UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

# CAMBIOS EN ESTRATEGIAS DE EXPLORACIÓN VISUAL DURANTE EXPERIMENTOS EN VISIÓN LIBRE

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN ELÉCTRICA

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL ELÉCTRICO

CAMILO IGNACIO ESPINOSA CURILEM

PROFESOR GUÍA:
Marcos Orchard Concha

PROFESORA CO-GUÍA:
Christ Devia Manríquez

MIEMBROS DE LA COMISIÓN:
Pedro Maldonado Arbogast
Jorge Silva Sánchez

Este trabajo ha sido parcialmente financiado por:
Fundación Guillermo Puelma
Proyecto FONDECYT 1210031

SANTIAGO DE CHILE
2023

RESUMEN DE LA TESIS PARA OPTAR
AL GRADO DE MAGÍSTER EN CIENCIAS
DE LA INGENIERÍA, MENCIÓN ELÉCTRICA
Y DE LA MEMORIA PARA OPTAR AL TÍTULO
DE INGENIERO CIVIL ELÉCTRICO.
POR: CAMILO IGNACIO ESPINOSA CURILEM
FECHA: 2023
PROF. GUÍA: MARCOS ORCHARD CONCHA
PROF. CO-GUÍA: CHRIST DEVIA MANRÍQUEZ

# CAMBIOS EN ESTRATEGIAS DE EXPLORACIÓN VISUAL DURANTE EXPERIMENTOS EN VISIÓN LIBRE

El estudio de dónde decidimos mirar puede revelar las prioridades de nuestro sistema visual al explorar nuestro entorno. Cuando se observa una escena, el procesamiento de los estímulos visuales ocurre durante períodos en que la mirada se encuentra relativamente inmóvil, llamados fijaciones. Debido a que las fijaciones suelen coincidir con los objetos o regiones que atendemos, es de interés estudiar dónde ocurren y cómo esta ubicación se relaciona con la información visual presente en la escena. Una de las formas en que la información visual se puede separar es en función de sus componentes *bottom up*, (propiedades puramente físicas) y *top down* (significado). Este trabajo investiga cambios en la estrategia de exploración en escenas estáticas a lo largo del tiempo, tomando como referencia estos componentes. Los datos utilizados fueron registrados durante experimentos de visión libre con sujetos humanos y los resultados muestran que los factores *top down* son muy relevantes al principio de los experimentos pero disminuyen su importancia con el tiempo. Por otro lado, los factores *bottom up* parten siendo muy poco relevantes, aumentando su influencia en la selección de fijaciones hasta aproximadamente 2 segundos después de la presentación de los estímulos.

*A Millaray, mi madre.*
*Todo lo bueno en mi vida ha sido posible gracias a ti.*

# Agradecimientos

Me gustaría agradecer Marcos Orchard por el apoyo y empuje entregados desde el inicio de este proceso, a Christ Devia y Pedro Maldonado por su gran disposición y confianza en mi trabajo y al equipo de Neurosistemas por ser un espacio profundamente enriquecedor, formado por personas extremadamente interesantes y generosas con su conocimiento.

Todas/os quienes nombro aquí aportaron a que disfrutara mi inserción en este espacio maravilloso, llamado Neurociencia, y a ellas/os les debo buena parte de lo que aprendí en esta tesis.

Finalmente, agradezco a la Fundación Guillermo Puelma y al apoyo de Marcos Orchard a través del Proyecto FONDECYT 1210031 por financiar el desarrollo de esta tesis.

# Tabla de Contenido

# Capítulo 1

# Introducción

## 1.1. Motivación

En humanos, la visión ocurre gracias a células fotosensibles ubicadas en la retina, las que se concentran con mayor densidad en una zona llamada *fóvea centralis* [1], al centro del campo visual. Esto produce que en dicho lugar se perciban los estímulos visuales en mayor resolución que en todo el resto del ojo. Por ello, al observar una escena, se debe llevar el centro de la mirada constantemente a las regiones u objetos que se desea ver con claridad. En consecuencia, una exploración visual común se conforma principalmente de dos eventos: períodos en que el ojo se encuentra relativamente inmóvil, llamados fijaciones; y movimientos rápidos que relocalizan la mirada sobre un nuevo punto de fijación, llamados sacadas.

Se cree que el cerebro humano ha desarrollado estrategias para priorizar y muestrear estímulos visuales con el fin de procesarlos eficientemente, el estudio de estos mecanismos se conoce como Atención Visual y la investigación en esta línea se centra en cómo se seleccionan, perciben e interpretan dichos estímulos. Los estudios sobre Atención Visual tienen el potencial de ayudar a identificar la presencia de trastornos del neurodesarrollo como el Trastorno de Espectro Autista, Déficit Atencional, entre otros [2, 3]. La investigación en esta línea también aporta a la formulación de modelos por computador para simular y predecir los procesos atencionales humanos, con aplicaciones que incluyen visión por computador, interfaces humano máquina y robótica [4].

En general la atención visual se estudia a través de la ubicación de las fijaciones, asumiendo que una región fijada está siendo explícitamente atendida por el sujeto. Con esta idea, un formato de modelo ampliamente utilizado son los llamados mapas de prominencia (*saliency maps*). Estos mapas indican, a través de una representación en dos dimensiones, qué regiones tienen mayor probabilidad de ser atendidas, para una escena visual dada. La forma en que se calculan estas probabilidades varía dependiendo del modelo y del componente de la atención que se desea modelar [4].

Una de las formas de estudiar la Atención Visual es a través de experimentos de visión libre, en que los sujetos son presentados con escenas visuales sin una tarea a realizar más que explorar la imagen a voluntad. Dichos experimentos representan un acercamiento más natural y representativo de los mecanismos de Atención Visual, por lo que han ganado fuerza en los últimos años. Durante estos experimentos se ha encontrado tendencias temporales en

la exploración ocular que indican que los sujetos cambian su comportamiento a medida que se familiarizan con el estímulo. Estas tendencias se manifiestan, para la mayoría de los estudios, como cambios sobre las fijaciones y sacadas que se realizan sobre las imágenes mostradas [5–9].

En base a lo anterior, una gran cantidad de trabajos se ha centrado en analizar la existencia de dos modos de exploración: un modo ambiental y un modo focal [10–17]. Esta separación se realiza a través de características como la duración de las fijaciones o la amplitud de las sacadas, encontrando evidencias de cambios discretos (no graduales) en dichas características [13] o relaciones entre cada uno de los modos y la activación de distintas regiones del cerebro humano [14].

El presente trabajo se centra en estudiar cómo evoluciona la atención visual durante experimentos de visión libre en función de las componentes *bottom up* y *top down*. Estos componentes se han propuesto como una forma de explicar y diferenciar qué tipos de información pueden atraer la atención en una escena dada. El primero se relaciona exclusivamente al estímulo y a cómo sus características físicas (bordes, texturas, colores, etc.) pueden llamar la atención de un observador. El segundo tiene que ver con decisiones intencionales, guiadas por factores cognitivos, interpretaciones, memoria u objetivos del sujeto, entre otros [4].

Para estudiar cómo cambian las prioridades atencionales en el tiempo, se estudiará qué tanto coinciden las regiones donde miran los sujetos con aquellas que contienen información *bottom up*/*top down*. Para cuantificar esta similitud se implementarán modelos matemáticos de Atención Visual que representen los componentes descritos y se compararán a las fijaciones realizadas a través de técnicas ampliamente utilizadas en la literatura para la evaluación de mapas de Atención Visual (*saliency maps*).

## 1.2. Hipótesis

*Durante la exploración visual libre de escenas estáticas, la conducta ocular exhibe una tendencia sistemática, relacionada a los componentes bottom up y top down de la atención visual. Este cambio se manifiesta a través de la asignación de fijaciones, cuya localización coincide inicialmente con un componente y, a medida que pasa el tiempo, se acerca progresivamente al otro.*

## 1.3. Objetivos

### 1.3.1. Objetivo General

Caracterizar las dinámicas temporales de exploración visual durante experimentos de visión libre a través de la localización de fijaciones y su similitud a modelos que representen las componentes *bottom up* y *top down* de la atención visual.

### 1.3.2. Objetivos Específicos

1. Representar atención top down y bottom up para cada escena visual mostrada en los experimentos.

2. Representar la atención real de los sujetos, manifestada a través de la ubicación de sus fijaciones, de manera que éstas puedan ser comparadas a los modelos de atención utilizados.

3. Evaluar la similitud entre los modelos de atención y las fijaciones realizadas por los sujetos, en función del tiempo de experimento.

4. Contrastar estas dinámicas de acuerdo al tipo de estímulo presentado a los sujetos.

5. Identificar momentos en que se observe un cambio de comportamiento en las medidas calculadas.

## 1.4. Estructura de la Tesis

El documento a continuación se estructura de la siguiente manera: en el Capítulo 2 "Marco Teórico y Estado del Arte" se describe el contexto en que se desarrolla el trabajo y las herramientas a aplicar. En el Capítulo 3 "Materiales y Métodos" se explican los datos utilizados y el procesamiento que se realiza sobre ellos para realizar el análisis propuesto. Los resultados obtenidos según lo explicado en este capítulo se presentan y discuten en el Capítulo 4. Finalmente, en el Capítulo 5 se concluye el trabajo realizado, los resultados obtenidos y el trabajo a futuro que se vislumbra a partir de éstos.

# Capítulo 2

# Marco Teórico y Estado del Arte

## 2.1. Estructura del ojo y movimientos oculares

En el ojo humano y de otros primates, la visión ocurre gracias a la existencia de células fotosensibles ubicadas en una región de su membrana interna, conocida como retina. Estas células, separadas en bastones y conos, cumplen la función de recibir la luz del exterior y transformarla en los potenciales eléctricos que serán transmitidos al cerebro para el procesamiento que da origen a la visión [1]. Los bastones, mucho más sensibles a la luz que los conos, se asocian a mecanismos de visión a baja luminosidad y en promedio se pueden encontrar en cantidades cercanas a los 130 millones en una retina humana. Los conos, por su lado, se relacionan con la visión a color y requieren una intensidad de luz mayor para funcionar, se encuentran a su vez en cantidades menores, alcanzando un promedio aproximado de 7 millones en cada ojo.

Bastones y conos se concentran con significativamente mayor densidad sobre una pequeña parte al centro del campo visual denominada fóvea. Esta región ocupa aproximadamente 0.4 mm (cerca de 1º visual) de diámetro en la retina y representa el lugar donde los estímulos se perciben en mayor resolución. A mayor distancia del centro de la fóvea, la densidad de células fotosensibles disminuye fuertemente y, por lo tanto, también lo hace la resolución visual.

Que la agudeza visual se concentre en una pequeña zona al centro del campo visual implica que, para obtener información del entorno en alta calidad, se deba mover constantemente los ojos, dirigiendo la fóvea hacia regiones que por distintos motivos puedan ser relevantes o interesantes en el momento. En consecuencia, al observar escenas estáticas de manera libre, la conducta ocular se puede dividir típicamente en dos tipos de evento, llamados fijaciones y sacadas. Las fijaciones corresponden a períodos durante los cuales los ojos se mantienen relativamente inmóviles, mientras que las sacadas representan movimientos cortos y de alta velocidad en los que se relocaliza la vista sobre un nuevo punto de fijación. El procesamiento visual ocurre durante los períodos de fijación y se cree que durante las sacadas éste se suspende momentáneamente [18].

## 2.2. Atención Visual

La corteza visual es la región del cerebro asociada al procesamiento de los estímulos percibidos por los ojos. Debido a que éstos estímulos llegan constantemente y en grandes cantidades,

se cree que el cerebro ha desarrollado ciertos mecanismos que permiten seleccionar y procesar partes relevantes de esta información de manera jerarquizada, agilizando y flexibilizando la tarea de dar sentido a lo que se ve [19]. El estudio de estos mecanismos y la generación de modelos computacionales que los repliquen se conoce como Atención Visual. La investigación en Atención Visual ha dado lugar a una variedad de teorías para explicar diferentes aspectos del comportamiento ocular y del procesamiento visual.

Para fines de este trabajo, será de principal interés estudiar el efecto de los componentes *bottom up* y *top down* de la atención visual sobre la conducta ocular y cómo estos efectos varían a medida que pasa el tiempo durante cada experimento. Estos componentes serán descritos en la sección siguiente.

### 2.2.1. Componentes *bottom up* y *top down*

Una forma de explicar cómo se distribuye la atención visual es a partir de sus componentes *bottom up* y *top down*. El componente *bottom up* se relaciona a operaciones básicas y rápidas que ocurren en las primeras etapas de la corteza visual de manera paralela y que se aplican sobre distintas características de la escena observada: color, orientación, movimiento, etc. Estas operaciones ocurren de manera muy rápida y en paralelo. Por otro lado, la componente *top down* depende de la tarea a desarrollar, de las interpretaciones del sujeto, de su estado mental, de sus experiencias previas o una mezcla de todas ellas. Se cree que la componente *top down* se procesa en regiones de mayor nivel de la corteza visual y que es mucho más lenta que la componente *bottom up*.

Si bien existe una gran cantidad de modelos para cuantificar la relación entre regiones de una escena visual y la atención *bottom up*, hacerlo para el componente *top down* es una tarea difícil, debido a los procesos de alto nivel que representa. En contexto de experimentos con instrucciones definidas se ha observado que la existencia de una tarea a desarrollar afecta significativamente la distribución de atención (expresada a través de la ubicación de las fijaciones) sobre una escena visual.

En el estudio realizado por Yarbus [1] y replicado por DeAngelus y Pelz [20], una escena con personas se mostró a un conjunto de observadores, a quienes se le solicitó contestar preguntas de distinta naturaleza de acuerdo a lo observado. En ambos experimentos se observó que el comportamiento ocular variaba significativamente con el tipo de pregunta realizada. Por ejemplo, lo que debían hacer incluía, entre otras cosas, estimar la edad de las personas en la escena o sus condiciones materiales y recordar las ropas que llevaban o la posición de objetos adicionales. Para cada una de estas preguntas las fijaciones se concentraron el lugares distintos; caras, la ropa, muebles, objetos en la mesa, etc.

La existencia de una tarea facilita la identificación de información *top down* en una escena. Sin embargo, el contexto de exploración libre representa un desafío especial ya que lo *top down* responderá a prioridades más difíciles de esclarecer sin un objetivo explícito por cumplir. Existen, de todas formas, distintos estudios que han identificado elementos que se relacionan a estos procesos de alto nivel de manera relativamente independiente de la existencia de una tarea o instrucción a seguir. Los modelos propuestos y los elementos de una imagen que se pueden utilizar para cuantificar información relacionada a los componentes *bottom up* y *top down* se discutirán en las siguientes secciones.

### 2.2.2. Exploración libre

Los experimentos de visión libre se basan en la presentación de estímulos visuales sin pedir a los sujetos una tarea específica a cumplir más que explorar la imagen según su interés. Siguiendo este esquema, distintos estudios han reportado tendencias sistemáticas e independientes del estímulo en el comportamiento ocular de los sujetos. Por ejemplo, se ha observado que las fijaciones tienden a durar más a medida que pasa el tiempo del experimento, a la vez que las sacadas van disminuyendo su amplitud [5–7, 9].

A partir de estos resultados, se ha propuesto la existencia de dos modos de exploración: uno ambiental, caracterizado por producir fijaciones de corta duración y sacadas de gran amplitud; y otro focal, relacionado a fijaciones largas y sacadas pequeñas. Esta separación en el comportamiento ocular ha encontrado creciente evidencia en otros tipos de medida. Por ejemplo, Velichkovsky et al. [14] observaron, a través de mediciones de resonancia magnética funcional (fMRI), que durante períodos de fijaciones largas se activan regiones del cerebro distintas a cuando ocurren fijaciones cortas. Si bien los umbrales para definir fijaciones cortas o largas y sacadas grandes o pequeñas varían de estudio en estudio, éstos rondan los 300 ms para las fijaciones y los 5º visuales para las sacadas.

