Universidad Zaragoza

Trabajo Fin de Grado

Implementación y análisis de modelos de predicción de movimientos oculares (scanpaths) en 2D

Implementation and analysis of 2D scanpath prediction models

Autor

Miguel Gómez Lahera

Directora

Sandra Malpica Mallo

Ponente

Diego Gutiérrez Pérez

Grado en Ingeniería Informática

ESCUELA DE INGENIERÍA Y ARQUITECTURA

2021

Agradecimientos

A Sandra, la directora de este trabajo, a la que expreso mi más sincera gratitud por su inestimable dirección, su ayuda y por todo lo que he aprendido de ella durante la realización del mismo.

A todos mis compañeros y amigos, con los que durante estos años he compartido tantos momentos y que de una forma u otra me han influido.

A Laura, quien me acompaña desde hace ya muchos años y es parte de mí. Y a sus padres por acogerme como uno más en su familia.

Y por último, pero no menos importante, a mis padres y mis hermanos, que son los que han estado siempre conmigo, en los mejores y peores momentos, y saben que no ha sido fácil llegar hasta aquí.

Por haber estado siempre conmigo, gracias.

Resumen

La atención visual engloba un conjunto de operaciones cognitivas que nos permiten obtener una representación mental del mundo que nos rodea. A su vez, los movimientos oculares son una manifestación cuantificable de los procesos de decisión que rigen dicha atención visual. En los últimos años han surgido diversos modelos para intentar predecir, modelar y analizar los movimientos oculares en una escena. Dos conceptos clave para intentar entender la atención visual son los mapas de saliencia y los scanpaths. Un mapa de saliencia resalta la región o regiones de una escena en la que es más probable que un observador se fije. Un scanpath es el camino concreto que los ojos de un individuo han seguido al observar una escena. Ambos conceptos están directamente relacionados, y es posible calcular un mapa de saliencia a partir de scanpaths o viceversa.

En particular, la cuestión de qué características de la imagen son las que afectan a la distribución espacial de las fijaciones ha sido el foco principal de un gran número de investigaciones hasta ahora. Los mapas de saliencia están directamente relacionados con los scanpaths, ya que se pueden considerar su antesala. Sin embargo, los mapas de saliencia dan una visión estática o final de la saliencia de una imagen, mientras que los *scanpaths* permiten estudiar de forma dinámica el recorrido visual de un observador particular para una escena. En la actualidad existen multitud de modelos de generación de mapas de saliencia, que se pueden calcular agregando el comportamiento estático de multitud de usuarios. En cambio, no existen tantos modelos de predicción de scanpaths, debido a su mayor complejidad. Sin embargo, en este trabajo nos centramos en el estudio de modelos de predicción de *scanpaths* ya que dan una información más completa y son capaces de simular el comportamiento variable de cada individuo al visualizar una escena.

Partiendo del estado del arte, se escogen los modelos con un mejor rendimiento en imágenes naturales, se implementan y se analizan en un banco de pruebas diseñado para estudiar la capacidad de generalización de los modelos utilizando un recopilatorio de bases de datos de un estudio hospitalario. Las métricas utilizadas nos permiten analizar de forma cuantitativa el comportamiento de los modelos, teniendo en cuenta las características espaciotemporales de los scanpaths generados, comprobando que no existe un único modelo que sea mejor en todas las categorías de imágenes analizadas.

Debido a la falta de datos que no se basen en tareas visuales de exploración libre, no es posible entrenar un nuevo modelo que supere el rendimiento de los modelos actuales en estas nuevas categorías. Debido a ello, se propone un metamodelo que se compone de un categorizador de imágenes basado en *Deep Learning* que nos permite escoger siempre el modelo con mejor rendimiento en cada categoría.

Abstract

Visual attention encompasses a set of cognitive operations that allow us to obtain a mental representation of the world around us. In turn, eye movements are a quantifiable manifestation of the decision-making processes that govern visual attention. In recent years, a number of models have emerged to attempt to predict, model and analyse eye movements in a scene. Two key concepts in trying to understand visual attention are saliency maps and scanpaths. A saliency map highlights the region or regions of a scene that an observer is most likely to notice. A scanpath is the specific path that an individual's eyes have followed when observing a scene. The two concepts are directly related, and it is possible to calculate a saliency map from scanpaths or vice versa.

In particular, the question of which image features affect the spatial distribution of fixations has been the main focus of a great deal of research so far. Saliency maps are directly related to scanpaths, as they can be considered their anteroom. However, saliency maps give a static or final view of the saliency of an image, whereas scanpaths allow to study dynamically the visual path of a particular observer for a scene. There are now many models for generating saliency maps, which can be calculated by aggregating the static behaviour of a multitude of users. On the other hand, there are not so many models for predicting scanpaths, due to their greater complexity. However, in this paper we focus on the study of scanpath prediction models as they provide more complete information and are able to simulate the variable behaviour of each individual when viewing a scene.

Based on the state of the art, the models with the best performance in natural images are chosen, implemented and analysed in a test bench designed to study the generalisation capacity of the models using a compilation of databases from a hospital study. The metrics used allow us to quantitatively analyse the behaviour of the models, taking into account the spatio-temporal characteristics of the scanpaths generated, proving that there is no single model that is better in all the categories of images analysed.

Due to the lack of data that are not based on free exploration visual tasks, it is not possible to train a new model that outperforms the current models in these new categories. Because of this, a metamodel is proposed that is composed of an image categoriser based on Deep Learning that allows us to always choose the best performing model in each category.

Índice

Índice de figuras

1. Introducción

1.1. Contexto del proyecto

La atención visual engloba un conjunto de operaciones cognitivas que nos permiten obtener una representación mental del mundo que nos rodea. A su vez, los movimientos oculares son una manifestación cuantificable de los procesos de decisión que rigen dicha atención visual.

En cuanto a los movimientos oculares hay dos conceptos de gran importancia: las fijaciones y las sacadas. La fijación visual es la habilidad monocular que tiene el ojo para mantener la mirada enfocada en un objeto. La fijación visual forma parte de los movimientos oculares, ya que nos ayuda a crear una imagen nítida y estable de la escena que vemos en cada momento. Las sacadas en cambio son movimientos rápidos, de trayectoria balística, que nos permiten dirigir nuestra atención a distintos objetos del entorno. A una ruta de movimientos oculares o sucesión de fijaciones y sacadas le llamamos *scanpath* [1].

Debido a las limitaciones anatómicas de los ojos, los seres humanos recopilamos información de alta densidad solo para una pequeña área de nuestro campo de visión. Esta limitación nos obliga a dirigir nuestra atención hacia lo que sea más relevante en cada momento. La forma en la que elegimos dónde mirar ha atraído muchas investigaciones a lo largo de las décadas. Una de las formas en que se trata de analizar el comportamiento visual es mediante lo que se conoce como mapas de saliencia (*saliency maps*). En particular, la cuestión de qué características de la imagen son las que afectan a la distribución espacial de las fijaciones ha sido el foco principal de un gran número de investigaciones hasta ahora. Los mapas de saliencia están directamente relacionados con los *scanpaths*, ya que se pueden considerar su antesala. Sin embargo, los mapas de saliencia dan una visión estática o final de la saliencia de una imagen, mientras que los *scanpaths* permiten estudiar de forma dinámica el recorrido visual de un observador particular para una escena. Ya que dan una información más completa, aunque de mayor complejidad, en este trabajo nos centramos en el estudio de modelos de predicción de *scanpaths*.

