understanding (REACTS'15).*, Lecture Notes in Computer Science. Springer, 2015.

[Li02] Zhaoping Li. A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, 6(1):9– 16, 2002.

[LLA+13] Jian Li, M. D. Levine, Xiangjing An, Xin Xu, and Hangen He. Visual saliency based on scale-space analysis in the frequency domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):996–1010, 2013.

[LMLCBT06] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau. A coherent computational approach to model bottom-up visual attention. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(5):802 –817, 2006.

[LQH+09] Huiying Liu, Xuekan Qiu, Qingming Huang, Shuqiang Jiang, and Changsheng Xu. Advertise gently - in-image advertising with low intrusiveness. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 3105–3108, 2009.

[LRD81] AG Leventhal, RW Rodieck, and B Dreher. Retinal ganglion cell classes in the old world monkey: morphology and central projections. *Science*, 213(4512):1139–1142, 1981.

[LSZ+07] Tie Liu, Jian Sun, Nan-Ning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object.

In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, 2007.

[LZX+09] Yin Li, Yue Zhou, Lei Xu, Xiaochao Yang, and Jie Yang. Incremental sparse saliency detection. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 3093–3096, 2009.

[MB13] Olivier Meur and Thierry Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods*, 45(1):251–266, 2013.

[MBMG04] Hugo Merchant, Alexandra Battaglia-Mayer, and AposotolosP. Georgopoulos. Neural responses in motor cortex and area 7a to real and apparent motion. *Experimental Brain Research*, 154:291–307, 2004.

[MDNBDN12] O. Muratov, Duc-Tien Dang-Nguyen, G. Boato, and F.G.B. De Natale. Saliency detection as a support for image forensics. In *Communications Control and Signal Processing (ISCCSP), 2012 5th International Symposium on*, pages 1–5, 2012.

[MF01] Tirin Moore and Mazyar Fallah. Control of eye movements and spatial attention. *Proceedings of the National Academy of Sciences*, 98(3):1273–1276, 2001.

[MGP07] Sophie Marat, Mickäel Guironnet, and Denis Pellerin. Video summarization using a visual attention model. In *Proceedings of the 15th European Signal Processing Conference*, pages 1784–1788, 2007.

[MGS+07] Thomas Michalke, Alexander Gepperth, Martin Schneider, Jannik Fritsch, and Christian Goerick. Towards a human-like vision system for resource-constrained intelligent cars. In *Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)*. Universitätsbibliothek Bielefeld, 2007.

[MHPG+09] Sophie Marat, Tien Ho Phuoc, Lionel Granjon, Nathalie Guyader, Denis Pellerin, and Anne Guérin-Dugué. Modelling spatio-temporal saliency to predict gaze direction for short videos. *International Journal of Computer Vision*, 82(3):231–243, 2009.

[Mil93] R. Milanese. *Detecting salient regions in an image: from biological evidence to computer implementation*. PhD thesis, 1993.

[MLZL02] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. A user attention model for video summarization. In *Proceedings of the tenth ACM international conference on Multimedia*, MULTIMEDIA '02, pages 533–542, New York, NY, USA, 2002. ACM.

[MM88] Carver A. Mead and M.A. Mahowald. A silicon model of early visual processing. *Neural Networks*, 1(1):91–97, 1988.

[MMTGM06] M. Mancas, C. Mancas-Thillou, B. Gosselin, and B. Macq. A rarity-based visual attention map - application to texture description. In *Image Processing, 2006 IEEE International Conference on*, pages 445 –448, 2006.

[Moz91] M.C. Mozer. *The perception of multiple objects: a connectionist approach*. Neural network modeling and connectionism. MIT Press, 1991.

[MP90] Jitendra Malik and Pietro Perona. Preattentive texture discrimination with early vision mechanisms. *J. Opt. Soc. Am. A*, 7(5):923–932, 1990.

[MRW95] S. Mannan, K.H. Ruddock, and D.S. Wooding. Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-d images. *Spatial Vision*, 9(3):363–386, 1995.

[MRW97] S.K. Mannan, K.H. Ruddock, and D.S. Wooding. Fixation patterns made during brief examination of two-dimensional images. *Perception-London*, 26:1059–1072, 1997.