En síntesis, una gran cantidad de evidencia apunta a que el comportamiento ocular es dinámico y que cambia a medida que los sujetos se familiarizan con el estímulo presentado. Los factores que afectan esta evolución en el tiempo y la forma en que ésta se manifiesta representan una línea de investigación en desarrollo. Un mayor entendimiento de los mecanismos que gobiernan el comportamiento ocular abre la posibilidad de entender en mayor detalle la forma en que el cerebro humano selecciona y procesa los estímulos visuales. Conocer estas estrategias permitiría, entre otras cosas, ayudar al estudio de trastornos como el Trastorno de Espectro Autista, Déficit Atencional [21], la esquizofrenia [22][23], el Alzheimer [24] y otras enfermedades neurodegenerativas o del neurodesarrollo [25], así como enriquecer modelos de visión por computador, interfaces humano máquina y aplicaciones en robótica.

### 2.2.3. Modelos de Atención Visual

Uno de los formatos en que se desarrollan los modelos computacionales para representar la atención visual es a través los mapas de prominencia (*saliency maps*), es decir, mapas topográficos en dos dimensiones que entregan la probabilidad de que una fijación ocurra sobre cierta región de una imagen dada. A mayor valor, mayor es la prominencia de dicha región. Las características con que se generan estos mapas, el tipo de procesamiento que se realiza sobre ellas y la teoría psicológica/fisiológica en la que se basan da lugar a una gran variedad de modelos de atención visual. Éstos serán discutidos en la sección siguiente.

#### 2.2.3.1. Modelos *bottom up*

Debido a que los procesos cerebrales relacionados al componente *bottom up* de la atención visual representan operaciones básicas en zonas de más bajo nivel del cerebro, existe una gran cantidad de modelos y propuestas para representarlo. A continuación, se presenta una síntesis de las teorías y modelos más utilizados en la literatura para la generación de mapas de prominencia, basados exclusivamente en las características físicas de la imagen.

Una de las teorías que ha generado más modelos de atención visual es la Teoría de In-

tegración de Características, basada en la noción de que varias propiedades visuales de una imagen son extraídas de manera involuntaria, rápida y en paralelo sobre la escena, antes incluso que exista un reconocimiento de los objetos presentes. Basándose en ella, Itti & Koch propusieron un procedimiento en el cual se aplican operaciones a distintas escalas sobre los componentes de color, intensidad y orientación de la escena visual [26]. Estas operaciones producen distintos mapas de características por cada componente, los que se combinan para generar un mapa de prominencia final. A partir de este modelo, distintas extensiones se han desarrollado para incluir movimiento (en videos) [27], otros aspectos biológicamente plausibles como la sensibilidad al contraste o el enmascaramiento visual [28] y características ligadas a la simetría de los objetos [29].

Entre los modelos con mejor rendimiento sobre la mayor parte de las métricas y bases de datos de la *MIT Saliency Benchmark* que no utilizan *deep learning* [30][31], encontramos el modelo *Boolean Map Saliency*. Éste se basa en la teoría de mapas booleanos y el principio Gestalt de segregación figura-fondo. Los principios de Gestalt, formulados para explicar la forma en que humanos agrupan elementos en una escena visual, proponen que figuras con formas cerradas tienen más probabilidad de ser atendidas que elementos del fondo [32]. De aquí que figuras con fronteras definidas se consideren prominentes (con mayor probabilidad de ser fijadas) y que el modelo use la *surroundedness* (o grado en que sus límites están definidos) para calcular la probabilidad de fijación sobre ellas. El nombre del modelo hace referencia a la utilización de mapas booleanos, formados por unos y ceros, que separan las regiones que pertenecen o no al fondo, basándose en distintos tipos de características: color, intensidad y orientación.

Otras propuestas involucran conceptos de Teoría de Información y cuantifican la prominencia visual de acuerdo a la rareza de un objeto o región, como el modelo de Bruce & Tsotos [33], donde la prominencia de una región se relaciona a la información o "sorpresa" que puede portar de acuerdo a las regiones que la contornan. En una línea diferente, se han propuesto modelos cuyo foco se inclina más hacia los datos (*data-driven*) que hacia la plausibilidad biológica. Por ejemplo, en modelos como el propuesto por Jiang et al. [34] y Liu et al. [35] las escenas son segmentadas en distintos niveles de detalle a partir de diferentes características y procesamientos. Con estas segmentaciones, basadas en contrastes y similitudes de color y texturas, los modelos aprenden a seleccionar características, ajustándose a imágenes ya anotadas con posiciones de objetos prominentes o de fijaciones realizadas por humanos.

Finalmente, en los últimos años se han implementados modelos de *deep learning*, es decir, redes neuronales con varias capas de neuronas y configuraciones que en algunos casos asemejan mecanismos de las etapas tempranas del procesamiento visual en humanos [36–38]. Estos modelos requieren una gran cantidad de imágenes, las que deben ir emparejadas con información sobre los lugares donde sujetos reales miraron. El entrenamiento con estos datos puede generar modelos de inferencia o predicción de fijaciones, así como extractores de características a utilizar con otro tipo de herramientas para obtener la prominencia de las regiones en una escena. Si bien estos modelos superan en general a la mayoría de los nombrados anteriormente, traen consigo ciertas desventajas relevantes: requieren gran cantidad de datos, tienen un alto costo computacional y, lo más importante, dificultan sustancialmente una interpretación para comparar con el comportamiento ocular humano.

#### 2.2.3.2. Modelos *top down*

De acuerdo a lo explicado al inicio de la sección de Atención Visual, el componente *top down* se asocia a procesos visuales de alto nivel en el cerebro humano. Elementos o regiones que se relacionan a este componente de la atención responden a distintos factores, como experiencias previas, el contexto de la escena u objetivos específicos durante la exploración. Éstos son, en su mayoría, muy difíciles de cuantificar. En general, experimentos donde se debe realizar una tarea específica facilitan la definición de regiones de interés alineadas con los objetivos y necesidades de dicha tarea. Como se mostró en [1] y [20], dependiendo de la pregunta que se deba responder, los sujetos fijarán su mirada en distintas regiones de una escena con el fin de recabar información que permita generar una respuesta adecuada.

En un contexto más natural, como lo son los experimentos de visión libre, cuantificar regiones que evoquen atención *top down* se hace más difícil al no existir elementos directamente identificables como útiles para la tarea en proceso. Existen, sin embargo, estudios que han logrado distinguir prioridades atencionales más allá de las características físicas de una escena dada o, inclusive, de la instrucción entregada. En 2017, Flechsenhar et al. [39] observaron que estímulos socialmente relevantes, como caras y cuerpos humanos, atraen en significativamente mayor medida la atención de los sujetos, independiente de la tarea que estaban realizando. Más aún, los experimentos realizados se diseñaron de tal forma que estos elementos sociales no sirvieran de ninguna forma en la resolución de las tareas y los modelos de atención *bottom up* aplicados asignaron una prominencia significativamente menor en estos elementos que en otras zonas en la escena visual mostrada. En esta misma línea, por ejemplo, modelos como el propuesto por Cerf et al. [40] utilizan detectores de caras para asignar una mayor probabilidad de fijación en las regiones donde éstas se encuentran.

Otros experimentos han entregado evidencia de que objetos con cualidades semánticas o que se relacionan con el contexto de las escenas observadas guían la visión, tanto durante exploración libre como en contexto de búsqueda de objetos [41]. Se ha observado también que pacientes con agnosia visual (incapacidad para reconocer objetos a pesar de tener una agudeza visual relativamente normal) se diferencian de sujetos sanos, dedicando mayor atención a características *bottom up* de la imagen [42]. En ambos estudio se hipotetizó que la existencia de objetos o regiones que entreguen información de contexto sobre la imagen atraen la atención visual en mayor medida que las características físicas por sí solas.

En la actualidad, gracias a los avances en modelos de *deep learning* y métodos de transferencia de aprendizaje, se ha podido incluir factores contextuales más allá de las caras y cuerpos humanos. Un ejemplo es el trabajo de Mahdi et al. [43], donde se utilizan redes neuronales convolucionales entrenadas sobre grandes bases de datos como *ImageNet* [44] para la clasificación de imágenes. La estimación de prominencia visual se realiza a partir de los mapas de activación de clase en la red neuronal. Estos mapas de activación representan las regiones de una imagen donde la red neuronal ha "aprendido" a "mirar" para inferir el tipo de entidad que existe en ella; un perro, un bote, una bicicleta, etc. A partir de una combinación ponderada de estos mapas de activación producidos por el modelo para cada clase disponible, se obtiene una representación en 2 dimensiones, donde objetos detectados reciben mayor valor de prominencia. Este acercamiento, sin embargo, está limitado por la cantidad de clases u objetos con los que se han entrenado las redes (1000 en el caso de *ImageNet*) y requiere un nivel de definición mínimo en estas regiones para generar una predicción aceptable: objetos

mostrados a medias o con partes ocultas por otros objetos no son detectados por los modelos.

## 2.3. Modelos Utilizados

El presente trabajo se centra en el cambio de prioridades atencionales a lo largo del tiempo. Específicamente, se evalúa el peso que tienen los componentes *bottom up* y *top down* en los distintos momentos de cada experimento. Para esta comparación se busca separar de la mejor manera posible un componente de otro, tomando en cuenta que ambos interactúan y coinciden frecuentemente en las escenas visuales.

En esta línea, para cuantificar la atención *bottom up* se evita utilizar modelos generados a partir del entrenamiento o ajuste sobre datos empíricos, ya que su objetivo principal es predecir dónde ocurrirá una fijación y dificultan la interpretación de las regiones que estiman prominentes. Por otro lado, se excluyen también modelos basados en sorpresa, rareza o auto información ya que los procesos asociados a estas características constituyen etapas de mayor nivel y, por lo tanto, más cercanas a la atención *top down*, dificultando la separación con los modelos utilizados para este componente.

En base a lo expuesto, se seleccionan dos modelos: el modelo propuesto por Itti & Koch, basado en la Teoría de Integración de Características y el modelo BMS, basado en la teoría Gestalt de segregación figura-fondo. Ambos cumplen con basarse exclusivamente en las características físicas de la escena y con tener una teoría neuropsicológica que facilita la interpretación de los resultados.

### 2.3.1. Itti & Koch

En este modelo, la imagen de entrada se submuestrea para obtener varias versiones de ésta en resoluciones menores. El submuestreo genera versiones a 9 escalas espaciales; desde 1:1 (escala 0) a 1:256 (escala 8). Cada escala se descompone en 3 características principales: color, intensidad y orientación.

Para el color se generan 4 canales:

$$\mathcal{R} = r - \frac{(g+b)}{2} \tag{2.1}$$

$$\mathcal{G} = g - \frac{(r+b)}{2} \tag{2.2}$$

$$\mathcal{B} = b - \frac{(r+g)}{2} \tag{2.3}$$

$$\mathcal{Y} = \frac{(r+g)}{2} - \frac{|r-g|}{2} - b \tag{2.4}$$

Donde los canales de cada color se denotan r: rojo, g: verde, b: azul e y: amarillo. El canal de luminancia se calcula como el promedio de los tres colores RGB:

$$I = \frac{(r+g+b)}{3} \tag{2.5}$$

Las características de orientación se calculan a través de filtros *Gabor* (Figura 2.1) aplicados en las 9 escalas para 4 orientaciones diferentes $\theta \in \{0^o, 45^o, 90^o, 135^o\}$.

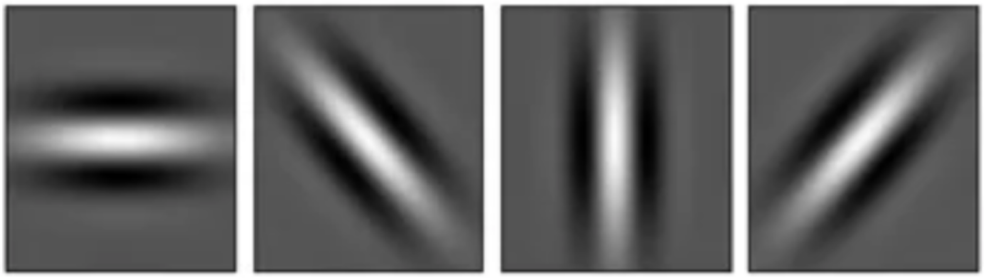


Figura 2.1: Ejemplos de filtros Gabor para las distintas orientaciones utilizadas, de izquierda a derecha: $0^o$, $45^o$, $90^o$ y $135^o$ respectivamente.

Para cada canal se aplican operaciones de centro-contorno (*center-surround*) entre los mapas obtenidos a distintas escalas. La operación consiste en sobremuestrear el mapa de menor escala (contorno) para tener las mismas dimensiones que el mapa de mayor escala (centro). La diferencia entre ambos entrega bordes y cambios abruptos de contraste, con los que se generarán los mapas de características del canal en la escala correspondiente. Para cada canal se calculan 6 mapas de características, comparando 3 escalas de centro $c \in \{2, 3, 4\}$ con 2 escalas de contorno $s \in \{3, 4\}$. El canal de intensidad se define como:

$$\mathcal{I}(c, s) = I(c) \ominus I(s) \tag{2.6}$$

Donde $\ominus$ representa la diferencia a través de escalas, en que el mapa de menor resolución se interpola para ajustarse al tamaño del mapa de mayor resolución y luego se computa una diferencia punto a punto entre ambos.

En canales de color se realiza la operación de centro-contorno sobre conjuntos de colores complementarios (rojo-verde y azul-amarillo):

$$\mathcal{RG}(c, s) = |(R(c) - G(c)) \ominus (G(s) - R(s))| \tag{2.7}$$
$$\mathcal{BY}(c, s) = |(B(c) - Y(c)) \ominus (B(s) - Y(s))| \tag{2.8}$$

En total se generan 42 mapas de características: 6 para intensidad, 12 para colores complementarios y 24 para las orientaciones. Todos los mapas se normalizan y suman, generando 3 mapas de sobresaliencia (*conspicuity*), los cuales se promedian para obtener el mapa de prominencia final:

$$\mathcal{S} = \frac{1}{3}(\mathcal{N}(\mathcal{I}) + \mathcal{N}(\mathcal{C}) + \mathcal{N}(\mathcal{O})) \tag{2.9}$$

Donde $\mathcal{N}$ es la operación de normalización, $\mathcal{I}$, $\mathcal{C}$ y $\mathcal{O}$ son los mapas de sobresaliencia basados en intensidad, color y orientación respectivamente.

El resultado es un mapa con las mismas dimensiones que la imagen donde se aplica. Este mapa tendrá asignado un valor numérico no negativo en cada pixel, donde a mayor valor, mayor probabilidad de fijación.

### 2.3.2. Boolean Map Saliency

El método de *Boolean Map Saliency* (BMS) consiste en generar un conjunto de mapas booleanos $B = \{B_1, B_2, ..., B_n\}$ a partir de una imagen de entrada $I$. Los mapas en $B$ se obtienen aplicando umbrales sobre componentes de la imagen (color, orientación, profundidad, movimiento, etc.) y asignando el valor 1 o 0 a cada pixel que esté sobre o bajo dicho umbral respectivamente [45].

Se aplican umbrales sobre 3 canales de color (espacio de color CIE Lab):

$$B_i = \mathbf{THRESH}(\sigma(I), \theta) \tag{2.10}$$

Donde **THRESH()** es la operación de umbral, $\sigma(I)$ es el canal de color donde se aplican los umbrales y $\theta$ es el valor del umbral, el cual se muestrea entre los valores 0 y 255 en pasos de tamaño fijo $\delta = 8$. Esta operación extrae distintos bordes para cada umbral aplicado, obteniéndose el mapa booleano correspondiente y para cada mapa $B_i$ se genera también un mapa invertido (los píxeles con valor 0 se cambian a 1 y vice versa). A partir de los mapas booleanos se generan mapas de atención $A_i$, donde se rellenan con el valor 1 regiones cerradas por los bordes extraídos. La lógica que sigue esta etapa es que una región es prominente en la medida que esté completamente cerrada y separada del fondo. Esto se realiza tanto para el mapa $B_i$ como para su versión invertida, utilizando el algoritmo *flood fill*.

Cada mapa de atención se normaliza para que áreas pequeñas y de alta prominencia se mantengan cuando se combine con los demás. Por otro lado, es deseable penalizar mapas de atención que generan varias regiones prominentes pequeñas y dispersas, para ésto, se aplica una dilación a través un *kernel* de tamaño $\omega_{d1} = 7$. Esto ensancha las regiones pequeñas, fomentando la agrupación de éstas en regiones más grandes y menos dispersas. Para obtener el mapa de atención final $\bar{A}$, se promedian todos los mapas $A_i$ obtenidos. Se realiza sobre $\bar{A}$ una última etapa de post procesamiento consistente en una dilación de tamaño $\omega_{d2} = 23$, para luego difuminar con un filtro gaussiano de desviación estándar $\sigma = 20$ pixeles. El resultado final es el mapa de prominencia $S$, que cumple con las mismas características del mapa Itti & Koch.