La utilidad de los modelos de predicción y generación de datos que simulen correctamente el comportamiento humano es fundamental cuando para los profesionales del sector o los investigadores no es posible recabar datos reales de usuarios. Sin embargo, tanto los modelos de predicción de mapas de saliencia como los de predicción de *scanpaths* existentes, están o bien diseñados y entrenados para imágenes naturales y tareas de *free viewing* (recorrido libre), o se han entrenado con muy pocos datos y un objetivo particular. Por este motivo queremos saber cómo los modelos grandes generalizan con diferentes tipos de imágenes y tareas visuales. Para

ello se han implementado [1] diferentes modelos existentes y se han extraído métricas, analizando su comportamiento tanto con bases de datos públicas, como con un banco de pruebas que hemos creado. También se propone un modelo nuevo para tratar de mejorar la capacidad de generalización de los modelos de predicción de *scanpaths* actuales. Por último se plantean una serie de aplicaciones de los modelos de predicción de *scanpaths* y se discuten los resultados obtenidos.

1.2. Objetivo del proyecto

El objetivo de este trabajo es implementar algunos modelos recientes y analizar su comportamiento ante diferentes escenarios y estímulos visuales, de cara a sentar las bases para un nuevo modelo futuro que sea capaz de generalizar a diferentes tipos de imágenes y tareas visuales. Partiendo de la implementación de varios de los mejores modelos de predicción de *scanpaths* actuales, hemos diseñado un banco de pruebas y analizado el comportamiento de cada modelo ante diferentes situaciones, con el objetivo de encontrar sus límites y poder analizar las posibles situaciones de fallo. Debido a la falta de datos complejos con los que entrenar desde cero un nuevo modelo de predicción, proponemos un modelo capaz de adaptarse automáticamente a diferentes contenidos visuales con el mejor rendimiento posible de entre los modelos ya existentes utilizando un algoritmo de categorización de imágenes basado en *deep learning*. Además se ha plantean algunas posibles aplicaciones de lo estudiado.

1.3. Alcance del proyecto

El alcance del proyecto incluye:

- Estudio y exploración de literatura y trabajos previos relacionados con la atención visual y los modelos de predicción de mapas de saliencia y *scanpath* (Sección 2).
- Identificación y exploración de bases de datos de *human gaze behavior* o *eye tracking* en 2D e implmentación de un banco de pruebas propio sobre el que realizar un análisis más detallado. (Sección 3).
- Identificación, exploración e implementación de los modelos de predicción de *scanpaths* en 2D, así como la categorización de estos: que datos de entrada requieren y en qué se basan. (Sección 2.4).
- Análisis de resultados (Sección 4).
- Propuesta de un nuevo modelo de predicción de *scanpaths* (Sección 5).

[1] Una implementación es la ejecución o puesta en marcha de una idea programada, ya sea, de una aplicación informática, un plan, modelo científico, diseño específico, estándar, algoritmo o política.`https://en.wikipedia.org/wiki/Implementation`

1.4. Organización del proyecto

Este trabajo de fin de grado ha sido realizado en colaboración con el grupo de investigación *Graphics and Imaging Lab*, más concretamente con Sandra Malpica Mallo, la directora, y Diego Gutiérrez Pérez, el ponente. La Sección 1.3 incluye las diferentes tareas realizadas íntegramente por el autor de este trabajo de fin de grado.

1.5. Planificación

El proyecto se ha dividido en una serie de tareas relacionadas con la estructura inherente al trabajo. A cada una de esas tareas se le ha dedicado un total de horas que puede consultarse en la Figura 1.1. Se ha seguido una metodología iterativa a la hora de buscar e implementar distintos modelos y métricas cuando ha sido posible.

Etapa	Proceso	Horas
Trabajo previo	Estudio de la literatura existente	47
Bases de datos	Identificación de bases de datos	15
	Diseño de un banco de pruebas propio	20
Modelos	Identificación y categorización de modelos scanpaths	23
	Implementación	75
Evaluación	Métricas de mapas de saliencia	24
	Métricas de scanpaths	65
	Análisis	15
Diseño de un modelo propio	Diseño de un modelo de predicción scanpath	20
Documentación	Redacción de la memoria	25
TOTAL		329

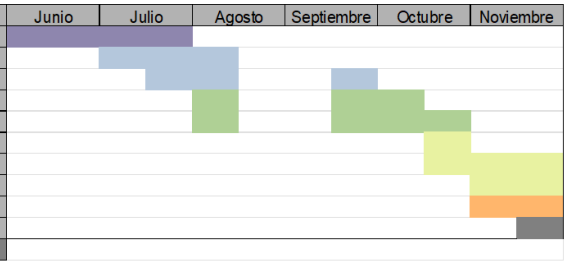

Figura 1.1: Horas dedicadas a cada tarea del proyecto.

Como se puede observar en el diagrama de Gantt, el trabajo tuvo una pausa entre los meses de agosto y septiembre. Este tiempo se tomó para finalizar con éxito las asignaturas pendientes del grado en Ingeniería Informática.

2. Marco teórico

2.1. Atención visual

La atención visual es un proceso cognitivo que facilita la detección de estímulos en una escena visual compleja, como la que habitualmente nos presenta el medio externo.

Manteniendo la mirada fija en un punto del campo visual somos capaces de atender a objetos situados en zonas periféricas al mismo, lo que se conoce como atención visual encubierta. Ésta implica la activación de conexiones frontales y parietales a la corteza visual, que aumentan su actividad y su capacidad perceptiva. Sin embargo, en este trabajo nos centramos en la atención visual directa (o *overt*), el tipo de atención relacionada con el punto al que se está mirando de forma explícita. La atención visual se ve modulada principalmente por dos factores: los estímulos visuales externos que se presentan ante el observador y los procesos cognitivos internos del mismo, incluyendo qué tarea está llevando a cabo en un momento determinado y su experiencia previa [2]. Aún no está claro cómo estos factores interactúan, pero sí que está claro que distintas tareas visuales se traducen en diferentes patrones de movimientos oculares, por lo que consideramos interesante estudiar cómo los modelos de predicción generalistas (entrenados mayoritariamente con datos de *free viewing*) se comportan a la hora de enfrentarse a estímulos visuales o tareas diferentes. A continuación, distinguimos los principales tipos de tareas visuales.

- Visualización libre (o *Free viewing*): se define como una tarea que no impone restricciones externas sobre qué ubicaciones o partes de un estímulo deben ser analizadas.

- Visualización guiada (o *Guided viewing*): consiste en guiar a los observadores el camino que debe seguir con la vista.

- Tareas relacionadas con la memoria (*Content Awareness Task* o CAT): Se pide al observador que recuerde alguna característica de la imagen para hacerles preguntas después o que siga un proceso determinado, normalmente relacionado con el contenido visual, a la vez que visualiza la imagen.

- IST (*Information Search Task*): al observador se le sugiere una tarea que le requiere realizar una búsqueda sobre el estímulo.

2.2. Mapa de saliencia

El objetivo de un mapa de saliencia (más conocido en inglés como *saliency map*) es reflejar el grado de importancia de un píxel para el sistema visual humano, es decir, la probabilidad de que el ojo se fije en esa región de la imagen. Este campo lleva siendo estudiado décadas, comenzando por el modelo seminal de *Itti & Koch* [3] hasta modelos más sofisticados y actuales como el modelo de *deep learning DeepGaze IIE* [4]. En la Figura 2.1 se puede ver un ejemplo de un mapa de saliencia.