[MS96] Michael C. Mozer and Mark Sitton. *Computational Modeling of Spatial Attention*. 1996.

[MS12] Stefan Mathe and Cristian Sminchisescu. Dynamic eye movement datasets and learnt saliency models for visual action recognition. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, Lecture Notes in Computer Science, pages 842–856. Springer Berlin Heidelberg, 2012.

[MS15] S. Mathe and C. Sminchisescu. Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(7):1408–1424, 2015.

[MSEB15] Irfan Mehmood, Muhammad Sajjad, Waleed Ejaz, and Sung Wook Baik. Saliency-directed prioritization of visual data in wireless surveillance networks. *Information Fusion*, 24(0):16 – 30, 2015.

[MSHH11] ParagK. Mital, TimJ. Smith, RobinL. Hill, and JohnM. Henderson. Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation*, 3(1):5–24, 2011.

[MWG+94] R. Milanese, H. Wechsler, S. Gill, J.-M. Bost, and T. Pun. Integration of bottom-up and top-down cues for visual attention using non-linear relaxation. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94. 1994 IEEE Computer Society Conference on*, pages 781 –785, 1994.

[MZH12] Keith A. May, Li Zhaoping, and Paul B. Hibbard. Perceived direction of motion determined by adaptation to static binocular images. *Current Biology*, 22(1):28 – 32, 2012.

[Nie07] E. Niebur. Saliency map. *Scholarpedia*, 2(8):2675, 2007.

[Not93] Hans-Christoph Nothdurft. The role of features in preattentive vision: Comparison of orientation, motion and color cues. *Vision Research*, 33(14):1937 – 1958, 1993.

[NSRM13] Anirvan S. Nandy, Tatyana O. Sharpee, John H. Reynolds, and Jude F. Mitchell. The fine structure of shape tuning in area {V4}. *Neuron*, 78(6):1102 – 1115, 2013.

[OAVE93] B A Olshausen, C H Anderson, and D C Van Essen. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, 13(11):4700–4719, 1993.

[OBH+01] N. Ouerhani, J. Bracamonte, H. Hugli, M. Ansorge, and F. Pellandini. Adaptive color image compression based on visual attention. In *Image Analysis and Processing, 2001.*

Proceedings. 11th International Conference on, pages 416–421, 2001.

[OBH06] Nabil Ouerhani, Alexandre Bur, and Heinz Hügli. Linear vs. nonlinear feature combination for saliency computation: A comparison with human vision. In Katrin Franke, Klaus-Robert Müller, Bertram Nickolay, and Ralf Schäfer, editors, *Pattern Recognition*, volume 4174 of *Lecture Notes in Computer Science*, pages 314–323. Springer Berlin Heidelberg, 2006.

[OSS09] Oyewole Oyekoya, William Steptoe, and Anthony Steed. A saliency-based method of simulating visual attention in virtual scenes. In *Proceedings of the 16th ACM Symposium on Virtual Reality Software and Technology*, VRST '09, pages 199–206, New York, NY, USA, 2009. ACM.

[OTCH03] A. Oliva, A. Torralba, M.S. Castelhano, and J.M. Henderson. Top-down control of visual attention in object detection. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 1, pages I – 253–6 vol.1, 2003.

[Oue04] N. Ouerhani. *Visual attention: from bio-inspired modeling to real-time implementation.* PhD thesis, 2004.

[PdHH90] R.Hans Phaf, A.H.C Van der Heijden, and Patrick T.W Hudson. Slam: A connectionist model for attention in visual selection tasks. *Cognitive Psychology*, 22(3):273–341, 1990.

[PIIK05] Robert J. Peters, Asha Iyer, Laurent Itti, and Christof Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18):2397 – 2416, 2005.

[PIW10] N. Parikh, L. Itti, and J. Weiland. Saliency-based image processing for retinal prostheses. *Journal of neural engineering*, 7(1):016006+, 2010.

[PLN02] Derrick Parkhurst, Klinton Law, and Ernst Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107– 123, 2002.

[Pro11] DIEM Project, 2011.

[PS00] C.M. Privitera and L.W. Stark. Algorithms for defining visual regions-of-interest: comparison with eye fixations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(9):970 –982, 2000.