### 2.3.3. Regiones Contextualmente Relevantes

Uno de los objetivos de este trabajo es evaluar la importancia de la atención *top down* a lo largo del tiempo durante experimentos de visión libre. De acuerdo a lo presentado en la sección de modelos *top down*, se asume que este componente influye más (de manera general) en la localización de la mirada que el componente *bottom up*. En consecuencia, regiones relacionadas a la atención *top down* deberían estar caracterizadas por una mayor cantidad de fijaciones y representar partes de la imagen que contienen estímulos sociales o que portan información relevante sobre el contexto de la escena observada. Con esto en mente, la propuesta para representar la información *top down* es utilizar la distribución empírica de las fijaciones realizadas por los sujetos para definir "Regiones Contextualmente Relevantes".

Las "Regiones Contextualmente Relevantes" se definen tomando en cuenta todas las fijaciones realizadas sobre la imagen a estudiar. El procedimiento seleccionará aquellas regiones donde se ha registrado una mayor cantidad de fijaciones. Como se explicó inicialmente, las

partes de la imagen que se obtengan a partir de este cálculo deberían coincidir con elementos o regiones que en la literatura se han relacionado al componente *top down*; caras y cuerpos de humanos y/o animales, así como objetos que porten información relevante sobre el contexto de la escena observada. El detalle de este proceso y su validación se explican en el capítulo Materiales y Métodos.

## 2.4. Métricas para modelos de atención visual

Para evaluar la evolución de la atención visual en el tiempo, se realizará una comparación entre conjuntos de fijaciones realizadas por los sujetos y modelos que representen los componentes de la atención visual *bottom up* y *top down*. El objetivo no es identificar el mejor modelo en función de su capacidad para predecir la ubicación de las fijaciones, si no representar dos componentes de la atención visual (*bottom up* y *top down*) y evaluar, para ciertas ventanas de tiempo, qué tanto se asemejan a la distribución de fijaciones realizadas por los sujetos. En la literatura existe una gran cantidad de métricas para medir similitud o distancia entre la distribución real de fijaciones y un modelo de prominencia, habiendo una buena variedad en la cuantificación e interpretación de esta medida. A continuación se nombran brevemente las 8 métricas más utilizadas en la literatura para evaluación de modelos de prominencia [46], éstas se presentan resumidas en la tabla 2.1. A partir de éstas, se seleccionan dos métricas en función de las necesidades y limitaciones del análisis a realizar.

Tabla 2.1: Métricas disponibles para la evaluación de modelos de prominencia

| **Métrica** | **Referencia** |
|---|---|
| Area under ROC Curve (AUC) | [47] |
| Shuffled AUC (sAUC) | [48] |
| Normalized Scanpath Saliency (NSS) | [49] |
| Information Gain (IG) | [50] |
| Coeficiente de Correlación de Pearson (CC) | [51] |
| Divergencia Kullback-Leibler (KL) | [52] |
| Similarity or histogram intersection (SIM) | [53] |
| Earth Mover's Distance (EMD) | [54] |

De acuerdo a lo explicado, la idea es comparar los modelos de atención con las fijaciones reales, separadas según el tiempo en que ocurren. Al no ser posible asegurar la misma cantidad de datos en todas las ventanas, una evaluación basada solamente en la ubicación de las fijaciones sesgaría los resultados, penalizando más duramente ventanas con menor cantidad de eventos. Por lo anterior, las ubicaciones obtenidas en cada ventana se utilizarán para generar distribuciones de probabilidad con las cuales comparar las predicciones de los modelos, esto permite sortear la diferencia en cantidad de datos por ventana.

Se descartan desde un principio 4 de las 8 métricas, debido a que evalúan los modelos como clasificadores de las ubicaciones exactas de cada fijación: miden qué tan correctamente el modelo predice la probabilidad de fijación de cada pixel. Estas métricas son *Area Under ROC Curve* (AUC), *shuffled AUC* (sAUC), *Normalized Scanpath Saliency* (NSS) e *Infor-*

*mation Gain* (IG). Se descarta también la Correlación de Pearson (CC) ya que compara de manera global la predicción y los datos empíricos, dificultando una interpretación detallada sobre la asignación de fijaciones en la escena.

Las 3 métricas restantes cumplen con las necesidades de este trabajo, sin embargo, se seleccionan solamente las métricas de Intersección de Histogramas (SIM) y *Earth Mover's Distance* (EMD). SIM permite interpretar directamente el nivel de traslape entre los mapas de prominencia y las fijaciones reales, siendo además una métrica acotada entre los valores 0 y 1. En contraste, EMD es una métrica no acotada que toma en cuenta, a diferencia de la mayoría de las otras, la distancia entre las ubicaciones predichas y aquellas realmente fijadas. Utilizar ambas métricas posibilita una mejor interpretación de los resultados, especialmente a través del contraste entre ambas. La Divergencia Kullbak-Leibler (KL) se excluye del análisis ya que funciona de manera similar a EMD, sin la ventaja de tomar directamente en cuenta la distancia de las predicciones. La implementación de las métricas seleccionadas se detalla a continuación.

### 2.4.1. Intersección de Histogramas (SIM)

Intersección de Histogramas es una medida de similitud acotada, compara dos distribuciones como si fueran histogramas y define esta similitud a partir de la intersección entre los *bins* de cada una. El puntaje se obtiene sumando el valor mínimo entre las dos distribuciones, evaluado en cada píxel o *bin*.

Dado un mapa de prominencia $P$ y una distribución de fijaciones $Q$:

$$SIM(P,Q) = \sum_i \min(P_i, Q_i) \tag{2.11}$$

$$\text{Donde } \sum_i P_i = \sum_i Q_i = 1 \tag{2.12}$$

SIM es una métrica acotada, toma valor 1 cuando ambas distribuciones son idénticas, mientras que un valor de 0 indica que no existe traslape entre ellas. Por su formulación, SIM es muy sensible a los falsos negativos, penalizando fuertemente cuando una distribución de referencia tiene probabilidad en lugares distintos a la otra, sin importar la distancia entre ambos.

### 2.4.2. Earth Mover's Distance (EMD)

Earth Mover's Distance (en español distancia del transportador de tierra) es una métrica que mide la diferencia entre dos distribuciones de probabilidad a partir de su distancia espacial. Como su nombre indica, la métrica representa estas distribuciones como acumulaciones de tierra y calcula el costo óptimo de mover la masa de una para obtener la forma de la otra. Al estar basada en la teoría de transporte óptimo [55], tiene la ventaja de incorporar la distancia entre las masas de probabilidad de cada distribución. Incorporar esta distancia entrega una evaluación más amplia de los modelos y más robusta frente a variaciones o ruidos en el registro de los datos empíricos.

Formalmente, EMD representa el costo total de transformar una distribución en otra. Este costo está conformado por la cantidad de masa de probabilidad que se mueve y la distancia

que ésta recorre. Debido a que existen infinitas formas de transformar una distribución para calzarla con otra, el cómputo de EMD requiere encontrar la forma óptima en que este procedimiento se puede realizar y, por lo tanto, resolver el siguiente problema de minimización:

$$\widehat{EMD}(P,Q) = \min_{f_{ij}} \sum_{i,j} f_{ij} d_{ij} \tag{2.13}$$

Sujeto a:

$$f_{ij} \geq 0 \tag{2.14}$$

$$\sum_{j} f_{ij} \leq P_i \tag{2.15}$$

$$\sum_{i} f_{ij} \leq Q_i \tag{2.16}$$

$$\sum_{ij} f_{ij} = \min(\sum_{i} P_i, \sum_{j} Q_j) \tag{2.17}$$

Donde $f_{ij}$ representa la cantidad de masa transportada desde la distribución $P$ a la distribución $Q$ y $d_{ij}$ es la distancia recorrida en este movimiento. La ecuación 2.13 minimiza el movimiento total de masa, moviendo sólo de $P$ a $Q$ (restricción 2.14). Las restricciones 2.15 y 2.16 evitan que se retire más masa de la que hay en $P$ y que se deposite más masa de la que se puede en $Q$ respectivamente. 2.17 fuerza a que el total de masa movida no supere la densidad total encontrada ya sea en $P$ como en $Q$.

Al ser una medida de distancia, su interpretación se basa en que, a mayores valores de EMD, más disimiles son las distribuciones comparadas, indicando la existencia de dos distribuciones idénticas cuando alcanza el valor cero. Es importante notar que, al requerir al resolución un problema de optimización global, el cálculo de esta métrica es computacionalmente intensivo. Además, debido a su formulación, EMD es una métrica no acotada. Sin embargo, debido a que las comparaciones se realizan sobre un mismo modelo a lo largo del tiempo y que la información de interés es la evolución de dicho puntaje, esta característica en especial no presenta un problema para el análisis realizado.

# Capítulo 3

# Materiales y métodos

## 3.1. Datos utilizados

### 3.1.1. Participantes

Los datos utilizados en este trabajo provienen de experimentos realizados para estudiar potenciales evocados (ERP) durante exploración libre de escenas visuales [56]. Un total de 16 sujetos participaron en el experimento (5 mujeres); 13 eran diestros, 8 tenían dominancia ocular derecha y su edad fue en promedio de 31.41 $\pm$ 7.17 años. Todos poseían visión normal o corregida a normal. Todos los sujetos aceptaron participar de forma voluntaria a través de un consentimiento informado escrito; el formulario de consentimiento y todos los protocolos experimentales fueron aprobados por el Comité de Ética para Investigación en Humanos de la Facultad de Medicina de la Universidad de Chile. Los registros se realizaron en un período de 32 días.

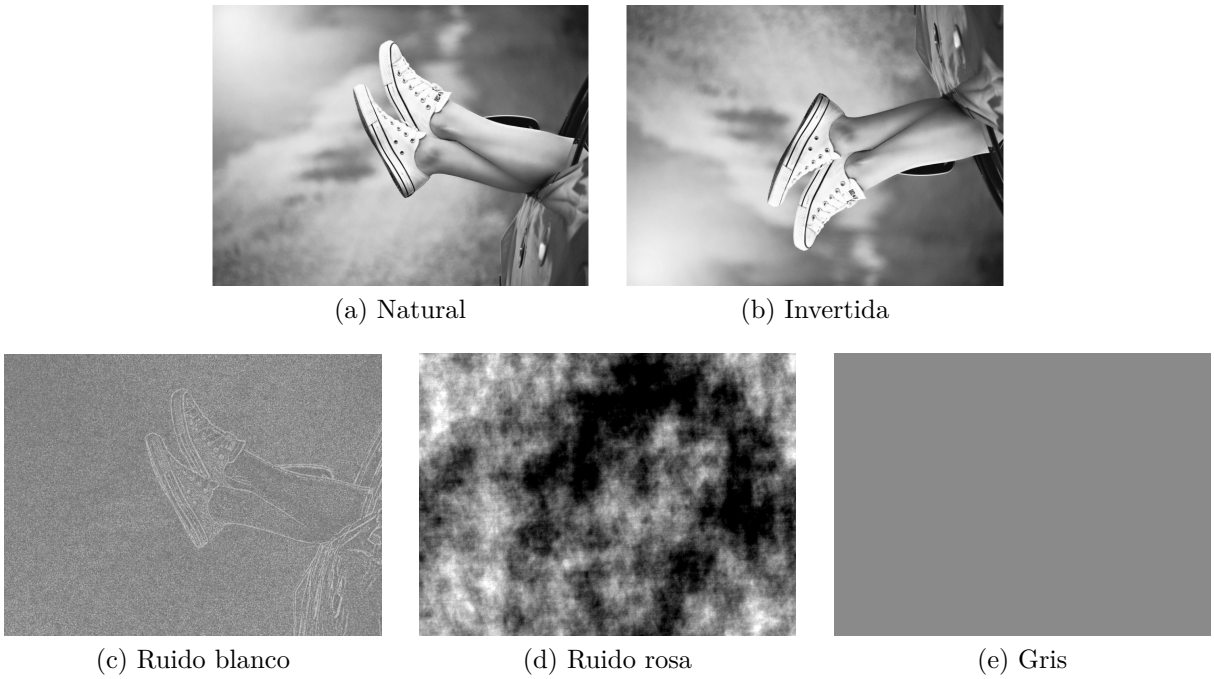
### 3.1.2. Descripción del experimento

Un total de 46 imágenes pertenecientes al Sistema Internacional de Imágenes Afectivas (IAPS) [57] fueron seleccionadas basándose en su valencia y valor de *arousal*, estando todas en el rango de valores medios de la base de datos (valencia media 6.62 $\pm$ 1.09; *arousal* medio 4.08 $\pm$ 0.96). El concepto de valencia hace referencia a la atracción (valencia positiva) o rechazo (valencia negativa) que puede producir un estímulo. Por otro lado, el *arousal* representa la intensidad con la que estas emociones puede producirse. En este caso, se seleccionaron imágenes que no provocaran reacciones emocionales fuertes hacia ninguna de las dos valencias. La luminancia de las imágenes seleccionadas fue corregida para su presentación en la pantalla utilizada. Este conjunto inicial se denomina *imágenes naturales* y para cada imagen natural se construyeron 4 imágenes de control: ruido rosa, ruido blanco, invertida y gris.

Las imágenes de ruido rosa y blanco se crearon considerando 2 parámetros de las imágenes naturales: su densidad espectral y su fase. Las imágenes de ruido rosa mantienen el mismo espectro de frecuencia que su contraparte natural, con la fase consistente en una versión mezclada de la original. Esto produce imágenes con manchas en distintas tonalidades de gris, donde no se puede discernir figuras u objetos de manera clara. Por otro lado, las imágenes ruido blanco se crearon aplanando la densidad espectral de la imagen natural original y manteniendo su contenido de fase. Realizar esta operación produce que los bordes de las figuras sobresalgan por sobre el resto, se pierde información de la imagen pero se conservan

las figuras generales a través de sus contornos (ver figura 3.1).

Las imágenes invertidas se crearon volteando las imágenes naturales con respecto al eje horizontal, quedando boca abajo. Las imágenes grises son imágenes planas (sin ningún tipo de contenido) que mantienen la misma luminancia promedio de sus contrapartes naturales. Finalmente, las imágenes blancas y negras se crearon como control de luminancia, con valores RGB máximos ([255, 255, 255]) y mínimos ([0, 0, 0]) en todos sus pixeles, respectivamente. Ejemplos de las modificaciones realizadas sobre las imágenes naturales se muestran en la figura 3.1.



(a) Natural (b) Invertida

(c) Ruido blanco (d) Ruido rosa (e) Gris

Figura 3.1: Ejemplo de modificaciones sobre las imágenes utilizadas. Imagen referencial, no corresponde a imágenes utilizadas en los experimentos.

Para el experimento, los sujetos se acomodaron de tal forma que sus ojos quedaran a 70 cm de una pantalla de tamaño correspondiente a 1920x1080 píxeles (ancho x alto) y la media de 32 pixeles por centímetro resulta en una equivalencia de 39.38 píxeles por ángulo visual. Las imágenes se presentaron con un tamaño de 1024 x 768 píxeles al centro de la pantalla. Los bordes sobrantes tenían un color gris plano ([173, 173, 173] RGB). Para la tarea principal se instruyó a los sujetos que exploraran las imágenes presentadas de manera libre, se les indicó también que habrían preguntas sobre su contenido al final del experimento. Se presentó un total de 322 imágenes, divididas en 2 bloques.

## 3.2. Preprocesamiento

La posición de la mirada y el diámetro pupilar se registraron utilizando un sistema de seguimiento ocular a 500 Hz (EyeLink 1000, SR-Research, ON, Canada). Sacadas y fijaciones se detectaron automáticamente a partir del algoritmo SaFiDe [58], clasificando sacadas como movimientos oculares mayores a $0.1^{o}$ visual, con una velocidad mínima de $30^{o}s^{-1}$ y una aceleración mínima de $4000^{o}s^{-2}$, ambas mantenidas durante al menos 4 ms. Pestañeos

ocurridos durante los experimentos se definieron a partir de la ausencia de datos pupilares. Debido al movimiento del ojo durante un pestañeo, el algoritmo utilizado detecta el inicio y final de estos eventos como una sacada, dichos datos fueron descartados del análisis. Sujetos con más del 10 % del tiempo total de experimento en pestañeo fueron descartados (3 sujetos). Finalmente, el tiempo donde no ocurren pestañeos ni sacadas se considera como fijación. De éstas, aquellas ubicadas fuera de la imagen presentada y con duración menor a 50 ms o mayor a 1 s, fueron excluidas del análisis. En total, se analizaron 45.507 fijaciones realizadas por 13 sujetos.