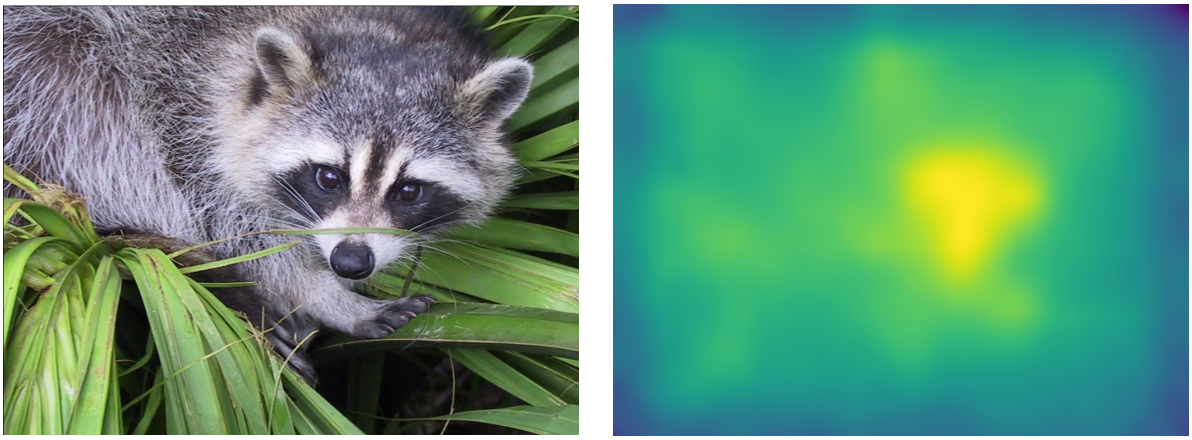

Figura 2.1: Ejemplo de generación de un mapa de saliencia. A la izquierda se muestra la imagen original y a la derecha el mapa de saliencia asociado, mostrado como un mapa de calor que indica qué zonas llaman más la atención.

2.3. Scanpath

Se denomina *scanpath* (Figura 2.2) a cualquier dato de movimiento ocular recopilado por un dispositivo de seguimiento de la mirada, donde se registra información sobre las trayectorias (o rutas) de los ojos cuando se escanea el campo visual, se visualiza y analiza cualquier tipo de información. Estos datos suelen consistir en la dirección de la mirada, la posición y duración de la fijación y la duración de la sacada. En nuestro caso, trabajamos con la secuencia temporal de coordenadas 2D en la imagen correspondientes al punto en el que se fija el usuario en cada momento, además de las duraciones o perfil temporal de cada coordenada. En la Figura 2.2 se puede ver un ejemplo de la generación de un *scanpath*.

Tras realizar un estudio en anchura del campo del comportamiento visual humano y sus diferentes modelos de predicción, hemos decidido crear una taxonomía basada en la inspiración que los autores han tomado a la hora de diseñar estos modelos. Al implementar los modelos de predicción utilizados en este trabajo se han escogido modelos actuales y representativos de cada una de estas categorías. La taxonomía es la siguiente:

- Inspiración biológica: se inspiran en los resultados de la neurociencia y la ciencia de la visión. Intentan replicar el comportamiento humano replicando procesos neurológicos, modelando estudios de comportamiento, etc.

Figura 2.2: Ejemplo de generación de un *scanpath* con el modelo *G-eymol (segundos=2, fps=40, coeff.alpha=0.3)*. En la figura de la izquierda se puede observar la imagen original, y en la figura de la derecha la misma imagen con el *scanpath* superpuesto.

- Inspiración estadística: intentan reproducir ciertas propiedades estadísticas de las rutas de exploración humanas. Se basan en las propiedades estadísticas generales de cada tipo de visualización, realizando ciertas asunciones sobre la generalización de comportamiento.
- Inspiración cognitiva: tratan de tener en cuenta factores cognitivos que afectan la posición de la mirada. Basadas en características de alto nivel o los procesos cognitivos que dirigen cómo presta atención el usuario a su entorno.
- Modelos de ingeniería: modelos basados en aprendizaje profundo que se ajustan a los datos. Estos modelos son agnósticos al proceso subyacente ya sea cognitivo o fisiológico. Dependen de obtener una cantidad de datos lo suficientemente grande para aprender comportamientos comunes.

2.4. Modelos de predicción

Después de un estudio en profundidad del campo [5], con la condición de que el código y los datos de cada modelo fueran públicos, y buscando cubrir las diferentes inspiraciones posibles de nuestra taxonomía (Sección 2.3), se han implementado en total siete modelos en este trabajo: tres modelos de predicción de saliencia y cuatro modelos de predicción de scanpaths. La importancia de los mapas de saliencia en este trabajo reside en que muchos de los modelos de predicción de *scanpaths* requieren de mapas de saliencia como parámetro de entrada para generar los *scanpaths*, Tal y como se muestra en el modelo *CLE* (Sección 2.4.6).

Cabe destacar que para todas las implementaciones, tanto para los modelos de predicción de saliencia, como para los modelos de predicción de *scanpaths*, ha sido necesaria la creación de programas en Python o Matlab para su puesta en marcha. En algunos casos ha sido necesario realizar cambios en el código base del modelo para adaptarlo. Como se menciona en la sección 4, se han implementado scripts de automatización para cada uno de estos modelos. Además también se ha automatizado el tratamiento de los datos de entrada y cálculo de métricas para el análisis de los bancos de pruebas. A continuación se presentan en detalle cada uno de los modelos

implementados y se muestran un conjunto de resultados generados con dichos modelos. En la Figura 2.3 se pueden ver para una imagen de un puerto, una cara, una imagen fractal y una de natualeza, como se domportan los diferentes modelos de saliencia implementados. Por otro lado, en la Figura 2.4 se puden ver para un cojunto de imágenes urbanas, el comportamiento obtenido por los diferentes modelos implementados y el *ground truth*[1] de las fijaciones de los usuarios.