[PS03] C.M. Privitera and L.W. Stark. Human-vision-based selection of image processing algorithms for planetary exploration. *Image Processing, IEEE Transactions on*, 12(8):917 – 923, 2003.

[PSD80] M.I. Posner, C.R. Snyder, and B.J. Davidson. Attention and the detection of signals. *Journal of Experimental Psychology: General; Journal of Experimental Psychology: General*, 109(2):160, 1980.

[PvdHH97] Eric O. Postma, H.Jaap van den Herik, and Patrick T.W. Hudson. Scan: A scalable model of attentional selection. *Neural Networks*, 10(6):993–1015, 1997.

[PVLA08] Jose L. Pardo-Vazquez, Victor Leborán, and Carlos Acuña. Neural correlates of decisions and their outcomes in the ventral premotor cortex. *The Journal of Neuroscience*, 28(47):12396–12408, 2008.

[PVLA09] Jose L Pardo-Vazquez, Victor Leborán, and Carlos Acuña. A role for the ventral premotor cortex beyond performance monitoring. *Proc Natl Acad Sci U S A*, 106(44):18815–9, 2009.

[PW09] O. Pele and M. Werman. Fast and robust earth mover's distances. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 460–467, 2009.

[RAS08] M.D. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, 2008.

[RB91] Raymond D. Rimey and Christopher M. Brown. Controlling eye movements with hidden markov models. *International Journal of Computer Vision*, 7:47–65, 1991.

[RBSdB05] Constantin A. Rothkopf, Dana H. Ballard, Brian T. Sullivan, and Kaya de Barbaro. Bayesian modeling of task dependent visual attention strategy in a virtual reality environment. *Journal of Vision*, 5(8):920, 2005.

[RCC98] Daniel L. Ruderman, Thomas W. Cronin, and Chuan-Chin Chiao. Statistics of cone responses to natural images: implications for visual coding. *J. Opt. Soc. Am. A*, 15(8):2036–2045, 1998.

[RDM+13] Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin, and Thierry Dutoit. A study of parameters affecting visual saliency assessment. 2013.

[RH09] E. Rahtu and J. Heikkila. A simple and efficient saliency detector for background subtraction. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 1137–1144, 2009.

[RKSH10] Esa Rahtu, Juho Kannala, Mikko Salo, and Janne Heikkilä. Segmenting salient objects from images and videos. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, volume 6315 of *Lecture Notes in Computer Science*, pages 366–379. Springer Berlin Heidelberg, 2010.

[RMC+12] Nicolas Riche, Matei Mancas, Dubravko Culibrk, Vladimir Crnojevic, Bernard Gosselin, and Thierry Dutoit. Dynamic saliency models and human attention: a comparative study on videos. In *Proceedings of the 11th Asian Conference on Computer Vision (ACCV)*, 2012.

[RMGD12] N. Riche, M. Mancas, B. Gosselin, and T. Dutoit. Rare: A new bottom-up saliency model. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 641–644, 2012.

[RMvdH+14] Alexander F. Russell, Stefan Mihalaş, Rudiger von der Heydt, Ernst Niebur, and Ralph Etienne-Cummings. A model of proto-object based saliency. *Vision Research*, 94(0):1 – 15, 2014.

[Ros99] Ruth Rosenholtz. A simple saliency model predicts a number of motion popout phenomena. *Vision Research*, 39(19):3157 – 3163, 1999.

[RP99] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, 1999.

[RR09] Fred Rieke and Michael E. Rudd. The challenges natural images pose for visual adaptation. *Neuron*, 64(5):605– 616, 2009.

[RRDU87] Giacomo Rizzolatti, Lucia Riggio, Isabella Dascola, and Carlo Umiltá. Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia*, 25(1, Part 1):31 – 40, 1987.

[RRP+03] E.D. Reichle, K. Rayner, A. Pollatsek, et al. The ez reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, 26(4):445–476, 2003.

[RSA08] Michael Rubinstein, Ariel Shamir, and Shai Avidan. Improved seam carving for video retargeting. *ACM Trans. Graph.*, 27(3):16:1–16:9, 2008.

[RvdLBC08] U. Rajashekar, I. van der Linde, A.C. Bovik, and L.K. Cormack. Gaffe: A gaze-attentive fixation finding engine. *Image Processing, IEEE Transactions on*, 17(4):564–573, 2008.