## 3.3. Test estadísticos

En varios de los análisis realizados se aplicarán tests para evaluar si las diferencias entre dos distribuciones son estadísticamente significativas. En casos donde los datos se puedan emparejar, se aplicarán test-t emparejados y cuando la cantidad de datos entre una distribución y otra no coincida, se realizarán comparaciones a través de las medianas calculadas a lo largo de los sujetos o de las imágenes, según corresponda.

### 3.3.1. Test-t pareado

El test t pareado o t-test para muestras emparejadas se utiliza para comparar las medias de dos distribuciones de datos relacionadas, se aplica cuando se tienen varios sujetos o variables que han sido medidos más de una vez. El objetivo es comparar las distribuciones de las diferentes medidas para evaluar un efecto o cambio en el tiempo. El test evalúa la probabilidad de una hipótesis nula, es decir, que las medias de las muestras emparejadas son iguales. En detalle, el procedimiento consiste en:

1. Para cada sujeto o variable se calcula la diferencia entre la primera y la segunda medición, generando una distribución a partir de todas las diferencias.

2. A la distribución obtenida se le calcula la media y la desviación estándar.

3. Con esta información se calcula el estadístico t a través de la siguiente ecuación:

$$t = \frac{\bar{x}}{\sigma \cdot \sqrt{\frac{1}{n}}} \tag{3.1}$$

   Donde $\bar{x}$ representa el promedio de la distribución, $\sigma$ su desviación estándar y $n$ el tamaño de la muestra.

4. Se define una distribución *t-student*, de acuerdo a los grados de libertad de la muestra, correspondientes a $n - 1$.

5. El nivel de significancia se determina a través de la distribución definida, evaluando la probabilidad de que el estadístico $t$ adopte el valor calculado.

6. Esta probabilidad indicará la significancia estadística de las diferencias. En la mayor parte de los casos, una probabilidad menor a 0.05 se considera estadísticamente significativa o, en otras palabras, que la hipótesis nula (medias son iguales) se puede rechazar.

Un test pareado requiere que la distribución de las diferencias corresponda a una distribución normal y que las muestras sean independientes. Estos supuestos se evaluarán en cada aplicación de los tests a través de un test de normalidad:

### 3.3.2. Test de normalidad Shapiro-Wilk

El test Shapiro y Wilk se utiliza para verificar la probabilidad de que un conjunto de mediciones haya sido muestreado de una distribución gaussiana. Se realiza a partir del cálculo de la *skewness* (sesgo) y kurtosis (grosor de las colas) de la distribución y requiere una cantidad de datos grande, en general mayor a 20. En detalle, se evalúa la probabilidad de la hipótesis nula, la que afirma que la muestra proviene de una distribución normal. Para obtener esta probabilidad se debe:

1. Ordenar datos de mayor a menor
2. Se calculan los coeficientes $a_i$:

$$(a_1, ..., a_n) = \frac{m^T V^{-1}}{\sqrt{(m^T V^{-1} V^{-1} m)}} \tag{3.2}$$

   Donde $m$ es un vector con $n$ medias de muestras obtenidas de distribuciones normales y $V$ es su correspondiente matriz de covarianza.

3. Cálculo del estadístico $W$:

$$W = \frac{(\sum_{i=1}^{n} a_i x_i)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{3.3}$$

   Donde $n$, $\bar{x}$ y $x_i$ corresponden al tamaño, la media y cada valor de la muestra.

4. $W$ puede tomar valores entre 0 y 1 y se analiza como un valor p.
5. En la mayor parte de los casos, una probabilidad menor a 0.05 se considera estadísticamente significativa o, en otras palabras, que la hipótesis nula (datos vienen de una distribución normal) se puede rechazar.

## 3.4. Análisis exploratorio

### 3.4.1. Tendencias sobre fijaciones y sacadas

De acuerdo a lo comentado en la sección 2.2.2, se ha observado en gran parte de la literatura que las fijaciones tienden a durar más a medida que pasa el tiempo del experimento y que, simultáneamente, las sacadas van disminuyendo su amplitud. Con el fin de verificar este fenómeno, se estudia la evolución de ambos eventos en el tiempo, evaluando las diferencias en duración o amplitud según el orden de ocurrencia de las fijaciones y sacadas, respectivamente.

La evolución en el tiempo se estudia agrupando fijaciones y sacadas según su orden de ocurrencia. Para cuantificar la significancia de las diferencias en duración o amplitud se realizan test-t pareados de la siguiente manera: en cada imagen y para cada uno de los conjuntos ordenados se calcula la mediana de la duración/amplitud a lo largo de los sujetos. El test

pareado se realiza entre las medianas de cada conjunto y el conjunto que representa la fijación/sacada siguiente. Este procedimiento se aplicará para estudiar y comparar la evolución en ambas variables frente a los distintos tipos de imagen utilizados (ver figura 3.1).

### 3.4.2. Tiempos de reacción

En los experimentos realizados por Flechsenhar et al. [39] se mostró que los sujetos miraban de manera inmediata y con mayor preferencia lugares correspondientes a caras humanas. En dicho estudio se hipotetizó que existe una región del cerebro encargada de la detección rápida de caras u otras señales sociales, la que permitiría prestarles atención velozmente cuando están presentes. En línea con esta idea, se estudia el efecto del contenido presente en cada imagen sobre el tiempo de reacción, es decir, el tiempo que demoran los sujetos en realizar la primera sacada desde la aparición del estímulo. La idea detrás de esto es que en imágenes donde es más fácil extraer un contexto debería ser más fácil definir un patrón de exploración y, por lo tanto, iniciar los movimientos de manera más expedita. En casos donde se pueden emparejar los datos uno a uno (e.g. imágenes naturales contra sus versiones invertidas o contra sus versiones ruido rosa) se realiza un test-t pareado, comparando el tiempo de reacción de todos los sujetos frente a cada imagen y su contraparte.

Además de las nombradas, algunas de las comparaciones se hacen entre conjuntos con distinta cantidad de imágenes. Por ejemplo, evaluar el efecto que tiene la existencia de contexto sobre el tiempo de reacción implica comparar imágenes naturales, invertidas y ruido blanco (donde es posible discernir objetos y figuras) con imágenes ruido rosa (consistentes en manchas grises). Debido a que los conjuntos de imágenes con contexto tienen 3 veces más imágenes que el conjunto de ruido rosa, es necesario agrupar estos datos de una manera que se puedan comparar. Para resolver esto, se calcula la mediana de los tiempos de reacción en cada sujeto, separando por contexto. Siguiendo el ejemplo expuesto, se calcula la mediana del tiempo de reacción a lo largo de las imágenes naturales, invertidas y ruido blanco para cada sujeto y se compara con la mediana de cada sujeto sobre las imágenes ruido rosa. Esta separación permite realizar un test-t pareado sobre las medianas de los sujetos entre cada contexto a analizar. Las otras dos comparaciones involucran imágenes planas e imágenes con o sin caras.

## 3.5. Distribuciones de probabilidad sobre las fijaciones

Una forma de estudiar la asignación de atención visual sobre una escena es a través de la ubicación de las fijaciones: se asume que una región fijada está siendo atendida por el observador. En presente trabajo se seguirá este lineamiento y se estudiará la posición de las fijaciones como indicador de dónde se encuentra la atención visual en cada momento de los experimentos. Para comparar las asignaciones de atención con respecto a los modelos de atención utilizados, se generarán distribuciones de probabilidad a partir de las fijaciones registradas, donde una mayor probabilidad de fijación representará a su vez una mayor atención, en promedio, dedicada a la región correspondiente por los sujetos o a lo largo del tiempo.

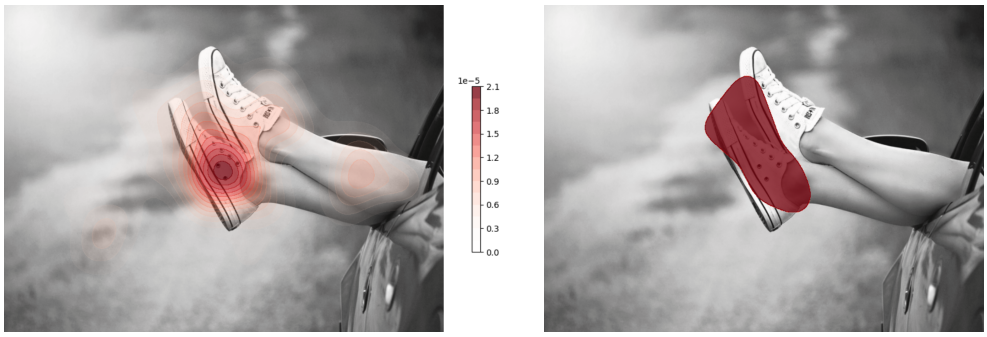
Por lo anterior, y de acuerdo a la práctica común [59], para cada conjunto de fijaciones a estudiar se generarán distribuciones de probabilidad aplicando una convolución entre cada fijación (ubicada en las coordenadas X e Y) y una distribución gaussiana cuya desviación

estándar corresponde a un ángulo visual. Luego de sumar las fijaciones, la matriz obtenida se estandariza para obtener una distribución de probabilidad en 2 dimensiones. De esta forma, las regiones donde se registra mayor cantidad de fijaciones tendrán asignado un mayor nivel de atención visual.

La convolución con una distribución gaussiana cumple la función de incorporar el campo visual en la representación de fijaciones. Es necesario este procedimiento ya que, al registrar, el sistema de seguimiento visual entrega una posición puntual de donde ocurrió cada fijación. De acuerdo a lo explicado en la sección 2.1, cuando se fija la mirada en una región, la luz que ingresa a nuestros ojos es captada en mayor resolución por la fóvea, la que representa un espacio significativamente mayor a un punto o pixel. Para tomar en cuenta este campo receptivo, se expande el punto donde ocurre la fijación con una distribución de tamaño similar a dicha parte de la retina.

## 3.6. Regiones Contextualmente Relevantes (RCR's)

De acuerdo a lo explicado en la Sección 2.3.3, las RCR's se calculan para cada imagen a partir de la distribución empírica de fijaciones, a lo largo de los sujetos e independiente del tiempo, es decir, se toman todas las fijaciones realizadas por todos los sujetos sobre dicha imagen. A partir de la distribución de probabilidad generada en este procedimiento y de acuerdo a lo explicado en la sección 2.2.3.2, se debe seleccionar la región de la imagen con mayor probabilidad de fijación. Ésta representaría el objeto o región que evoca la atención del observador por razones más allá de sus características físicas. Un ejemplo de las RCR obtenidas se muestra en la figura 3.2.




(a) Distribución Empírica de fijaciones

(b) Región Contextualmente Relevante

Figura 3.2: Cálculo de Regiones Contextualmente Relevante. Imagen referencial, no corresponde a imágenes utilizadas en los experimentos.

Para obtener la RCR de cada imagen, se selecciona un área correspondiente al 5 % de su tamaño total donde hay mayor probabilidad de fijación. Esto se calcula tomando los pixeles que pertenecen al percentil 95 de la distribución empírica de fijaciones. El tamaño de la RCR es el mismo para todas las imágenes.

### 3.6.1. Validación de las RCR's

De acuerdo a lo explicado en la sección 2.3.3, las RCR's deberían coincidir con objetos o regiones relacionadas a cuerpos y caras animales (incluyendo humanos), así como cualquier

elemento que aporte a dar contexto a la escena observada. Con el fin de validar las RCR's en función de su coincidencia con este tipo de regiones, se aplicará un modelo para segmentación visual llamado *Segment Anything Model* (SAM) [60]. El modelo es capaz de segmentar de manera automática cualquier imagen que se le entregue, identificando una gran cantidad de objetos y entidades en ella. El estudio se realiza de la siguiente manera:

1. Se seleccionan imágenes naturales que contengan caras y cuerpos de animales y humanos, así como productos humanos: prendas de vestir, construcciones, medios de transporte, etc. Este conjunto contiene 25 de las 46 imágenes naturales utilizadas en este trabajo.
2. Para cada imagen se obtiene un conjunto de segmentos a través del modelo SAM.
3. En paralelo, se separan las fijaciones realizadas sobre la imagen entre aquellas que pertenecen a una RCR y aquellas que no.
4. Para cada segmento, se contabiliza la cantidad de fijaciones ocurridas en su interior.
5. Los segmentos se clasifican a mano, generando 48 tipos de regiones, separadas en 5 clases:
    - **caras y rasgos:** nariz, colmillo, oreja, ojo, trompa, cara, boca.
    - **cuerpos y miembros:** pata, piernas, cuerpo, ala, mano.
    - **vegetación y paisaje:** árbol, verdura, flor, rama, césped, arbustos.
    - **productos humanos:** globo, canasto, puerta, barco, letra, chaqueta, construcción, sombrilla, casa, ancla, calzado, ventana, antena, techo, calle, banca, muro, cuerdas, pared, peldaños.
    - **fuera de foco:** regiones desenfocadas o donde no se entiende lo que hay.
    - **paisaje:** cielo, mar, acantilado, nube, tornado, lago, montañas
    - Se reporta la proporción de fijaciones que cae en cada tipo de objeto o región, comparando aquellas que pertenecen a una RCR y aquellas que no.

## 3.7. Procedimiento para la comparación entre fijaciones y modelos de atención

En este trabajo se estudia la evolución de la atención visual en el tiempo, comparando dos de sus componentes: *bottom up* y *top down*. En esta línea, para todas las imágenes se obtiene una representación de cada componente a partir de su modelo computacional (BMS, *Itti & Koch* o RCR). Estos modelos entregan mapas en dos dimensiones que serán comparados con las distribuciones de probabilidad de fijación empíricas. La evolución en el tiempo se obtiene separando las fijaciones realizadas sobre cada imagen según su orden de ocurrencia (primera fijación, segunda, tercera, etc.), desde la primera hasta la doceava fijación, incluyendo a todos los sujetos en cada conjunto.

Para cada imagen se definen doce conjuntos de fijaciones, los que se comparan con los modelos de atención. La comparación se realiza a través de las métricas explicadas en la sección 2.4, primero calculando distribuciones de probabilidad partir de cada conjunto de fijaciones

y luego comparando dichas distribuciones con los modelos de atención visual nombrados. El objetivo de obtener estas medidas es analizar el cambio en el peso de cada componente entre una fijación y la siguiente. En esta línea, no es de interés el puntaje de un modelo o qué tan bien predice la localización de las fijaciones en general; lo interesante es cómo cambia este puntaje en el tiempo y qué representa esto en función de las prioridades atencionales de los observadores a medida que se desarrolla cada experimento. En la figura 3.3 se muestra un diagrama explicativo del proceso.

La significancia de los cambios se obtiene a través de un test de hipótesis pareado (explicado en 3.3.1), donde se compara la diferencia de puntaje entre un orden de fijación y el siguiente a lo largo de todas las imágenes. En este caso las mediciones son los puntajes en cada imagen y las muestras sucesivas corresponden a los conjuntos de fijaciones ordenados según su ocurrencia.