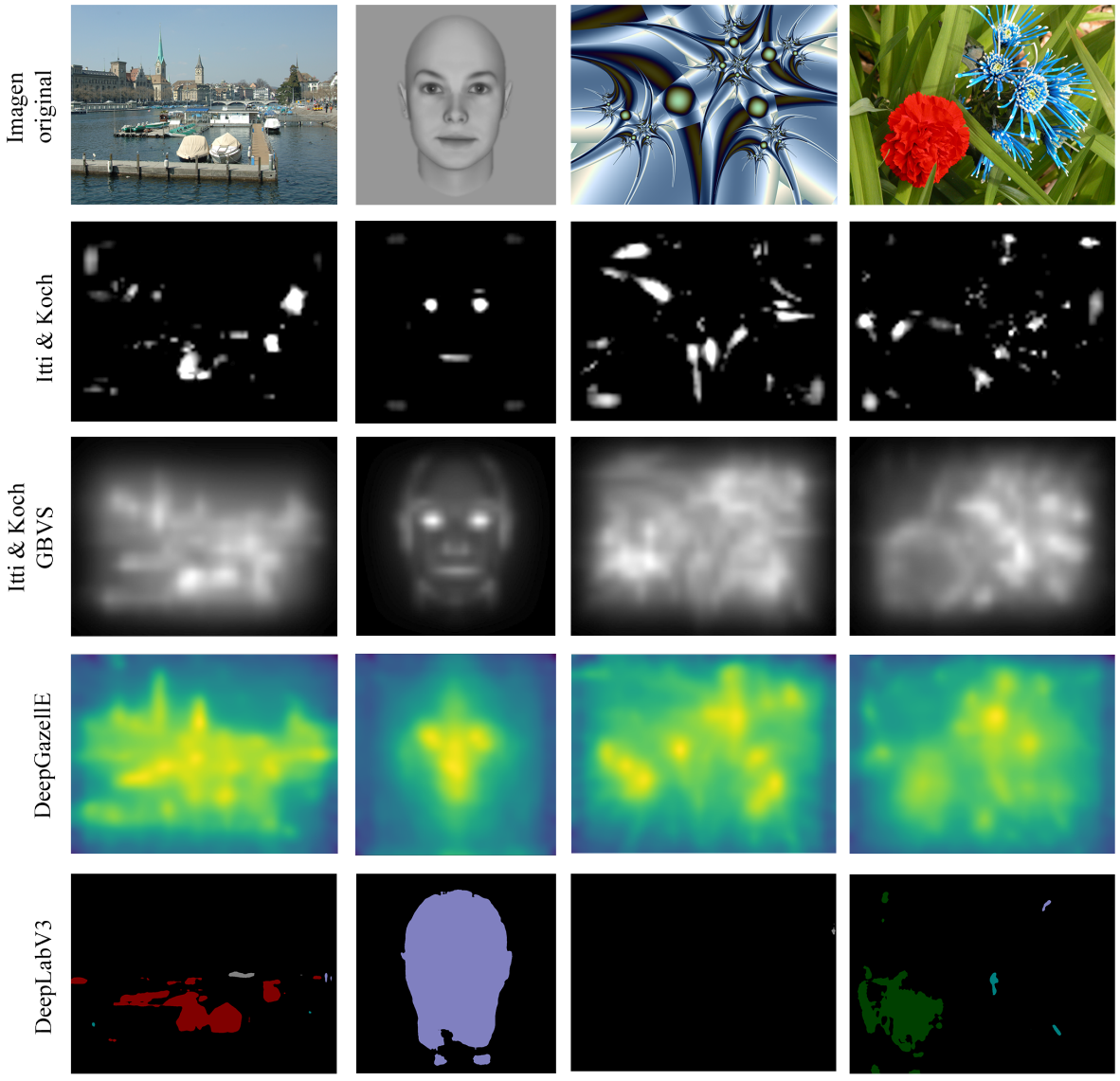

Figura 2.3: Imágenes originales y sus respectivos mapas de saliencia para cada modelo de predicción de saliencia implementado.

[1]Información que se sabe que es real o verdadera, proporcionada por observación y medición directas (es decir, evidencia empírica). `https://en.wikipedia.org/wiki/Ground_truth`

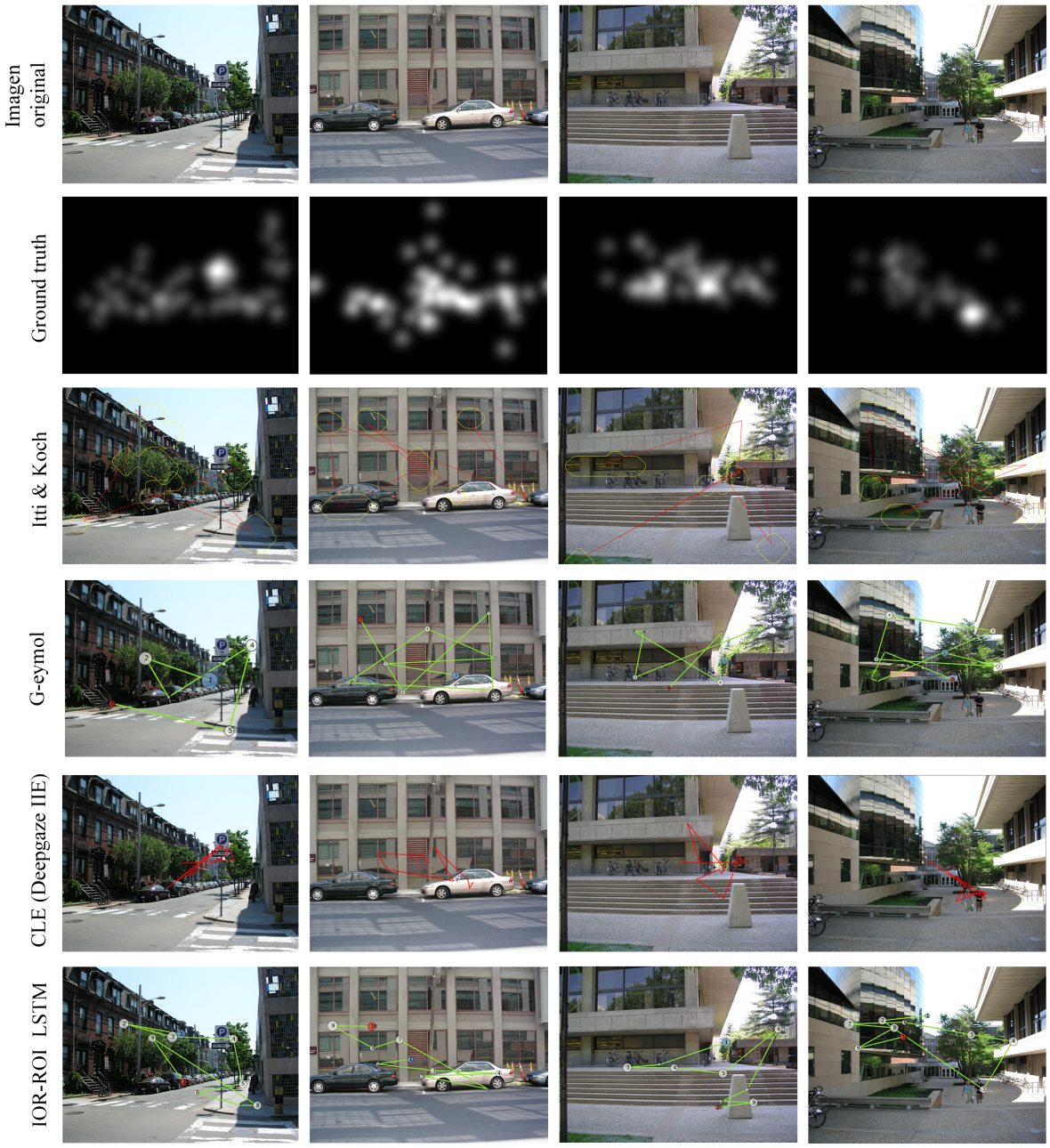

Figura 2.4: Imágenes originales, y sus respectivos *ground truth* y scanpaths generados por cada modelo de predicción de *scanpaths*.

2.4.1. Itti & Koch

El modelo seminal propuesto por *Laurent Itti* y *Christof Koch* se trata de un modelo de inspiración biológica de predicción de mapas de saliencia. Las características de la imagen (color, intensidad, orientación) se combinan en un solo mapa de prominencia topográfica. Luego, una red neuronal dinámica biológicamente plausible selecciona las ubicaciones atendidas en orden de

prominencia decreciente utilizando una red *WTA (Winner-Take-All).*

El *WTA* es un principio computacional que se aplica en modelos de redes neuronales mediante el cual las neuronas de una capa compiten entre sí por la activación. En la forma clásica, solo la neurona con la activación más alta permanece activa mientras que todas las demás neuronas se apagan; sin embargo, otras variaciones permiten que más de una neurona esté activa, por ejemplo, el *soft winner*[2], mediante el cual se aplica una función de potencia a las neuronas.

Este modelo de predicción de saliencia es antiguo y ha sido superado por los modelos más recientes, tal y como indica el *MIT/Tübingen Saliency Benchmark* [6], un banco de pruebas de mapas de saliencia existente. Sin embargo, se incluye en este trabajo por dos motivos: como base histórica, y porque experimentalmente se ha comprobado que su tiempo de ejecución es notablemente bajo en comparación con otros modelos más pesados, como los de *deep learning.* En la Figura 2.5 se muestra un ejemplo de un mapa saliencia generado con este modelo.