[RZ99a] Pamela Reinagel and Anthony M Zador. Natural scene statistics at the centre of gaze. *Network: Computation in Neural Systems*, 10(4):341–350, 1999.

[RZ99b] Pamela Reinagel and Anthony M Zador. Natural scene statistics at the centre of gaze. In *Network: Computation in Neural Systems*, pages 341–350, 1999.

[Sal00] D.D. Salvucci. A model of eye movements and visual attention. In *Proceedings of the International Conference on Cognitive Modeling*, pages 252–259, 2000.

[San90] Peter A. Sandon. Simulating visual attention. *J. Cognitive Neuroscience*, 2(3):213–231, 1990.

[SB03] Nathan Sprague and Dana H. Ballard. Eye movements for reward maximization. In *NIPS*, 2003.

[SCDM10] Christian Sandor, A. Cunningham, Arindam Dey, and V.-V. Mattila. An augmented reality x-ray system based on visual saliency. In *Mixed and Augmented Reality (ISMAR), 2010 9th IEEE International Symposium on*, pages 27–36, 2010.

[SG00] Dario D. Salvucci and Joseph H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, ETRA '00, pages 71–78, New York, NY, USA, 2000. ACM.

[SH02] Robert Shapley and Michael Hawken. Neural mechanisms for color perception in the primary visual cortex. *Current Opinion in Neurobiology*, 12(4):426 – 432, 2002.

[SHD07] O. Schwartz, A. Hsu, and P. Dayan. Space and time in visual context. *Nature Reviews Neuroscience*, 8(7):522–535, 2007.

[SJT10] Cristina Savin, Prashant Joshi, and Jochen Triesch. Independent component analysis in spiking neurons. *PLoS Comput Biol*, 6(4):e1000757, 2010.

[SKR13] Tatyana O. Sharpee, Minjoon Kouh, and John H. Reynolds. Trade-off between curvature tuning and position invariance in visual area v4. *Proceedings of the National Academy of Sciences*, 2013.

[SM06] M. Shahram and P. Milanfar. Statistical and Information-Theoretic Analysis of Resolution in Imaging. *Information Theory, IEEE Transactions on*, 52(8):3411–3437, 2006.

[SM09a] Hae Jong Seo and P. Milanfar. Nonparametric bottom-up saliency detection by self-resemblance. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, pages 45 –52, 2009.

[SM09b] Hae Jong Seo and Peyman Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12), 2009.

[SM10] Hae Jong Seo and P. Milanfar. Training-free, generic object detection using locally adaptive regression kernels. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1688–1704, 2010.

[SM11] Hae Jong Seo and P. Milanfar. Action recognition from one example. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(5):867–882, 2011.

[SO01] Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24(1):1193–1216, 2001.

[Sop09] Marat Sophie. *Modèles de saillance visuelle par fusion d'informations sur la luminance, le mouvement et les visages pour la prédiction de mouvements oculaires lors de l'exploration de vidéos.* PhD thesis, 2009.

[SP67] Robert Sekuler and Allan Pantle. A model for after-effects of seen movement. *Vision Research*, 7(5–6):427 – 439, 1967.

[SRS11] N Sharmili, P S Ramaiah, and G Swamynadhan. Image compression and resizing for retinal implant in bionic eye. *International Journal of Computer Science and Engineering Survey*, 2(2):31–37, 2011.

[SS77] Walter Schneider and Richard M Shiffrin. Controlled and automatic human information processing: I. detection, search, and attention. *Psychological Review*, 84(1):1–66, 1977.

[SSP+09] A. Sur, S.S. Sagar, R. Pal, P. Mitra, and J. Mukhopadhyay. A new image watermarking scheme using saliency based visual attention model. In *India Conference (INDICON), 2009 Annual IEEE*, pages 1–4, 2009.

[Str35] J. R. Stroop. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18(6):p643 – 662, 1935.

[SU88] A. Sha'asua and S. Ullman. Structural saliency: The detection of globally salient structures using a locally connected network. In *Computer Vision., Second International Conference on*, pages 321–327, 1988.

[SWZW10] Junge Sun, Yunhong Wang, Zhaoxiang Zhang, and Yiding Wang. Salient region detection in high resolution remote sensing images. In *Wireless and Optical Communications Conference (WOCC), 2010 19th Annual*, pages 1–4, 2010.