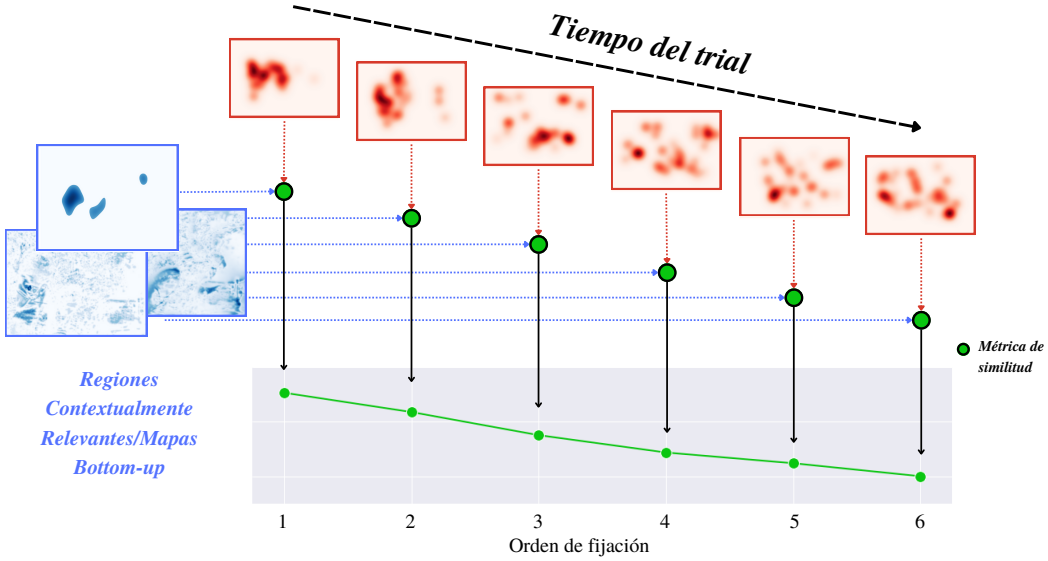



Figura 3.3: Análisis sobre la evolución de atención visual a lo largo de los experimentos. En azul, el modelo a comparar (modelos *bottom up* o RCR). En rojo, las distribuciones empíricas a partir de las fijaciones según su orden de ocurrencia. Nodos verdes representan la aplicación de las métricas EMD o SIM para comparar similitud entre las distribuciones empíricas y los modelos de atención.

Se realiza también una comparación entre los distintos tipos de imagen, principalmente sobre cómo cambia el comportamiento entre imágenes planas (sin información), imágenes ruido rosa (sólo contenido de texturas, sin información de contexto) e imágenes con contenido contextual (naturales, invertidas y versiones ruido blanco). En este sentido, durante el análisis se agrupa las imágenes según la modificación que las genera: imágenes naturales, ruido rosa, ruido blanco, invertidas y planas (grises, blancas y negras). Se espera que la diferencia en tipo de contenido implique cambios de comportamiento entre un conjunto de imágenes y otro.

## 3.8. Tiempos de los cambios en estrategias de exploración

De acuerdo al objetivo 5 de la tesis, se estudian las dinámicas de exploración ocular para identificar un tiempo específico en que se observen cambios en las tendencias observadas. Ya que el orden de fijaciones no entrega una medida directa del tiempo en que se observan las dinámicas, en caso de identificarse cambios relacionados al orden de ocurrencia se calcularán las distribuciones de tiempo sobre el inicio de la fijación o sacada correspondiente para definir un umbral temporal aproximado en este cambio.

## 3.9. Implementación y librerías

Los análisis e implementaciones de modelos se realizaron en *Visual Studio Code*[1] utilizando el lenguaje *Python*. Para los modelos de atención *bottom up* (*Itti & Koch* y BMS), así como para la métrica de similitud de histogramas (SIM) se utilizó la librería *pysaliency*[2], que provee una interfaz unificada de algoritmos y métricas para la investigación de modelos de prominencia visual. Para la métrica de distancia de movimiento de tierra (EMD) se utiliza la librería *Python Optimal Transport*[3] [61]. Esta librería contiene una gran variedad de algoritmos para la resolución de problemas de optimización relacionados al Transporte Óptimo. Ambas librerías son de código abierto. Finalmente, para los test estadísticos se utilizan las funciones *ttest_rel*, *shapiro*, y *mannwhitneyu* de la librería *SciPy*[4].

[1] https://code.visualstudio.com/
[2] https://github.com/matthias-k/pysaliency
[3] https://pythonot.github.io/
[4] https://scipy.org/

# Capítulo 4

# Resultados y Discusiones

## 4.1. Tendencias sobre fijaciones y sacadas

De acuerdo a lo explicado en la sección 3.4.1, se estudia la existencia de tendencias sobre la duración de las fijaciones y la amplitud de las sacadas a medida que avanza el tiempo de los experimentos. Esto se muestra en las figuras 4.1 y 4.2, donde se presenta el comportamiento de las fijaciones y de las sacadas, respectivamente. En cada orden de ocurrencia se calcula la mediana de las duraciones (para las fijaciones) o las amplitudes (para las sacadas). Las bandas representan un intervalo de confianza del 95 % sobre la estimación del estadístico (marcador circular). Diferencias significativas entre un conjunto y el siguiente se muestran como: $>$ o $<$, y * ($p$<0.05); ** ($p$<0.01); *** ($p$<0.001).

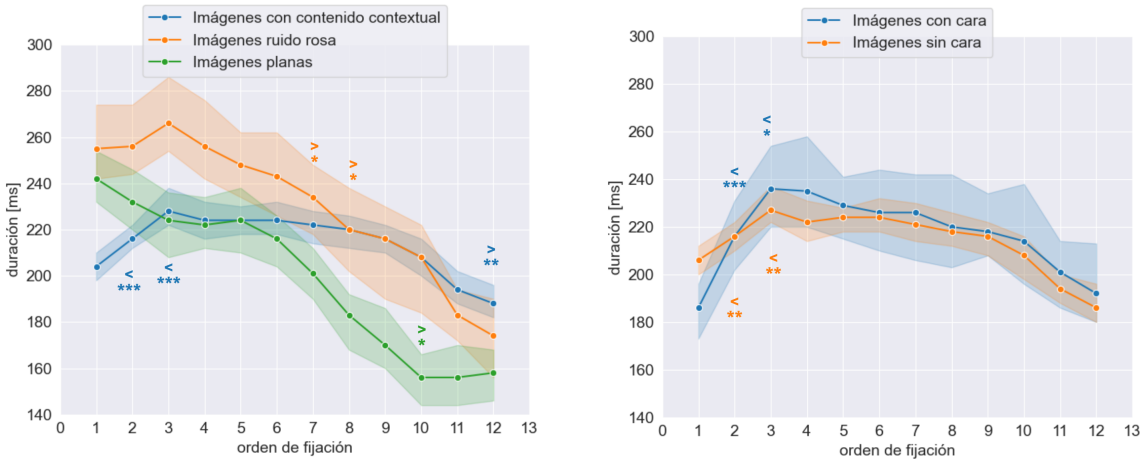
Se evaluó esta evolución para experimentos donde se presentaron imágenes que muestran distintos niveles de información: imágenes con contenido contextual (naturales, invertidas y versiones ruido blanco), imágenes ruido rosa (manchas grises) e imágenes planas. Además, se compara el comportamiento entre experimentos donde se usaron imágenes que contenían caras y experimentos donde las imágenes, teniendo contenido contextual, no las incluyen (se omiten imágenes planas y ruido rosa).

### 4.1.1. Duración de fijaciones

En la figura 4.1 se grafica la duración de las fijaciones en función de su orden de ocurrencia. Se compara el comportamiento en el tiempo entre experimentos donde se mostraron imágenes planas, ruido rosa y con contenido contextual (figura 4.1.a). Lo primero a notar es que la tercera fijación coincide con un cambio de tendencia para los tres tipos de imágenes: ante imágenes con contenido contextual, se evidencia un aumento en la duración desde la primera hasta la tercera ocurrencia; se observa también que frente a imágenes ruido rosa se alcanza un máximo de duración en la tercera fijación, seguido por un descenso monotónico hasta el final del experimento; para imágenes planas se observa una disminución en la duración de fijaciones desde la primera ocurrencia, este decrecimiento se interrumpe en la tercera fijación y se retoma a partir de la sexta. Si bien se observa esta coincidencia de cambios en la tercera fijación, en el único escenario donde se observan cambios significativos es frente a imágenes con contenido contextual. Cambios significativos frente a las otras imágenes se encuentran de manera aislada y no representan un hallazgo interpretable, sin embargo, tanto para imágenes planas como ruido rosa, se observa una disminución progresiva en la duración

de las fijaciones hasta el final de los experimentos.

Una comparación entre experimentos donde se mostraron imágenes con o sin caras se presenta en la figura 4.1.b. Es posible observar que, al igual que para la figura 4.1.a, existe una tendencia al aumento de duración, que se mantiene hasta la tercera fijación realizada. Este cambio es más notorio cuando hay caras que cuando no las hay: en experimentos con caras los sujetos parten con fijaciones más cortas y llegan a producir fijaciones más largas para la tercera ocurrencia. Luego de esto, no existe una tendencia de interés. Los resultados obtenidos se diferencian de lo reportado en la literatura, ya que se ha observado en la mayoría que la tendencia de las fijaciones es a aumentar su duración. Es importante notar, sin embargo, que en los otros estudios todas las imágenes utilizadas presentaban información y contexto, por lo que el comportamiento frente a imágenes planas o ruido rosa es un escenario novedoso.




(a) (b)

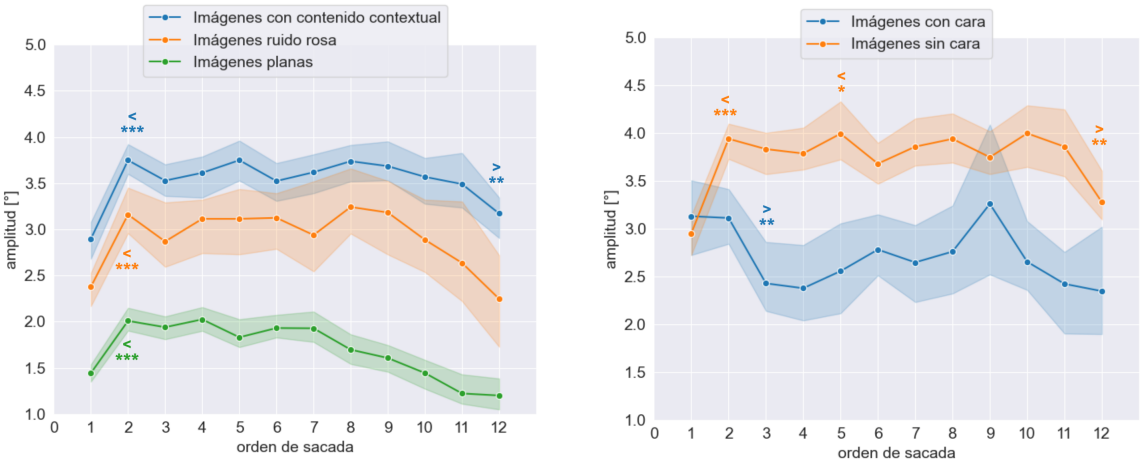
Figura 4.1: Evolución de la duración de fijaciones en función del orden de ocurrencia. A la izquierda se grafican experimentos que mostraron imágenes con contenido contextual, ruido rosa o imágenes planas; a la derecha, experimentos donde se usaron imágenes con caras e imágenes sin caras. Diferencias significativas entre un conjunto de fijaciones y el siguiente se muestran como: $>$ o $<$, y * ($p$<0.05); ** ($p$<0.01); *** ($p$<0.001)

### 4.1.2. Amplitud de sacadas

En la figura 4.2 se muestra la amplitud de las sacadas en función de su orden de ocurrencia. Se compara el comportamiento de las sacadas en el tiempo entre experimentos con imágenes planas, ruido rosa y con contenido contextual (figura 4.2.a). Las sacadas muestran una clara diferencia en amplitud, dependiendo del tipo de imagen que se presenta: si no hay contenido (imágenes planas) tienden a ser muy pequeñas, rondando entre 1 y 2 grados visuales; en contraste, frente a ruido rosa y contenido contextual, las sacadas son notoriamente mayores, tendiendo a ser más grandes cuando hay contenido contextual. Además del aumento entre la primera y la segunda sacada (observable para los tres tipos de imagen), no se observan tendencias claras o de interés.

De acuerdo a lo mostrado en la figura 4.2.b, se observa que la presencia de caras hace que

las sacadas disminuyan en amplitud entre la segunda y tercera ocurrencia. Fuera de esto, no se observan otras tendencias debido a la gran variabilidad de los datos y a que cambios significativos aparecen aislados. Cuando no se muestran caras, el comportamiento es muy similar a lo observado para imágenes con contenido contextual en la figura 4.2.a. Similar a lo observado para las fijaciones, los resultados obtenidos no se alinean con la literatura, donde se ha visto que las sacadas tienden a disminuir su amplitud en el tiempo. En este caso, las sacadas mantienen su amplitud en ciertos rangos, sin cambios significativos durante el experimento.




(a) (b)

Figura 4.2: Evolución de la amplitud de las sacadas en función del orden de ocurrencia. A la izquierda se grafican experimentos que mostraron imágenes con contenido contextual, ruido rosa o imágenes planas; a la derecha, experimentos donde se usaron imágenes con caras e imágenes sin caras. Diferencias significativas entre un conjunto de sacadas y el siguiente se muestran como: $>$ o $<$, y * ($p$<0.05); ** ($p$<0.01); *** ($p$<0.001)

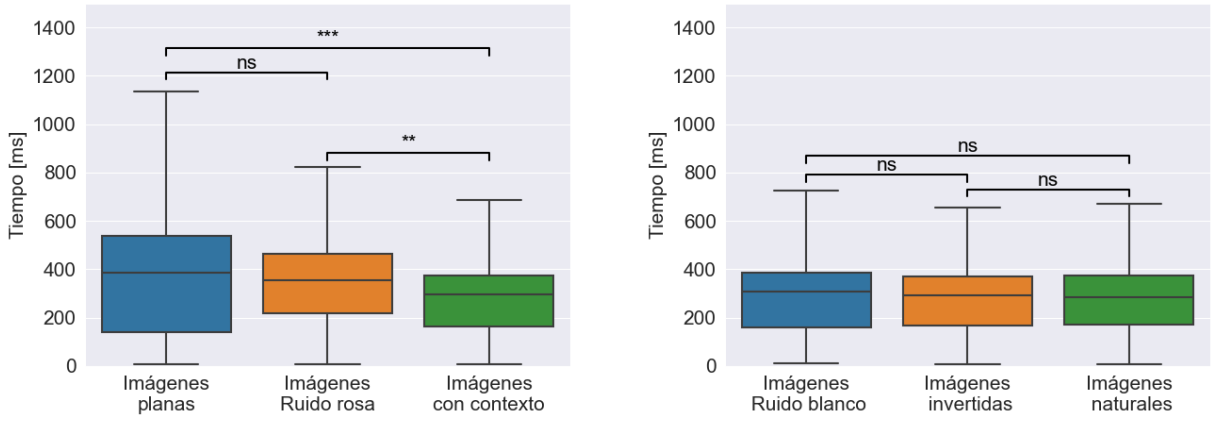
### 4.1.3. Discusión

Una gran cantidad de estudios sobre el comportamiento ocular durante exploración visual caracteriza su evolución a través de la duración de las fijaciones y la amplitud de las sacadas. Específicamente, en dichos experimentos se ha observado con frecuencia que la duración de las fijaciones tiende a aumentar con el tiempo, a la vez que la amplitud de las sacadas tiende a disminuir. Los resultados expuestos en esta sección no presentan estas tendencias de manera clara o con la misma intensidad que en la literatura, lo que puede deberse en parte a diferencias en la configuración de los experimentos: diferencia en la duración (otros experimentos en general tienden a ser más largos, llegando a ser mayores a 10 segundos), los tipos de imágenes que se presentan y las instrucciones entregadas a los sujetos. Se espera que, a través del análisis de los modelos de atención propuestos, se pueda evidenciar tendencias o cambios de estrategia de manera más clara durante la exploración ocular libre.

## 4.2. Tiempos de reacción

Siguiendo el procedimiento explicado en la sección 3.4.2, se comparan los tiempos de reacción de acuerdo al tipo de contenido presente en la escena mostrada. Se presenta, en las figuras 4.3 y 4.4, gráficos de caja representando las distribuciones en los tiempos de reacción. Diferencias significativas entre los conjuntos se calculan a través del test t explicado en la sección 3.3.1 y se reportan como * ($p$<0.05); ** ($p$<0.01); *** ($p$<0.001).