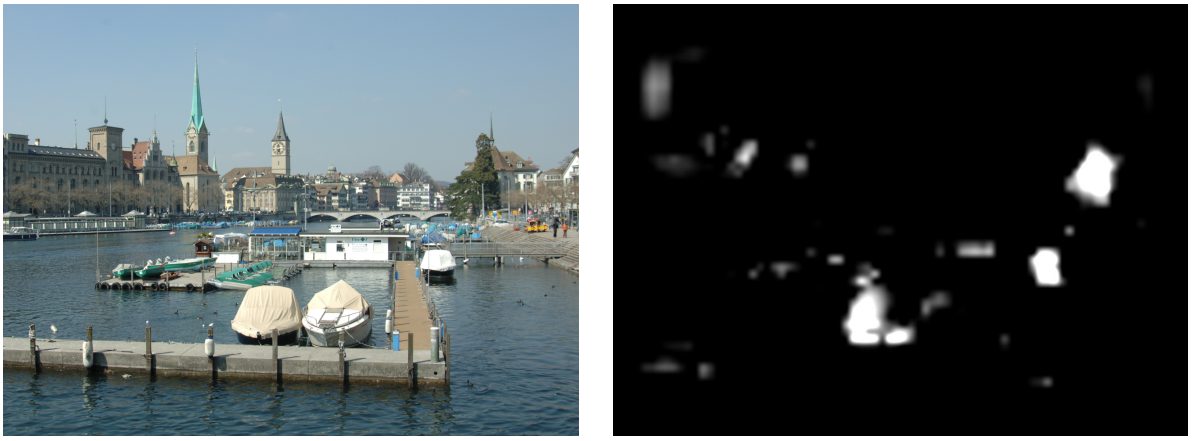

Figura 2.5: En la imagen de la izquierda se puede observar la imagen original, y en la imagen de la derecha el mapa de saliencia obtenido con una configuración estándar de *Itti & Koch.*

En cuanto a la implementación de este modelo, hemos empleado el código fuente que se encuentra en el paquete *Saliency Toolbox* [7] escrito en *Matlab.* Este paquete proporciona una colección de funciones y scripts para calcular mapas de saliencia. En este caso se han utilizado los parámetros estándar ya que son ampliamente utilizados.

2.4.2. Itti & Koch GBVS

El modelo de inspiración biológica *Itti & Koch GBVS (Graphic-Based Visual Saliency)* es una re-implementación del modelo mencionado previamente y que está incluido en el paquete *GBVS Toolbox* [8] para *Matlab.* Como en el modelo anterior, hemos utilizado los parámetros estándar. Este modelo adaptado funciona más rápido y proporciona mejores resultados (Figura 2.6) que su predecesor tal y como se muestra en el *MIT/Tüebingen Saliency Benchmark* [6].

[2] `https://en.wikipedia.org/wiki/Winner-take-all_(computing)`

Figura 2.6: En la imagen de la izquierda se puede observar la imagen original, y en la imagen de la derecha el mapa de saliencia obtenido con una configuración estándar de *Itti & Koch GBVS*.

2.4.3. DeepGaze II y DeepGaze IIE

El modelo de *Itti & Koch* fue el primero en predecir un mapa de saliencia a partir de cualquier imagen arbitraria sin la necesidad de precalcular características elementales y que permite una amplia gama de aplicaciones. Esto allanó el camino para muchos modelos interesantes de predicción de saliencia que conducen a la actualidad, donde los modelos de *Deep Learning* dominan el campo impulsados por conjuntos de datos de saliencia a gran escala.

Los modelos *DeepGaze II* [9] y *DeepGaze IIE* [4] son algunos de ellos. La diferencia entre estos dos modelos es que el segundo es una versión más reciente que se compone de un cojunto (*Ensemble*) de varias redes neuronales. Su predecesor *DeepGazeII* solo está compuesta por una. Tal y como se muestra en el *MIT/Tüebingen Saliency Benchmark* [6], *DeepGaze IIE* es hasta la fecha el modelo de predicción de saliencia con mejores resultados, por lo que se ha optado por implementar esta versión del modelo. En la Figura 2.8 se muestra un ejemplo de un mapa saliencia generado con este modelo.

En la imagen de la derecha de la Figura 2.7, se puede observar el *pipeline* general del modelo *DeepGaze IIE*, donde el modelo final se deriva de la combinación de múltiples redes neuronales. En la parte de la izquierda de la misma figura se puede observar la aquitectura para *DeepGaze II*, donde primero se procesa una imagen con una *CNN (Convolutional Neural Network)* para extraer activaciones profundas, que posteriormente se procesan en una red de lectura de convoluciones 1×1. El canal único de salida de la red de lectura se difumina, se combina con un *centerbias* [3] y se alimenta a través de una función *softmax* [4] para producir una distribución de fijación bidimensional (Figura 2.7). Esencialmente, *DeepGaze IIE* es una adaptación de la arquitectura de *DeepGaze II* llevándola a un nuevo estado, donde se combinan algunos de los *backbones* de ImageNet [5] de última generación, aprovechando la complementariedad entre modelos.

[3] Traducido como sesgo central.

[4] En matemáticas, la función softmax, o función exponencial normalizada, es una generalización de la función logística. `https://es.wikipedia.org/wiki/Funci\%C3\%B3n_SoftMax`

[5] El proyecto ImageNet es una gran base de datos diseñada para la investigación de software de reconocimiento de objetos. `https://en.wikipedia.org/wiki/ImageNet`

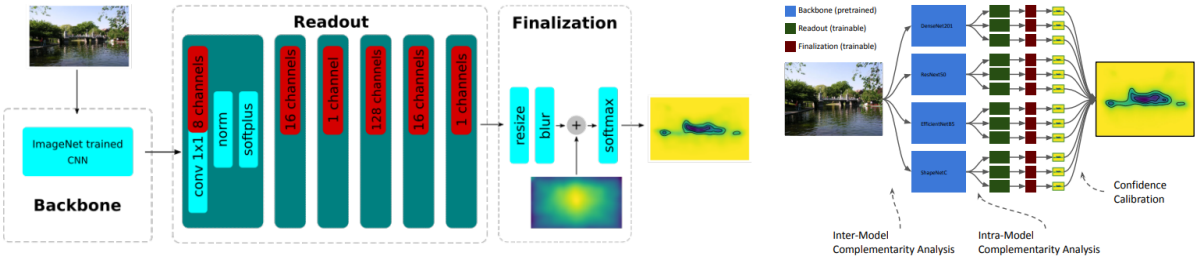

Figura 2.7: En la parte izquierda se muestra la arquitectura del modelo *DeepGaze II* y a la derecha la arquitectura del modelo *DeepGaze IIE*.