[SZ14] Khurram Soomro and Amir R Zamir. Action recognition in realistic sports videos. In *Computer Vision in Sports*, pages 181–208. Springer, 2014.

[Tat07] Benjamin W. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14), 2007.

[TBG05] Benjamin W. Tatler, Roland J. Baddeley, and Iain D. Gilchrist. Visual correlates of fixation selection: effects of scale and time. *Vision Research*, 45(5):643 – 659, 2005.

[TBV06] B.W. Tatler, R.J. Baddeley, and B.T. Vincent. The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision research*, 46(12):1857–1862, 2006.

[TCW+95] John K. Tsotsos, Scan M. Culhane, Winky Yan Kei Wai, Yuzhong Lai, Neal Davis, and Fernando Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78(1–2):507 – 545, 1995.

[TG80] Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97– 136, 1980.

[TG88] Anne Treisman and Stephen Gormican. Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, 95(1):15–48, 1988.

[TLMT+05] John K. Tsotsos, Yueju Liu, Julio C. Martinez-Trujillo, Marc Pomplun, Evgueni Simine, and Kunhao Zhou. Attending to visual motion. *Computer Vision and Image Understanding*, 100(1–2):3 – 40, 2005.

[TOCH06] A. Torralba, A. Oliva, M.S. Castelhano, and J.M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766, 2006.

[Toe11] A. Toet. Computational versus psychophysical bottom-up image saliency: A comparative evaluation study. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(11):2131–2146, 2011.

[Tor03] Antonio Torralba. Modeling global scene factors in attention. *J. Opt. Soc. Am. A*, 20(7):1407–1418, 2003.

[TPL+02] John Tsotsos, Marc Pomplun, Yueju Liu, Julio Martinez-Trujillo, and Evgueni Simine. Attending to motion: Localizing and classifying motion patterns in image sequences. In Heinrich Bülthoff, Christian Wallraven, Seong-Whan Lee, and Tomaso Poggio, editors, *Biologically Motivated Computer Vision*, volume 2525 of *Lecture Notes in Computer Science*, pages 439–452. Springer Berlin / Heidelberg, 2002.

[Tre85] Anne Treisman. Preattentive processing in vision. *Computer Vision, Graphics, and Image Processing*, 31(2):156– 177, 1985.

[TS99] K. G. Thompson and J. D. Schall. The detection of visual signals by macaque frontal eye field during masking. *Nature neuroscience*, 2(3):283–288, 1999.

[Tso11] John K. Tsotsos. *A Computational Perspective on Visual Attention.* MIT Press, 2011.

[TWK+10] B W Tatler, N J Wade, H Kwan, J M Findlay, and B M Velichkovsky. Yarbus, eye movements, and vision. *iPerception*, 1(1):7–27, 2010.

[TWY07] Minghui Tian, S. Wan, and Lihua Yue. A novel approach for change detection in remote sensing image based on saliency map. In *Computer Graphics, Imaging and Visualisation, 2007. CGIV '07*, pages 397–402, 2007.

[TZKM11] Thomas Töllner, Michael Zehetleitner, Joseph Krummenacher, and Hermann J. Müller. Perceptual basis of redundancy gains in visual pop-out search. *J. Cognitive Neuroscience*, 23(1):137–150, 2011.

[Und98] G. Underwood. *Eye guidance in reading and scene perception.* Elsevier Science, 1998.

[UoSC12] ILAB University of Southern California, 2012. http://ilab.usc.edu/toolkit/home.shtml.

[VEI12] R.C. Voorhies, L. Elazary, and L. Itti. Neuromorphic bayesian surprise for far-range event detection. In *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*, pages 1 –6, 2012.

[vZDT04] Wieske van Zoest, Mieke Donk, and Jan Theeuwes. The role of stimulus-driven and goal-driven control in saccadic visual selection. *Journal of Experimental Psychology: Human Perception and Performance*, 30(4):746–759, 2004.

[Wan95] B A Wandell. *Foundations of Vision*, volume 21. Sinauer Associates, 1995.