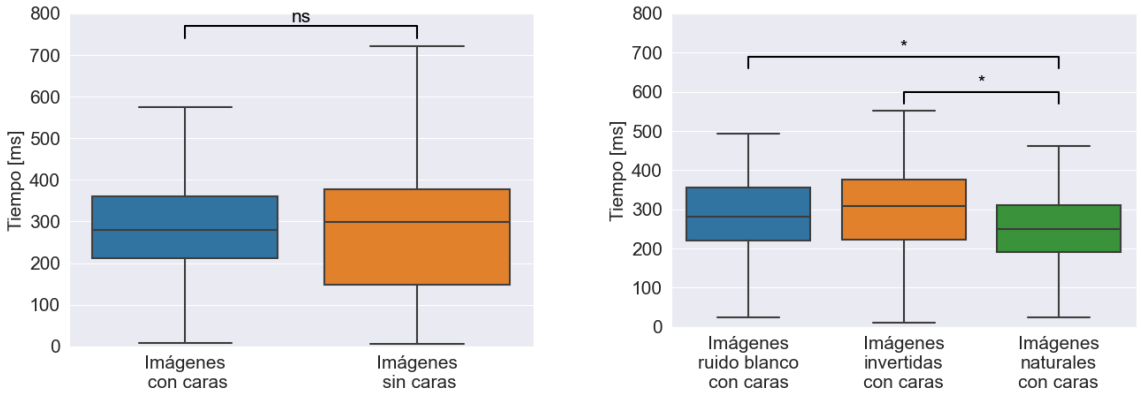
En la figura 4.3.a se observa que imágenes con contenido contextual se relacionan a tiempos de respuesta significativamente más cortos que los observados cuando se presentan imágenes planas ($t(15) = 4.373$, $p = 0.000272$) o ruido rosa ($t(15) = 3.665$, $p = 0.00114$). Entre imágenes naturales y las dos modificaciones que mantienen el contexto de la escena (ruido blanco e invertidas) no se observan diferencias significativas sobre los tiempos de reacción, esto se presenta en la figura 4.3.b. Los resultados parecen indicar que a mayor cantidad de información, más rápida es la reacción de los sujetos.




(a) (b)

Figura 4.3: Tiempo de reacción (tiempo desde la presentación del estímulo hasta primera sacada) para cada trial. Se muestra la distribución de los tiempos de reacción [ms] comparando imágenes planas, ruido rosa y con contenido contextual (a); imágenes naturales, invertidas y con ruido blanco (b). Diferencias significativas entre los conjuntos se muestran como ns ($p$>0.05); * ($p$<0.05); ** ($p$<0.01); *** ($p$<0.001).

Entre imágenes con caras y sin caras tampoco se observa diferencias significativas ($t(15) = -0.912$, $p = 0.188$, figura 4.4.a). Finalmente, entre las imágenes con caras, se observan diferencias significativas en los tiempos de reacción para cada variante, habiendo reacciones más rápidas cuando se presentan imágenes naturales que cuando se muestran sus versiones ruido blanco ($t(7) = -2.015$, $p = 0.0231$) o sus versiones invertidas ($t(7) = -2.259$, $p = 0.0129$), esto se muestra en la figura 4.4.b.




(a) (b)

Figura 4.4: Tiempo desde presentación de estímulo hasta primera sacada para cada trial. Se muestra la distribución de los tiempos de reacción [ms] comparando imágenes con y sin cara (a); imágenes con cara naturales, invertidas y con ruido blanco (b). Diferencias significativas entre los conjuntos se muestran como ns ($p$>0.05); * ($p$<0.05); ** ($p$<0.01); *** ($p$<0.001).

### 4.2.1. Discusión

Los resultados de esta sección muestran que el tipo de información presente en la imagen (con contenido contextual, sólo texturas o vacío) afecta los tiempos de reacción de los sujetos. Se puede interpretar que a mayor nivel de contenido el tiempo de reacción disminuye debido a una mayor facilidad al momento de identificar regiones de interés físico o semántico. Lo observado para imágenes con caras, donde modificaciones a la imagen aumentan significativamente el tiempo de reacción, se puede relacionar con la evidencia de que seres humanos exhiben patrones de exploración específicos de exploración facial [62], los que se verían entorpecidos por el cambio de orientación (imágenes invertidas) o de frecuencias (imágenes ruido blanco).

## 4.3. Validación de las RCR's

De acuerdo al procedimiento descrito en la sección 3.6.1, se reporta en la figura 4.5 la proporción de fijaciones realizadas sobre los distintos tipos de objetos en las imágenes, separando aquellas que pertenencen a una Región Contextualmente Relevante (RCR) de las que no.

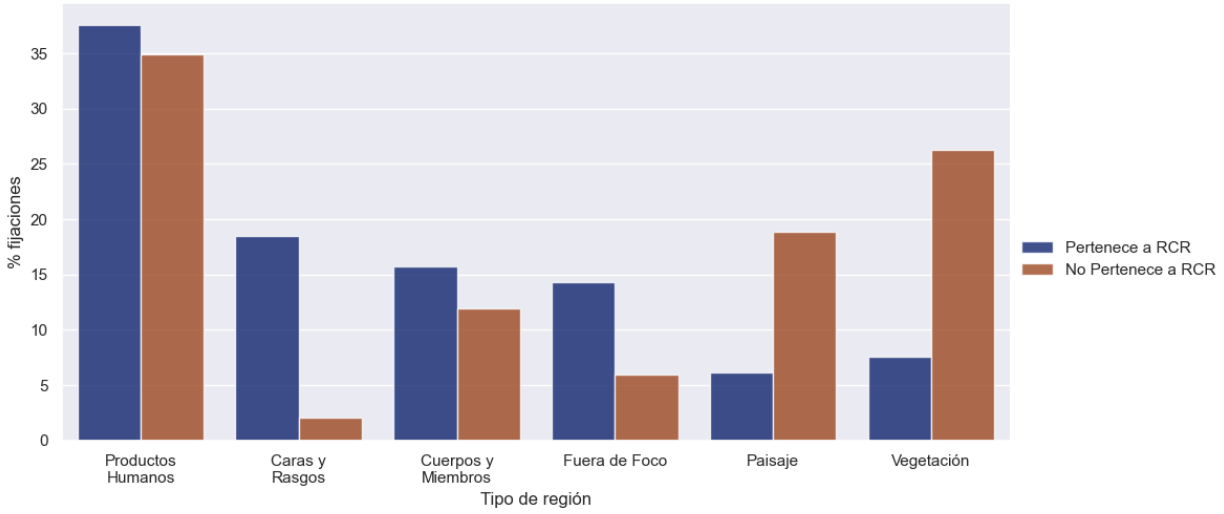



Figura 4.5: Contraste de la proporción de fijaciones realizadas sobre cada tipo de objeto encontrado por SAM, según su pertenencia a una Región Contextualmente Relevante. En azul, fijaciones pertenecientes a RCR's. En café, fijaciones que ocurren fuera de estas regiones.

Se observa que el porcentaje de fijaciones realizadas sobre productos humanos (prendas, construcciones, etc.) es muy similar entre ambos conjuntos: 37.59 % perteneciente a RCR vs 34.88 % fuera de éstas. En contraste, en regiones donde hay caras o rasgos faciales la diferencia en proporción es mucho mayor, involucrando un 18.50 % de las fijaciones pertenecientes a las RCR pero sólo un 2.03 % de las fijaciones externas a este conjunto. Frente a cuerpos y miembros, los porcentajes son similares: 15.72 % (RCR) versus 11.96 %. El último tipo de zona cuya mayor proporción de fijaciones pertenece a RCR's son aquellas fuera de foco o de fondo, alcanzando un 14.36 % contra el 5.98 % de fijaciones que no pertenecen a RCR's.

Las regiones que contienen vegetación o paisaje muestran una mayor correlación con fijaciones no pertenecientes a RCR's, donde el 26.29 % de éstas cae en regiones con vegetación y un 18.85 % coincide con zonas del paisaje en las escenas, lo que contrasta fuertemente con el porcentaje de fijaciones pertenecientes a RCR's que caen en las mismas zonas (7.61 % y 6.19 % respectivamente).

De los resultados mostrados se evidencia que regiones con mayor probabilidad de fijación (pertenecientes a RCR's) se relacionan en su mayoría a productos humanos, rasgos, caras, cuerpos o miembros, sumando un 71.81 % del total de fijaciones. En contraste, las fijaciones que no pertenecen a este conjunto se concentran sobre regiones con vegetación, paisaje o productos humanos (75.53 % del total). El principal cambio entre fijaciones que pertenecen o no a las RCR's está en que disminuye la coincidencia con caras y rasgos faciales y aumenta la coincidencia con vegetación y paisaje.

### 4.3.1. Discusión

La validación realizada permite confirmar lo expuesto en la sección 2.2.3.2 ya que muestra cómo ciertas zonas de la imagen, normalmente relacionadas a lo *top down* en la literatura (caras y rasgos faciales, cuerpos y miembros), coinciden con el modelo RCR (18.50 %, 15.72 %

respectivamente). En contraste, partes de la imagen fijadas que no coinciden con el modelo son en su mayoría vegetación o paisajes (26.29 % y 18.85 % respectivamente). De todas formas, las regiones que se llevan la mayor cantidad de fijaciones, sin importar si éstas pertenecen o no a una RCR, son los productos humanos. Este último resultado requiere un análisis en mayor profundidad, ya que los objetos pertenecientes a esta categoría van desde prendas de ropa a construcciones, barcos y otro tipo de objetos que podrían clasificarse como parte del paisaje y, si bien estos resultados permiten utilizar el modelo propuesto con cierta confianza, queda pendiente generar una representación del componente *top down* que no dependa de los datos empíricos producidos por los sujetos.

## 4.4. Evolución en la atención visual

A continuación se reportan los resultados obtenidos al realizar el procedimiento explicado en la sección 3.6. Sobre todas las imágenes (exceptuando imágenes planas) se aplican tres modelos de atención (RCR, BMS e *Itti&Koch*), estos modelos se comparan con las fijaciones, previamente separadas según su orden de ocurrencia, a través de las métricas EMD y SIM. En imágenes planas sólo se aplica RCR ya que éste se calcula a partir de las fijaciones registradas y los modelos *bottom up* no pueden calcular una probabilidad de fijación a partir de imágenes sin ningún tipo de contenido. Las figuras 4.6, 4.7 y 4.8 muestran cómo evoluciona la similitud entre cada modelo de atención y la ubicación de las fijaciones a lo largo del tiempo. En las tres figuras, cada columna representa una de las métricas utilizadas: EMD en la izquierda, SIM en la derecha (sección 2.4).

Se obtienen tres puntajes de similitud/disimilitud (uno por modelo) para cada uno de los conjuntos de fijaciones realizadas sobre la imagen (primera hasta doceava fijación). Se grafican estos puntajes en función del orden de ocurrencia de las fijaciones, cada línea muestra la evolución de similitud entre éstas y el modelo de atención correspondiente. Los marcadores circulares representan la mediana sobre los puntajes obtenidos a lo largo de las imágenes para cada subconjunto de fijaciones. Las bandas representan un intervalo de confianza del 95 % sobre la estimación de dicho estadístico. Cambios significativos en la similitud entre fijaciones y modelos se obtienen a través de test t pareados (sección 3.3.1) y se reportan aumentos o disminuciones ($<$ o $>$) de la siguiente manera: * ($p$<0.05); ** ($p$<0.01); *** ($p$<0.001).

### 4.4.1. Imágenes con contenido contextual

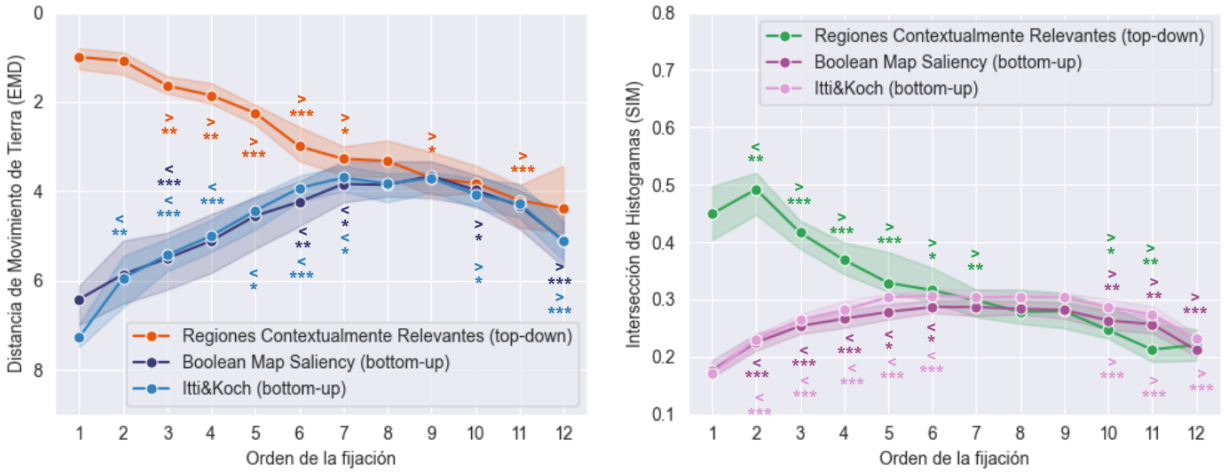



Figura 4.6: Similitud entre modelos de atención y distribución de fijaciones agrupadas según su orden de ocurrencia. Se grafican los datos obtenidos en experimentos donde se presentan imágenes con contenido contextual (naturales, ruido blanco e invertidas). Cada columna reporta una de las métricas utilizadas. Cuando la diferencia ($>$ o $<$) entre un conjunto de fijaciones y el siguiente es significativa, ésta se reporta de la siguiente manera: * ($p$<0.05); ** ($p$<0.01); *** ($p$<0.001).

Según lo reportado en la figura 4.6, a medida que aumenta el orden de ocurrencia existe una marcada disminución de similitud entre las fijaciones y las Regiones Contextualmente Relevantes. Este comportamiento se observa en todos los tipos de imagen analizados, al menos hasta la séptima fijación. En imágenes con contexto se reporta un descenso significativo entre la tercera y la séptima fijación para la métrica EMD, lo que se repite bajo la métrica SIM. Bajo esta segunda métrica, sin embargo, existe un aumento de similitud, que ocurre entre la primera y la segunda fijación. Por la formulación de las métricas, este último resultado es indicador de que las fijaciones se encuentran más concentradas inicialmente, lo que resulta en un menor traslape con las RCR's, si bien ocurren dentro de ellas. Esto afecta más el puntaje SIM que EMD, y a partir de la segunda ocurrencia, las ubicaciones se encuentran más dispersas, provocando que el comportamiento de ambas métricas empiece a coincidir.

Frente a este mismo tipo de imágenes, la similitud respecto al modelo de *Itti & Koch* aumenta de manera significativa y para ambas métricas desde la segunda a la séptima fijación. Si bien las medianas en los puntajes del modelo BMS siguen una tendencia parecida, la variabilidad es mayor en este caso y sólo se observan cambios significativos en la segunda y sexta fijaciones. Para los tres modelos existe una disminución de similitud cerca de las últimas fijaciones.

### 4.4.2. Imágenes ruido rosa

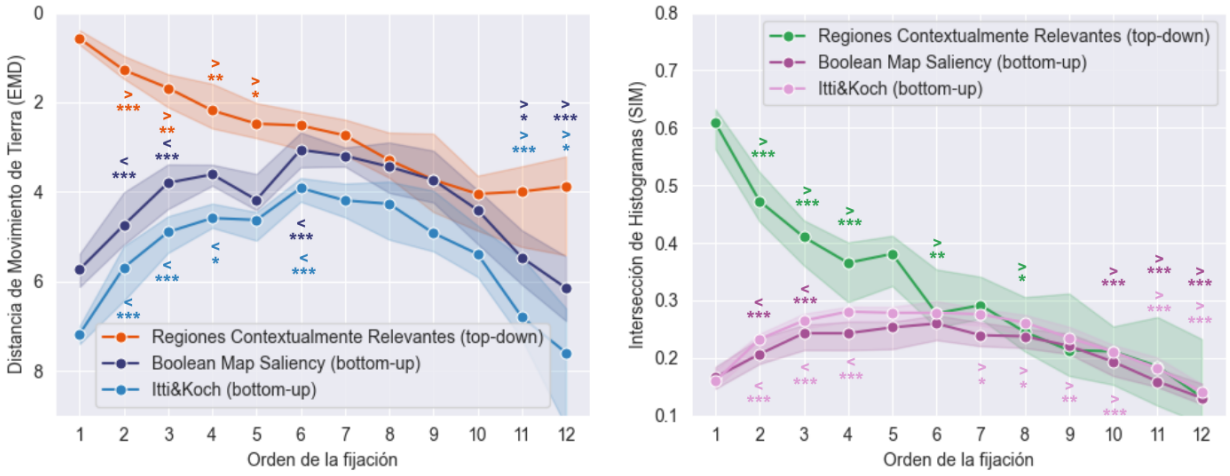



Figura 4.7: Similitud entre modelos de atención y distribución de fijaciones agrupadas según su orden de ocurrencia. Se grafican los datos obtenidos en experimentos donde se presentan imágenes ruido rosa. Cada columna reporta una de las métricas utilizadas. Cuando la diferencia ($>$ o $<$) entre un conjunto de fijaciones y el siguiente es significativa, ésta se reporta de la siguiente manera: * ($p$<0.05); ** ($p$<0.01); *** ($p$<0.001).