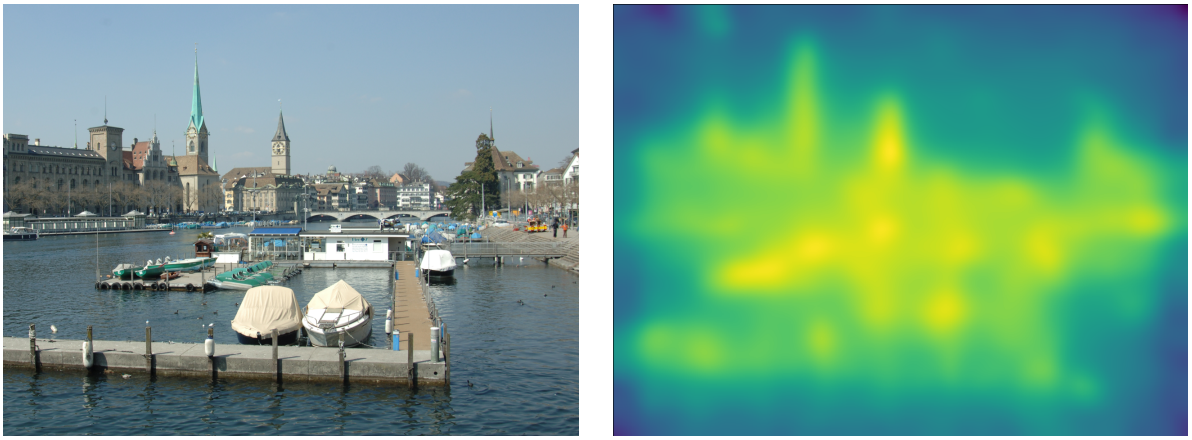

Figura 2.8: En la imagen de la izquierda se puede observar la imagen original, y en la imagen de la derecha el mapa de saliencia obteniedo con el modelo *DeepGaze IIE* entrenado con la base de datos *MIT1003*.

En cuanto a su implementación, hemos empleado el código fuente proporcionado por sus autores en Python que utiliza la librería *PyTorch*. Se ha utilizado el modelo ya entrenado que se encuentra en el código fuente ya que habia sido entrenado con la base de datos *MIT1003*, que constituye el caso base para el análisis de los modelos implementados. También se ha utilizado un *centerbias* para esta base de datos que al igual que el modelo, ha sido entrenado para la base de datos *MIT1003*.

2.4.4. Itti & Koch

El modelo seminal de Itti-Koch se utiliza habitualmente hoy en día como un modelo de prominencia espacial que predice mapas de prominencia. Pero el modelo en realidad incluye un mecanismo de selección de fijación en el que se desarrolla un mapa de saliencia con una red *WTA* (*Winner-Take-All*) con inspiración biológica y sobre el cual se itera para generar un scanpath. En la Figura 2.9 se muestra un ejemplo de un *scanpath* generado con este modelo.

Figura 2.9: En la imagen de la izquierda se puede observar la imagen original, y en la imagen de la derecha la misma imagen con el dibujo del *scanpath* obtenido con el modelo *Itti & Koch* con una red *WTA*.

2.4.5. G-eymol

Otro modelo de predicción de *scanpaths* implementado, también de inspiración biológica es el modelo de nombre *G-eymol* [10]. Define un *scanpath* como un proceso dinámico que puede interpretarse como una ley variacional relacionada de alguna manera con los procesos cognitivos tras la atención visual, donde el foco de atención está sujeto a un campo gravitacional. La masa virtual distribuida que impulsa los movimientos oculares está asociada con la presencia de detalles y movimiento en el video. A diferencia de la mayoría de los modelos actuales, este enfoque propuesto no estima directamente el mapa de saliencia, pero la predicción de los movimientos oculares permite integrar a lo largo del tiempo las posiciones de interés.

G-eymol modela el movimiento de la mirada mediante un sistema de ecuaciones diferenciales de segundo orden que tiene en cuenta el gradiente de brillo y el gradiente de flujo óptico (que es cero para imágenes estáticas). Las rutas de exploración se generan simulando la posición de la mirada de acuerdo con la ecuación diferencial y luego aplicando algoritmos de detección de fijaciones. En la Figura 2.10 se pude ver el ejemplo de un *scanpath* generado con este modelo.

2.4.6. CLE

El modelo de predicción de *scanpaths* y de inspiración estadística *CLE* [11] (*Constrained Levy Exploration*), modela movimientos sacádicos mediante *Levy Flight* [6] utilizando una distribución *Cauchy* [7]. Sin embargo, a diferencia de utilizar una distribución de *Levy flight* pura, *CLE* modula la distribución de saltos con un mapa de saliencia: el punto de inicio de la distribución de saltos se mueve desde la útlima ubicación de fijación a lo largo del gradiente buscando otras fijaciones potenciales. Además los objetivos sacádicos con mayor saliencia tienen una mayor probabilidad de ser seleccionados.

Puesto que este modelo requiere como entrada un mapa de saliencia, además de la imagen

[6] tipo de paseo aleatorio en el cual los incrementos son distribuidos de acuerdo a una distribución de probabilidad de cola pesada. `https://en.wikipedia.org/wiki/L\%C3\%A9vy_flight`

[7] distribución de probabilidad continua. `https://en.wikipedia.org/wiki/Cauchy_distribution`

Figura 2.10: En la imagen de la izquierda se puede observar la imagen original, y en la imagen de la derecha la misma imagen con el dibujo del *scanpath* obtenido con el modelo *G-eymol (segundos=2, fps=40, coeff.alpha=0.3)*. Cabe destacar que este modelo predice la duración de cada una de las fijaciones (el tamaño del círculo correspondiente a cada fijación).

original, se han utilizado como base para la generación de *scanpaths* los diferentes modelos de generación de mapas de saliencia implementados, explicados anteriormente en esta sección. Para ello se ha adaptado parte del código fuente del modelo implementado en Python y se han creado nuevos scripts para automatizar las tareas. En la figura 2.11 se puede ver un ejemplo de un *scanpath* generado con este modelo.

Figura 2.11: En la imagen de la izquierda se puede observar la imagen original, y en la imagen de la derecha la misma imagen con el dibujo del *scanpath* obtenido con el modelo *CLE (utilizando DeepgGazeIIE como generador del mapa de saliencia previo) (numSteps=15)*.

2.4.7. IOR-ROI LSTM

Como modelo representativo de la inspiración cognitiva y basado en datos nos encontramos con el *IOR-ROI LSTM* [12]. Utiliza modelos de *Deep Learning* para codificar una imagen en características y máscaras semánticas predichas previamente. Estos datos alimentan una arquitectura de red neuronal recurrente que inhibe ciertas características de la imagen y predice la siguiente región de interés mediante una mezcla de gaussianos. Esta predicción se combina con

sesgos para seleccionar la siguiente ubicación de fijación. En la Figura 2.13 se puede ver un ejemplo de un *scanpath* generado con este modelo.

Como los autores no proporcionan un segmentador funcional para su modelo, ha sido necesario buscar, implementar y adaptar otro segmentador de características similares para hacer funcionar el modelo. En concreto, se ha empleado el *DeepLabV3 ResNet50* [13]. El cual funciona considerablemente bien, tal y como muestra la Figura 2.12, donde es capaz de reconcoer los botes de la imagen además de algunas personas al fondo a la derecha.

Figura 2.12: En la imagen de la izquierda se puede observar la imagen original, y en la imagen de la derecha el mapa semántico obtenido a partir de la imagen de la izquierda con el modelo de generación semántico *DeepLabV3 ResNet50*.

2.5. Métricas

Se han evaluado los resultados recogidos por los modelos implementados en base a diferentes bancos de pruebas, comparando los *scanpaths* generados con los datos reales del usuario (*ground truth*). La Sección 3 explica los datos utilizados con más detalle. A continuación, se describen las métricas empleadas, ampliamente utilizadas a la hora de evaluar mapas de saliencia o *scanpaths* [14]:

2.5.1. AUC

La curva *ROC* [8] (*Receiver Operating Characteristic*) y el *AUC* (*Area Under the Curve*) se utilizan a menudo para evaluar los pros y los contras de un clasificador binario. La curva *ROC* se denomina curva característica de funcionamiento del receptor. *AUC* es el área bajo la curva *ROC*, y su área no será nunca mayor que 1. Dado que la curva *ROC* no se puede emplear para juzgar apropiadamente la calidad del modelo de clasificador, el valor *AUC* se utiliza para evaluar y comparar el modelo de clasificador. Generalmente, cuanto mayor sea el valor AUC, mejor será el rendimiento del clasificador.