[WBKK11] Niklas Wilming, Torsten Betz, Tim C. Kietzmann, and Peter König. Measures and limits of models of fixation selection. *PLoS ONE*, 6(9):e24038, 2011.

[Web11] Michael A Webster. Adaptation and visual coding. *Journal of Vision*, 11(5):1–23, 2011.

[WFC+13] Pengfei Wan, Yunlong Feng, G. Cheung, I.V. Bajic, and O.C. Au. 3-d motion estimation for visual saliency modeling. *Signal Processing Letters, IEEE*, 20(10):972–975, 2013.

[WH04] J. M. Wolfe and T. S. Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews. Neuroscience*, 5:495–501, 2004.

[WM97] Michael A. Webster and J. D. Mollon. Adaptation and the color statistics of natural images. *Vision Research*, pages 3283–3298, 1997.

[WM03] Jonathan D. Wallis and Earl K. Miller. From Rule to Response: Neuronal Processes in the Premotor and Prefrontal Cortex. *J Neurophysiol*, 90(3):1790–1806, 2003.

[Wol94] Jeremy Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic Bulletin and Review*, 1:202–238, 1994.

[Wol01] JeremyM. Wolfe. Asymmetries in visual search: An introduction. *Perception and Psychophysics*, 63(3):381–389, 2001.

[WRKP05] Dirk Walther, Ueli Rutishauser, Christof Koch, and Pietro Perona. Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. *Computer Vision and Image Understanding*, 100(1-2):41– 63, 2005.

[WSA03] Thomas Wachtler, Terrence J Sejnowski, and Thomas D Albright. Representation of color stimuli in awake macaque primary visual cortex. *Neuron*, 37(4):681 – 691, 2003.

[WTS+10] Melanie Wilke, Janita Turchi, Katy Smith, Mortimer Mishkin, and David A. Leopold. Pulvinar inactivation disrupts selection of movement plans. *The Journal of Neuroscience*, 30(25):8650–8659, 2010.

[WTSL08] Yu-Shuen Wang, Chiew-Lan Tai, Olga Sorkine, and Tong-Yee Lee. Optimized scale-and-stretch for image resizing. *ACM Trans. Graph.*, 27:118:1–118:8, 2008.

[WY14] Qi Wang and Yuan Yuan. Learning to resize image. *Neurocomput.*, 131:357–367, 2014.

[XHZ11] Donna K. McClish Xiao-Hua Zhou, Nancy A. Obuchowski. *Statistical Methods in Diagnostic Medicine*, volume chapter VI. John Wiley and Sons, 2011.

[XXL+13] Weichen Xue, Dong Xing, Ming Lin, Jing Wang, Bin Sheng, and Lizhuang Ma. Depth-of-field rendering with saliency-based bilateral filtering. In *Computer-Aided Design and Computer Graphics (CAD/Graphics), 2013 International Conference on*, pages 399–400, 2013.

[Yar67] A L Yarbus. *Eye movements and vision*, volume chapter VI. Plenum Press, 1967.

[ZK11] Qi Zhao and Christof Koch. Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*, 11(3), 2011.

[ZL13] Liming Zhang and Weisi Lin. *Selective visual attention: Computational models and applications*. John Wiley & Sons, 2013.

[ZLR+13] Sheng-hua Zhong, Yan Liu, F Ren, Jinghuan Zhang, and Tongwei Ren. Video Saliency Detection via Dynamic Consistent Spatio-Temporal Attention Modelling. *AAAI*, pages 1063–1069, 2013.

[ZOM11] Xiao-Hua Zhou, Nancy A. Obuchowski, and Donna K. McClish. *Appendix B: Jackknife and Bootstrap Methods of Estimating Variances and Confidence Intervals*. John Wiley and Sons, Inc., 2011.

[ZP90] Hagit Zabrodsky and Shmuel Peleg. Attentive transmission. *Journal of Visual Communication and Image Representation*, 1:189–198, 1990.

[ZTM+08] Lingyun Zhang, Matthew H. Tong, Tim K. Marks, Honghao Shan, and Garrison W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7), 2008.

[ZTW09] Lingyun Zhang, Matthew H. Tong, and Garrison W. Sunday: Saliency using natural statistics for dynamic analysis of scenes. In *Proceedings of the Thirty-first Annual Cognitive Science Society Conference*, 2009.