Cuando se muestran imágenes ruido rosa (figura 4.7), la similitud entre las fijaciones y las RCR's disminuye de manera significativa desde la primera hasta la quinta fijación, según EMD, y hasta la cuarta según SIM. Esta disminución es mucho más marcada de acuerdo a la métrica SIM y, bajo esta misma métrica, existe una interrupción en la tendencia durante la quinta fijación. Este comportamiento coincide con el observado para lo modelos *bottom up* según la métrica EMD. En estos modelos, el aumento de similitud parte desde la primera fijación y llega hasta la tercera (para BMS) o hasta la cuarta (para *Itti&Koch*). En ambos modelos la tendencia cambia durante la quinta fijación, disminuyendo su similitud para posteriormente aumentar una última vez hacia la sexta ocurrencia. Excluyendo el descenso exhibido durante la quinta fijación, el comportamiento de los modelos *bottom up* es muy similar para la métrica SIM. Una diferencia adicional es que el modelo *Itti&Koch* tiende a tener mejor puntajes que BMS bajo la métrica SIM, lo que ocurre a la inversa según EMD.

### 4.4.3. Imágenes planas

Sobre imágenes planas se aplica exclusivamente el modelo de Regiones Contextualmente Relevantes. La similitud a las RCR sólo muestra una tendencia bajo la métrica SIM, la que exhibe cambios significativos entre la primera y cuarta fijación, luego el descenso se mantiene en menor medida. Bajo la métrica EMD no se exhiben tendencias claras y se observa una gran dispersión en los datos desde la sexta fijación.


Figura 4.8: Similitud entre modelos de atención y distribución de fijaciones agrupadas según su orden de ocurrencia. Se grafican los datos obtenidos en experimentos donde se presentan imágenes planas (blancas, negras y grises). Cada columna reporta una de las métricas utilizadas. Cuando la diferencia ($>$ o $<$) entre un conjunto de fijaciones y el siguiente es significativa, ésta se reporta de la siguiente manera: * ($p$<0.05); ** ($p$<0.01); *** ($p$<0.001).

### 4.4.4. Discusión

El resultado más importante de esta sección muestra que en imágenes ruido rosa e imágenes con contenido contextual las fijaciones realizadas inicialmente distan mucho de los modelos de atención *bottom up* y que, a medida que avanza el experimento, van asemejándose progresivamente a dichos modelos. Esta tendencia se ve interrumpida en cierto momento, alcanzando un máximo de similitud para luego comenzar a decaer hacia el final del registro (la disminución se observa para los tres modelos de atención en todos los casos, pero la existencia de un máximo sólo se observa para modelos *bottom up*). La principal diferencia en ambas situaciones es que esta interrupción ocurre durante la séptima fijación cuando se muestran imágenes con contenido contextual. En contraste, frente a imágenes ruido rosa y dependiendo de la métrica y el modelo analizados, la tendencia exhibida se interrumpe entre una y dos fijaciones antes que lo observado en el primer caso. Debido a que las fijaciones duran más cuando se presentan imágenes ruido rosa, la temporalidad de esta interrupción coincide para estos dos tipos de imagen, lo que será analizado en la siguiente sección.

## 4.5. Tiempos de inicio de las fijaciones

En la figura 4.6 se hace notorio que las fijaciones disminuyen su similitud al modelo *top down* mientras se parecen progresivamente más al modelo *bottom up*. Este comportamiento aparece bajo las dos métricas utilizadas y para los dos tipos de modelos utilizados ocurre hasta la séptima fijación. Cuando se presentan imágenes ruido rosa (figura 4.7), la tendencia no se exhibe de manera consistente entre las métricas y modelos. Sin embargo, se observa que la similitud máxima con los modelos *bottom up* se alcanza en la sexta fijación.

De acuerdo a los resultados nombrados y a lo descrito en la sección 3.8, es de interés

identificar el instante de tiempo en que ocurren un cambio en las tendencias exhibidas por los datos. En este sentido, se estudia la relación entre los ordenes de fijación y el tiempo que éstas ocurren. En la figura 4.9 se muestra el promedio en el tiempo de inicio de las fijaciones, separadas según su orden de ocurrencia y por el tipo de imagen mostrada durante los experimentos. Se observa que, aproximadamente entre el orden 3 y el orden 8 existe una diferencia relativamente constante en el tiempo en que se inician las fijaciones, dependiente del tipo de imagen que se presenta: en este lapso, el tiempo en que ocurre una fijación frente a imágenes planas o ruido rosa tiende a ser el mismo que la fijación siguiente cuando se presentan imágenes con contenido contextual. Por ejemplo, la cuarta fijación ocurre en promedio a los 1.5 segundos cuando se muestran imágenes planas o ruido rosa, y este instante de tiempo coincide (en promedio) con el inicio de la quinta fijación cuando se presentan imágenes con contenido contextual.

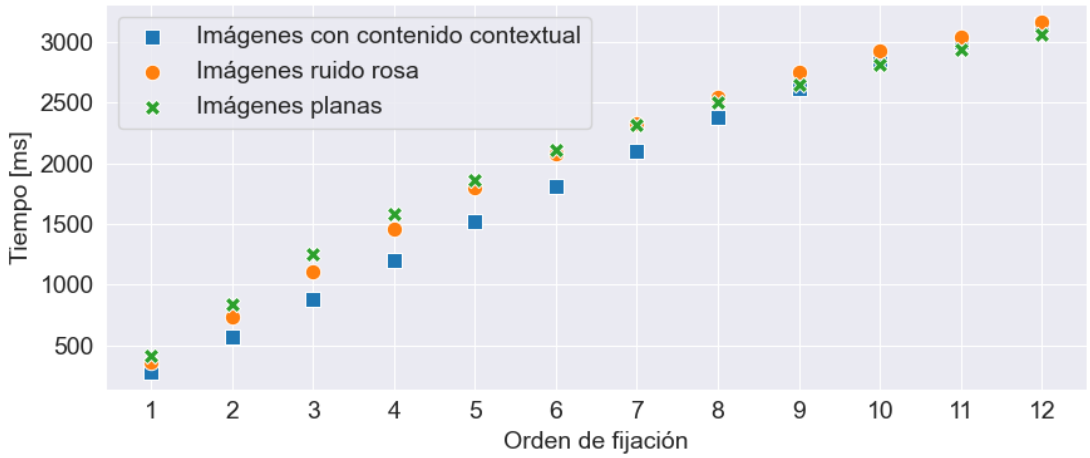



Figura 4.9: Tiempo promedio en que ocurren las fijaciones según su orden. Los marcadores (cuadrados para contenido contextual, círculos para ruido rosa y"X" para imágenes planas) indican la media a lo largo de las fijaciones.

Una inspección más detallada se presenta en la figura 4.10, donde se muestra la distribución de tiempos para la séptima fijación frente a imágenes con contenido contextual (arriba) y lo mismo para las fijaciones 5 (abajo) y 6 (al medio) realizadas sobre imágenes ruido rosa. Se puede observar que cuando se presentan imágenes con contenido contextual, la distribución de tiempos para la séptima fijación es muy similar a la distribución de la sexta fijación frente a ruido rosa. Se realiza un test-t pareado sobre las medianas de tiempo obtenidas por sujeto, comparando la distribución de la séptima fijación en imágenes con contenido contextual y la sexta fijación frente a ruido rosa. El test indica que no existe una diferencia significativa entre ambas distribuciones ($t(15) = -0.110$, $p = 0.9134$).

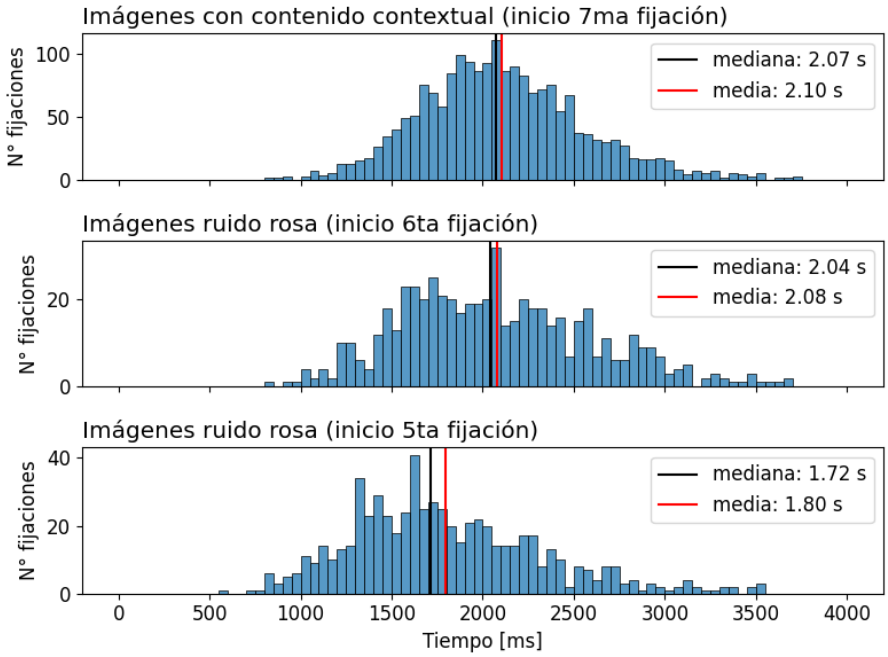



Figura 4.10: Distribución de tiempo para la séptima fijación durante experimentos con imágenes que muestran contenido contextual (arriba) y para la sexta y quinta fijaciones realizadas sobre imágenes ruido rosa (al medio y abajo respectivamente). Histogramas con bines de 50 ms, se reporta la mediana (línea vertical negra) y la media (línea vertical roja) de las distribuciones.

### 4.5.1. Discusión

La existencia de un máximo de similitud, seguido por una disminución monotónica hacia el final de los experimentos, respecto de los tres modelos utilizados (figuras 4.6 y 4.7), puede indicar una pérdida de interés sobre el estímulo, resultando en una disminución progresiva de la atención general de parte de los sujetos. Debido a que las fijaciones realizadas sobre imágenes ruido rosa tienden a durar más que frente a contenido contextual (figura 4.9), el tiempo promedio en que ocurre el máximo es muy parecido para ambas (figura 4.10). Esto podría indicar un marcador temporal claro en que los sujetos cambian de prioridades atencionales o simplemente se comienzan a desacoplar de la escena presentada, el cual estaría cercano a los 2.1 segundos desde la presentación del estímulo.

# Capítulo 5

# Conclusiones

En este trabajo se estudió cómo cambia la atención visual, separada en sus componentes *bottom up* y *top down*, a lo largo de experimentos de visión libre. Se utilizaron datos de posición ocular registrados en sujetos instruidos a explorar escenas visuales de manera libre por períodos cercanos a los 4 segundos. Para estudiar estos cambios en la atención, se definieron modelos que representan cada uno de sus componentes y se compararon éstos con las fijaciones registradas, separándolas según su orden de ocurrencia temporal.

Para el componente *bottom up* se aplicaron dos modelos de atención, basados en la Teoría de Integración de Características (*Itti & Koch*) y La Teoría Gestalt de segmentación figura-fondo (*Boolean Map Saliency*). Modelar el componente *top down* representó una dificultad mayor al no existir una gran variedad de propuestas, las que presentaban limitaciones considerables en su aplicación e interpretación. Se optó por definir regiones de la imagen cuya relevancia se manifiesta a través de una mayor cantidad de fijaciones respecto del resto de la escena (Regiones Contextualmente Relevantes). Se realizó una validación del modelo a través de un estudio sobre los distintos objetos y regiones que contiene cada imagen. Los resultados obtenidos se alinean satisfactoriamente con la literatura, permitiendo apoyarse con más confianza en la propuesta.

Siguiendo los objetivos 1, 2 y 3, para cada imagen analizada se calculan tres modelos: los dos modelos de atención *bottom up* (*Itti & Koch* y *Boolean Map Saliency*) y el modelo de atención *top down* (Regiones Contextualmente Relevantes). Las fijaciones se agruparon temporalmente y se compararon con dichos modelos a través de métricas comúnmente utilizadas en la literatura: Intersección de Histogramas y Distancia de Movimiento de Tierra. De acuerdo al objetivo 4, la evolución de la atención se estudió para imágenes con distintos niveles de contenido: imágenes planas (sin contenido), imágenes ruido rosa (contenido sólo físico, sin contexto) e imágenes con contenido contextual.

El análisis realizado permitió confirmar la presencia de las tendencias propuestas en la hipótesis, específicamente cuando se presentaron imágenes ruido rosa y con contenido (no se presentan tendencias claras frente a imágenes planas). Esto podría indicar un cambio de prioridad atencional a lo largo del tiempo, pasando de una dominancia inicial del componente *top down* a un equilibrio entre ambos componentes, cerca de los 2 segundos. Este equilibrio ocurre luego de una disminución constante en la similitud entre el componente *top down* y las fijaciones, las que se van asemejando progresivamente más al componente *bottom up*. Luego

de este punto, ambos modelos disminuyen su similitud a las fijaciones, monotónicamente y hasta el final del tiempo de experimento.

La disminución en la similitud entre fijaciones y modelos de atención hacia el final de los experimentos puede indicar una pérdida de interés sobre el estímulo, resultado en una disminución progresiva de la atención general de parte del sujeto. Este resultado podría tomarse como un marcador temporal claro en que los sujetos cambian de prioridades atencionales o simplemente se comienzan a desacoplar de la escena presentada. El determinar este umbral de cambio responde exitosamente al objetivo 5 de la tesis.

Se encontró que el tipo de contenido presente en las imágenes afecta significativamente los tiempos de reacción de los sujetos: frente a imágenes con contenido contextual el tiempo entre la presentación del estímulo y la primera sacada es significativamente menor que cuando no existen objetos o regiones discernibles (imágenes planas o ruido rosa), lo que puede indicar que la presencia de contexto entrega suficiente información para definir un patrón de exploración de manera más rápida. Los resultados en esta línea también muestran que imágenes con caras invertidas se relacionan a tiempos de reacción mayores que imágenes con caras en su posición natural. Esto es consistente con la evidencia de que seres humanos poseen un patrón específico para la exploración de caras [62], el cual posiblemente se vea obstaculizado por el cambio de orientación.

En la literatura existe una gran cantidad de estudios sobre las dinámicas del comportamiento ocular durante experimentos de visión libre. Esta evolución se ha caracterizado, en la gran mayoría de los casos, a través de la duración de las fijaciones, la amplitud de las sacadas y el cambio de estas variables a lo largo del tiempo. Si bien los datos analizados en este trabajo no exhiben las tendencias sobre la duración de las fijaciones o amplitud de sacadas observadas en la literatura, el estudio a través de los modelos de atención sí permite identificar tendencias sistemáticas en las dinámicas de exploración visual, a través de nuevas variables basadas en la interacción entre el sujeto y la información disponible en las escenas mostradas.

En base a lo anterior, la principal contribución de esta tesis radica en la propuesta de una caracterización novedosa para estudiar la evolución de la atención visual, no utilizada anteriormente en el estudio de las dinámicas del comportamiento ocular. La temporalidad con la que cambia la similitud entre los modelos de atención *bottom up* y las fijaciones se puede utilizar como referencia para realizar comparaciones entre las dos etapas observadas (antes y después de los 2 segundos) para variables adicionales del tipo conductual o cerebral. Los resultados obtenidos representan un nuevo acercamiento a la incorporación de la variable de tiempo para la predicción de fijaciones, donde la alternancia de los componentes de la atención visual podría dar mejores explicaciones sobre las preferencias visuales en comparación a modelos estáticos de prominencia visual.