[8] `https://en.wikipedia.org/wiki/Receiver_operating_characteristic`

Figura 2.13: Ejemplo de generación *scanapth* con el modelo *IOR-ROI LSTM* y *DeepLabV3 ResNet50*. Los números indican el orden de las fijaciones, y el tamaño la duración de las mismas.

Hemos implementado esta métrica para evaluar los diferentes modelos de predicción de saliencia, es decir, cómo de bien predicen la saliencia para un conjunto de datos en concreto, siguiendo la forma de analizar mapas de saliencia más común que hemos observado en el campo.

2.5.2. NSS

La métrica *NSS* (*Normalized Scanpath Saliency*) [15], junto con la mencionada previamente, *AUC*, son dos métricas ya establecidas en el campo de la predicción de mapas de saliencia. *NSS* es una métrica de correspondencia normalizada, calcula la saliencia promedio de las ubicaciones fijadas después de normalizar el mapa de saliencia para que tenga media cero y varianza unitaria. Cuanto mayor sea su valor, mejor es el modelo.

2.5.3. Euclidean distance

En cuanto a métricas de *scanpaths* en primer lugar se implementó la distancia euclídea, que aplicado a este caso de uso se trata de la distancia en el plano imagen entre *scanpaths*. No es una métrica que aporte mucha información aplicado a *scanpaths*, ya que no tiene en cuenta otras

cosas como el tiempo o la correlación entre las fijaciones, pero se ha elegido incluirla de todas formas para poder proporcionar una análisis más completo. En este caso, cuanto mayor sea el valor obtenido, peor es el modelo.

2.5.4. ScanMatch

Otra métrica para *scanpaths* ampliamente reconocida es la métrica *ScanMatch* [16], que emplea el algoritmo *Needleman-Wunsch* [9] para realizar la alineación de los *scanpaths* teniendo en cuenta las distancias entre las distintas fijaciones, así como de la duración de cada una de ellas. Para esta métrica cuanto mayor sea su valor, mejor es el modelo.

2.5.5. DTW

La métrica *DTW* (*Dynamic Time Warping*) empleada para *scanpaths* es una métrica de similitud entre series temporales, en este caso, dos *scanpaths*. Conceptualmente, es similar al algoritmo *Needleman-Wunsch*, en tanto a que ambos realizan una matriz de disimilitud, con las distancias entre todos los miembros de una relación, como manera de calcular la distancia óptima entre los miembros de un grupo. Cuanto mayor sea su valor, peor es el modelo. Cabe destacar que la métrica DTW es de particular interés ya que incluye la dimensión temporal a la hora de comparar varios scanpaths y no sólo su distancia espacial, al igual que ScanMatch.

[9] `https://es.wikipedia.org/wiki/Algoritmo_Needleman-Wunsch`

3. Diseño de un banco de pruebas propio

Uno de los *benchmarks* más conocidos en el campo de los mapas de saliencia es el *MIT/Tübingen Saliency Benchmark* [6]. Este banco de pruebas ofrece una serie de bases de datos relacionados con la atención visual. Las bases de datos que ofrecen están compuestas principalmente por imágenes naturales y datos de usuario durante tareas de *free viewing*. En concreto, la base de datos *MIT300* la emplean para realizar su *benchmarking* y obtener métricas de rendimiento para todo el que quiera enviarles su modelo. Para mantener la veracidad de sus resultados, este subconjunto de datos no está disponible públicamente por lo que no se ha podido utilizar para las métricas: el *ground truth* no es público para evitar que los modelos que propongan investigadores se entrenen con los datos que se usan de test. Durante este trabajo hemos estudiado el banco de pruebas del MIT para escoger los mejores modelos de mapas de saliencia y poder utilizarlos como base para el modelo CLE. La diferencia entre este banco de pruebas y el nuestro, es que su objetivo principal consiste en la evaluación de modelos de mapas de saliencia.

3.1. Bases de datos existentes

Aunque las imágenes naturales y los datos de *free viewing* no forman parte del objetivo principal de este trabajo, se decidió emplear la base de datos *MIT1003* [17] para probar el correcto desempeño de los modelos y poder analizarlos. El *MIT300* es un subconjunto de esta base de datos, por lo que su utilidad a la hora de evaluar el rendimiento de modelos de predicción de la atención visual está más que demostrado. Además, el modelo de generación de mapas de saliencia *DeepGaze IIE*, está entrenado con esta base de datos. Utilizando la base de datos *MIT1003* como parte de nuestro banco de pruebas nos aseguramos una primera validación ecológica de la implementación de todos nuestros modelos antes de probar su capacidad de generalización ante estímulos de distinta naturaleza. El *benchmark* del *MIT/Tübingen Saliency Benchmark* también referencia otras bases de datos de movimientos oculares, pero la más usada es la del MIT1003, por lo que es la que hemos escogido como base.

Esta base de datos esta compuesta por datos de seguimiento ocular (*eye tracking*[1]) de 15 observadores, de entre 18 y 35 años, en 1003 imágenes naturales durante 3 segundos en cada una.

[1]Proceso de evaluar y registrar el punto donde se fija la mirada en cada momento.

3.2. Diseño de un banco de pruebas propio

Debido a que los modelos existentes estan entrenados para imágenes naturales y principalmente para tareas de visualización libre, hemos decidido crear nuestro propio banco de pruebas. De esta forma podemos someter a los modelos implementados a imágenes más complejas y ver cómo se comportan, si fallan y en caso de hacerlo por qué.

Para ello hemos añadido a la base del MIT 1003 parte de los datos recapitulados por un macroestudio hospitalario, publicado en un artículo de *Dryad* [18]. De esta manera, se han escogido 126 imágenes, siendo estas representativas de varias categorías en general alejadas de las imágenes naturales o con tareas visuales novedosas.

- Faces: 32 imágenes de rostros modelados en 3D.
- Fractals: 10 imágenes fractales, es decir, objetos geométricos cuya estructura básica, fragmentada o aparentemente irregular, se repite a diferentes escalas.
- Illusion: 24 imágenes lineales (dibujos en blanco y negro) que producen ilusiones ópticas al tener más de una posible interpretación.
- Landscape: 10 imágenes de paisajes.
- Nature: 10 imágenes de natualeza.
- Pink noise: 10 imágenes de ruido rosa. Imágenes abstractas de colores vivos.
- Urban: 20 imágenes de paisajes urbanos.
- Web: 10 imágenes de páginas web.

En la Figura 3.1 se pueden observar las categorías que componen el estudio.