Modelos computacionales que logren capturar exitosamente ambos componentes atencionales de las imágenes tienen el potencial de predecir con mayor precisión las regiones de interés de los sujetos, permitiendo el desarrollo de interfaces humano máquina más efectivas, así como la implementación de experimentos en visión que se beneficien de una manipulación de estos componentes para estudiar la respuesta de los sujetos frente a imágenes alteradas o cambiantes. Por otro lado, cualquier caracterización nueva sobre la conducta ocular se puede

utilizar para estudiar patologías neurodegenerativas o del neurodesarrollo, especialmente la agnosia visual (alteración en la capacidad de reconocer objetos visuales) y esquizofrenia [63].

## 5.1. Trabajo Futuro

De acuerdo a lo expuesto anteriormente, una primera extensión de este trabajo podría centrarse en el estudio de la dilatación pupilar, analizando su evolución en el tiempo o directamente comparando el antes y después del punto de quiebre encontrado. Debido a que la pupilometría se utiliza como indicador del nivel de atención o acople entre el sujeto y el estímulo [64], esta variable se puede relacionar de manera directa al análisis de los modelos utilizados. Por ejemplo, dado que las fijaciones comienzan a distanciarse de todos los modelos de atención después de los 2 segundos, la pupilometría podría ayudar a dilucidar si el cambio observado responde efectivamente a una disminución general en la atención sobre la escena mostrada o si son los modelos de atención los que no logran explicar adecuadamente las prioridades atencionales del sujeto durante los últimos segundos del experimento.

Además de variables conductuales, mediciones como electroencefalogramas (EEG) o imágenes por resonancia magnética funcional (fMRI) permiten identificar cambios en la actividad cerebral, asociados a distintos estados cognitivos o regiones del cerebro [65–67]. Estos cambios podrían relacionarse, por un lado, a los componentes de atención visual y los períodos en que cada uno está siendo atendido con mayor intensidad. Por otro lado, las mediciones se pueden analizar temporalmente, buscando tendencias o cambios que coincidan con aquellos identificados en este estudio.

Una última extensión se deberá enfocar en la definición de modelos de atención *top down* confiables, que permitan identificar regiones u objetos contextualmente relevantes exclusivamente a partir de los contenidos de la imagen analizada. En esta línea, un acercamiento con resultados cada vez más prometedores se apoya en el uso de redes neuronales y *deep learning* para cuantificar e identificar objetos o regiones de interés [43, 68–70]. Utilizar estos modelos requerirá un tratamiento cuidadoso de las predicciones obtenidas ya que, si bien los modelos incorporan características de más alto nivel, diferenciar qué parte corresponde a la atención *top down* y cuál sigue siendo *bottom up* no es una tarea trivial.

# Bibliografía


[1] Yarbus, A., Eye movements and Vision. New York: Plenum Press, 1967.

[2] Kamensek, T., Wong, E., Leung, C., Iarocci, G., y Oruc, I., "The face diet of adults with autism spectrum disorder," Journal of Vision, vol. 22, no. 14, p. 4483, 2022.

[3] Nagarajan, K., Luo, G., Narasimhan, M., y Satgunam, P., "Children with amblyopia make more saccadic fixations when doing the visual search task," Investigative Opthalmology &amp Visual Science, vol. 63, no. 13, p. 27, 2022.

[4] Gide, M. y Karam, L., "Computational visual attention models," Foundations and Trends® in Signal Processing, vol. 10, no. 4, pp. 347–427, 2017.

[5] Tatler, B. y Vincent, B. T., "Systematic tendencies in scene viewing," Journal of Eye Movement Research, vol. 2, no. 2, pp. 1–18, 2008.

[6] Antes, J., "The time course of picture viewing," Journal of Experimental Psychology, vol. 103, no. 1, pp. 62–70, 1974.

[7] Unema, P., Pannash, S., Joos, M., y Velichkovsky, B., "The time course of information processing during scene perception: The relationship between saccade amplitude and fixation duration," Visual Cognition, vol. 12, no. 3, pp. 473–494, 2005.

[8] Henderson, J., "Human gaze control during real-world scene perception," TRENDS in Cognitive Sciences, vol. 7, no. 11, pp. 498–504, 2003.

[9] Tatler, B. y Brockmole, J., "Latest: A model of saccadic decisions in space and time," TRENDS in Cognitive Sciences, vol. 7, no. 11, pp. 498–504, 2003.

[10] Trevarthen, C., "Two mechanisms of vision in primates," Psychologische Forschung, vol. 31, pp. 299–337, 2017.

[11] Velichkovsky, B., Joos, M., Helmert, J., y Pannasch, S., "Two visual systems and their eye movements: Evidence from static and dynamic scene perception," Applied Cognitive Research/Psychology III, pp. 2283–2288, 2005.

[12] Pannasch, S. y Velichkovsky, B., "Distractor effect and saccade amplitudes: Further evidence on different modes of processing in free exploration of visual images," Visual Cognition, vol. 17, no. 6, pp. 1109–1131, 2009.

[13] Ito, J., Yamane, Y., Suzuki, M., Maldonado, P., Fujita, I., Tamura, H., y Grün, S., "Switch from ambient to focal processing mode explains the dynamics of free viewing eye movements," Scientific Reports, 2017.

[14] Velichkovsky, B., Korosteleva, A., Panasch, S., Helmert, J., Orlov, V., Sharaev, M., Velichkovsky, B., y Ushakov, V., "Two visual systems and ther eye movements: a fixation-based event-related experiment with ultrafast fmri reconciles competing views," Advan-

ced Researches, vol. 11, no. 4, 2019.

[15] Milisavljevic, A., Le Bras, T., Mancas, M., Petermann, C., Gosselin, B., y Doré-Mazars, K., "Towards a better description of visual exploration through temporal dynamica of ambient and focal modes," 2019 Symposium on Eye Tracking Research and Applications, 2019.

[16] Malem-Shinitski, N., Opper, M., Reich, S., Schwetlick, L., Seelig, S., y Engbert, R., "A mathematical model of local and global attention in natural scene viewing," PLoS Computational Biology, vol. 16, no. 12, 2020.

[17] Le Meur, O. y Fons, P., "Predicting image influence on visual saliency distribution: the focal and ambient dichotomy," 2020 Symposium on Eye Tracking Research and Applications, 2020.

[18] Wurtz, R. H., "Neuronal mechanisms of visual stability," Vision Research, vol. 48, no. 20, pp. 2070–2089, 2008.

[19] Desimone, R. y Duncan, J., "Neural mechanisms of selective visual attention," Annual Review of Neuroscience, vol. 18, no. 1, pp. 193–222, 1995.

[20] DeAngelus, M. y Pelz, J. B., "Top-down control of eye movements: Yarbus revisited," Visual Cognition, vol. 17, no. 6-7, pp. 790–811, 2009.

[21] Howard, P. L., Zhang, L., y Benson, V., "What can eye movements tell us about subtle cognitive processing differences in autism?," Vision, vol. 3, no. 2, p. 22, 2019.

[22] Wolf, A., Ueda, K., y Hirano, Y., "Recent updates of eye movement abnormalities in patients with schizophrenia: A scoping review," Psychiatry and Clinical Neurosciences, vol. 75, no. 3, pp. 82–100, 2021.

[23] Shishido, E., Ogawa, S., Miyata, S., Yamamoto, M., Inada, T., y Ozaki, N., "Application of eye trackers for understanding mental disorders: Cases for schizophrenia and autism spectrum disorder," Neuropsychopharmacology Reports, vol. 39, no. 2, pp. 72–77, 2019.

[24] Readman, M. R., Polden, M., Gibbs, M. C., Wareing, L., y Crawford, T. J., "The potential of naturalistic eye movement tasks in the diagnosis of alzheimer's disease: A review," Brain Sciences, vol. 11, no. 11, p. 1503, 2021.

[25] Bueno, A., Sato, J., y Hornberger, M., "Eye tracking – the overlooked method to measure cognition in neurodegeneration?," Neuropsychologia, vol. 133, p. 107191, 2019.

[26] Itti, L., Koch, C., Niebur, E., y et al, "A model of saliency-based visual attention for rapid scene analysis," IEEE transactions on pattern analysis and machine intelligence, vol. 20, no. 11, pp. 1254–1259, 1998.

[27] Itti, L., Dhavale, N., y Pighin, F., "Realistic avatar eye and head animation using a neurobiological model of visual attention," en Applications and Science of Neural Networks, Fuzzy Systems, and Evolutionary Computation VI (Bosacchi, B., Fogel, D. B., y Bezdek, J. C., eds.), SPIE, 2004.

[28] Meur, O. L., Callet, P. L., Barba, D., y Thoreau, D., "A coherent computational approach to model bottom-up visual attention," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 5, pp. 802–817, 2006.

[29] Kootstra, G., Nederveen, A., y de Boer, B., "Paying attention to symmetry," en Procedings of the British Machine Vision Conference 2008, British Machine Vision Association,

2008.

[30] Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., y Durand, F., "What do different evaluation metrics tell us about saliency models?," arXiv preprint arXiv:1604.03605, 2016.

[31] Judd, T., Durand, F., y Torralba, A., "A benchmark of computational models of saliency to predict human fixations," en MIT Technical Report, 2012.

[32] Mazza, V., Turatto, M., y Umilt, C., "Foreground?background segmentation and attention: A change blindness study," Psychological Research Psychologische Forschung, vol. 69, no. 3, pp. 201–210, 2004.

[33] Bruce, N. D. B. y Tsotsos, J. K., "Saliency, attention, and visual search: An information theoretic approach," Journal of Vision, vol. 9, no. 3, pp. 5–5, 2009.

[34] Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., y Li, S., "Salient object detection: A discriminative regional feature integration approach," en Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

[35] Liu, Z., Meur, L., y Luo, S., "Superpixel-based saliency detection," en 2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), IEEE, 2013.

[36] Wen, S., Han, J., Zhang, D., y Guo, L., "Saliency detection based on feature learning using deep boltzmann machines," en 2014 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2014.

[37] Vig, E., Dorr, M., y Cox, D., "Large-scale optimization of hierarchical features for saliency prediction in natural images," en Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[38] Shen, C. y Zhao, Q., "Learning to predict eye fixations for semantic contents using multi-layer sparse network," Neurocomputing, vol. 138, pp. 61–68, 2014.

[39] Flechsenhar, A. F. y Gamer, M., "Top-down influence on gaze patterns in the presence of social features," PLOS ONE, vol. 12, no. 8, p. e0183799, 2017.

[40] Cerf, M., Harel, J., Einhäuser, W., y Koch, C., "Predicting human gaze using low-level saliency combined with face detection," Advances in neural information processing systems, vol. 20, 2007.

[41] Hwang, A. D., Wang, H.-C., y Pomplun, M., "Semantic guidance of eye movements in real-world scenes," Vision Research, vol. 51, no. 10, pp. 1192–1205, 2011.

[42] Foulsham, T., Barton, J. J., Kingstone, A., Dewhurst, R., y Underwood, G., "Fixation and saliency during search of natural scenes: The case of visual agnosia," Neuropsychologia, vol. 47, no. 8-9, pp. 1994–2003, 2009.

[43] Mahdi, A., Qin, J., y Crosby, G., "DeepFeat: A bottom-up and top-down saliency model based on deep features of convolutional neural networks," IEEE Transactions on Cognitive and Developmental Systems, vol. 12, no. 1, pp. 54–63, 2020.

[44] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., y Fei-Fei, L., "Imagenet: A large-scale hierarchical image database," en 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255, 2009.

[45] Zhang, J. y Sclaroff, S., "Saliency detection: A boolean map approach," en Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013.

[46] Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., y Durand, F., "What do different evaluation metrics tell us about saliency models?," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 3, pp. 740–757, 2019.

[47] Fawcett, T., "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.

[48] Tatler, B. W., "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," Journal of Vision, vol. 7, no. 14, p. 4, 2007.

[49] Peters, R. J., Iyer, A., Itti, L., y Koch, C., "Components of bottom-up gaze allocation in natural images," Vision Research, vol. 45, no. 18, pp. 2397–2416, 2005.

[50] Kümmerer, M., Wallis, T. S. A., y Bethge, M., "Information-theoretic model comparison unifies saliency metrics," Proceedings of the National Academy of Sciences, vol. 112, no. 52, pp. 16054–16059, 2015.

[51] Meur, O. L., Callet, P. L., y Barba, D., "Predicting visual fixations on video based on low-level visual features," Vision Research, vol. 47, no. 19, pp. 2483–2498, 2007.

[52] Joyce, J. M., "Kullback-leibler divergence," en International Encyclopedia of Statistical Science, pp. 720–722, Springer Berlin Heidelberg, 2011.

[53] Swain, M. J. y Ballard, D. H., "Color indexing," International Journal of Computer Vision, vol. 7, pp. 11–32, 1991.

[54] Rubner, Y. International Journal of Computer Vision, vol. 40, no. 2, pp. 99–121, 2000.

[55] Rubner, Y., Tomasi, C., y Guibas, L. J., "The earth mover's distance as a metric for image retrieval," International journal of computer vision, vol. 40, no. 2, p. 99, 2000.

[56] Devia, C., Montefusco-Siegmund, R., Egaña, J. I., y Maldonado, P. E., "Precise timing of sensory modulations coupled to eye movements during active vision," 2017.

[57] Lang, P. J., Bradley, M. M., Cuthbert, B. N., *et al.*, International affective picture system (IAPS): Affective ratings of pictures and instruction manual. NIMH, Center for the Study of Emotion & Attention Gainesville, FL, 2005.

[58] Madariaga, S., Babul, C., Egaña, J. I., Rubio-Venegas, I., Güney, G., Concha-Miranda, M., Maldonado, P. E., y Devia, C., "SaFiDe: Detection of saccade and fixation periods based on eye-movement attributes from video-oculography, scleral coil or electrooculography data," MethodsX, vol. 10, p. 102041, 2023.

[59] Meur, O. L. y Baccino, T., "Methods for comparing scanpaths and saliency maps: strengths and weaknesses," Behavior Research Methods, vol. 45, no. 1, pp. 251–266, 2012.

[60] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., y Girshick, R., "Segment anything," 2023.

[61] Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., y Vayer, T., "Pot: Python optimal transport," Journal of Machine Learning Research, vol. 22, no. 78, pp. 1–8, 2021.

[62] Rogers, S. L., Speelman, C. P., Guidetti, O., y Longmuir, M., "Using dual eye tracking

to uncover personal gaze patterns during social interaction," Scientific Reports, vol. 8, no. 1, 2018.

[63] Gold, J. M., Fuller, R. L., Robinson, B. M., Braun, E. L., y Luck, S. J., "Impaired top–down control of visual search in schizophrenia," Schizophrenia Research, vol. 94, no. 1-3, pp. 148–155, 2007, doi:10.1016/j.schres.2007.04.023.

[64] Mathôt, S., "Pupillometry: Psychology, physiology, and function," Journal of Cognition, vol. 1, no. 1, 2018.

[65] Liu, Y., Bengson, J., Huang, H., Mangun, G. R., y Ding, M., "Top-down modulation of neural activity in anticipatory visual attention: Control mechanisms revealed by simultaneous EEG-fMRI," Cerebral Cortex, p. bhu204, 2014.

[66] Ahirwal, M. K. y Londhe, N., "Power spectrum analysis of eeg signals for estimating visual attention," International Journal of computer applications, vol. 42, no. 15, pp. 22–25, 2012.

[67] Busch, N. A. y VanRullen, R., "Spontaneous EEG oscillations reveal periodic sampling of visual attention," Proceedings of the National Academy of Sciences, vol. 107, no. 37, pp. 16048–16053, 2010.

[68] Yan, F., Chen, C., Xiao, P., Qi, S., Wang, Z., y Xiao, R., "Review of visual saliency prediction: Development process from neurobiological basis to deep models," Applied Sciences, vol. 12, no. 1, p. 309, 2021.

[69] Tliba, M., Kerkouri, M. A., Ghariba, B., Chetouani, A., Coltekin, A., Shehata, M. S., y Bruno, A., "SATSal: A multi-level self-attention based architecture for visual saliency prediction," IEEE Access, vol. 10, pp. 20701–20713, 2022.

[70] Paramanandam, K. y Kanagavalli, R., "A review on deep learning techniques for saliency detection," en Information and Communication Technology for Competitive Strategies (ICTCS 2021), pp. 279–289, Springer Nature Singapore, 2022.