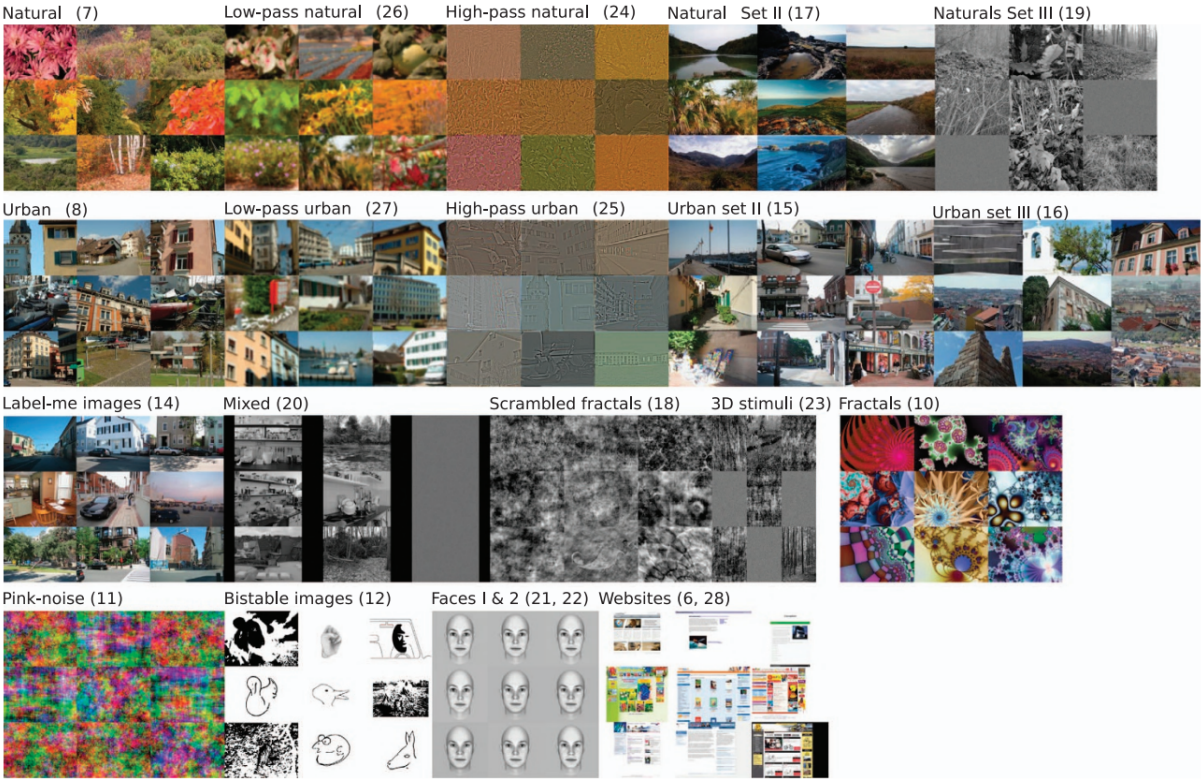

Figura 3.1: Categorías de las imágenes reacapituladas por el artículo del *Dryad*: *An extensive dataset of eye movements during viewing of complex images.*

4. Resultados

Para obtener todas las métricas mostradas a continuación se han escrito *scripts* y programas de automatización en los lenguajes de *Python* o *Matlab*, según los modelos lo requiriesen. Se ha uniformizado el formato de los datos de entrada y del proceso de alimentación de cada uno de los modelos, así como la generación de *scanpaths* para cada una de las imágenes del banco de pruebas y el cálculo de todas las métricas escogidas para el análisis.

Cabe destacar que para la mayoría de métricas se requería que las longitudes de los *scanpaths* fueran iguales, por lo que tratamos siempre de generar suficientes fijaciones con los modelos de predicción para luego normalizar la longitud del *scanpath*, adaptándonos a la que el *ground truth* nos requirese. El motivo de esta variabilidad se encuentra en el comportamiento humano natural. Dos observadores distintos, o incluso un mismo observador en diferentes ocasiones resultarán en *scanpaths* diferentes, lo que incluye no sólo las coordenadas si no también el número de fijaciones.

En primer lugar se decidió analizar y comparar los resultados de los diferentes modelos de mapas de saliencia para la base de datos *MIT1003*. Aunque ya teníamos una idea aproximada de su desempeño [6], se obtuvieron las métricas de estos modelos para corroborar que el modelo basado en *Deep Learning*, *DeepGaze IIE*, es sin lugar a duda el mejor modelo de mapas de saliencia hasta la fecha, siendo incluso dos veces mejor que el modelo seminal de *Itti & Koch* (basandonos en la métrica AUC). Estos resultados se pueden apreciar con más detalle en la Tabla 4.1.

Una vez obtenidas las métricas para los mapas de saliencia, se diseñaron e implementaron los *scripts* para calcular ahora las de los modelos de predicción de *scanpaths*. Para los casos del modelo determinista *G-eymol*, se calculó para cada imagen de la base de datos *MIT1003* con el parámetro del *coeficiente alpha* con valor 0.3 tal y como sus autores proponen. Pero el parámetro de los *fps* lo redujimos de 60 a 30 debido a que el tiempo de cálculo era muy elevado. Por último para el parámetro de los segundos de simulación de la generación del *scanpath* establecimos 3

	Datos	AUC↑	NSS↑
Itti&Koch	MIT1003	0.4505	0.5194
Itti&Koch (GBVS)	MIT1003	0.8065	1.3656
DeepGazeIIE	MIT1003	**0.9108**	**1.6108**

Tabla 4.1: Métricas obtenidas con los diferentes modelos de predicción de mapas de saliencia para la base de datos *MIT1003*.

	MIT1003		
Modelo	Euc. Dis.↓	ScanMatch↑	DTW↓
G-eymol	**361.58**	0.3701	**1329.96**
CLE (Itti & Koch)	1030.13	0.2397	2534.04
CLE (Itti & Koch GBVS)	897.85	0.2521	2242.01
CLE (DeepGaze IIE)	876.61	0.2583	2191.99
IOR-ROI LSTM	812.93	**0.4014**	1728.75

Tabla 4.2: Métricas obtenidas con los diferentes modelos de predicción *scanpath* para el base de datos *MIT1003*.

	Faces		
Modelo	Euclidean.distance↓	ScanMatch↑	DTW↓
G-eymol	604.50	0.1929	1072.64
CLE (Itti&Koch)	510.15	0.3681	685.85
CLE (Itti&Koch GBVS)	**437.79**	**0.3946**	**537.34**
CLE (DeepGazeIIE)	448.55	0.3762	564.74
IOR-ROI LSTM	614.36	0.2553	1063.18

Tabla 4.3: Métricas obtenidas con los diferentes modelos de predicción *scanpath* para lascategorías *Faces* y *Fractal* del banco de pruebas diseñado.

segundos, igual que los usuarios que conformaban el *ground truth*.

Para los modelos no deterministas *CLE* y *IOR-ROI LSTM*, se generaron por cada imagen un total de 10 *scanpaths*, y se compararon con los 15 *scanpaths* de cada imagen de la base de datos de *MIT1003*. Además para el modelo *CLE* alimentamos la entrada del mapa de saliencia con los tres modelos implementados.

A continuación, se muestran y discuten los resultados obtenidos para los modelos de predicción de *scanpaths* con la base de datos *MIT1003*. Como se puede observar en la tabla 4.2, los mejores resultados obtenidos serían por los modelos *G-eymol* y *IOR-ROI LSTM*. Aunque este segundo obtiene un valor mejor que el primero en la métrica *ScanMatch*, es peor a la hora de comparar sus distancias euclídeas y el *DTW*.

En cuanto al modelo *CLE*, se puede apreciar como para una base de datos general como lo es *MIT1003*, se obtienen mejore resultados utilizando *DeepGazeIIE* como entrada de mapas de saliencia. Este resultado es esperable, ya que *DeepGazeIIE* genera el mejor mapa de saliencia posible de los tres modelos implementados.

En las siguientes tablas (Tabla 4.3, 4.4, 4.5, 4.6, 4.7, y 4.8), se muestran los resultados obtenidos con los modelos implementados para el banco de pruebas diseñado. Como se puede observar, entre todas la categorías, no hay un modelo superior al resto, sino que cada modelo funciona mejor que otro dependiendo de las circunstancias. Estos resultados fueron clave a la hora de diseñar la arquitectura de la propuesta de modelo (Sección 5).

Cabe destacar que las categorías de *Illusion* y *Landscape* del banco de pruebas no han sido añadidas debido a que la mayoría de modelos no conseguían leer las imagenes por problemas

	Fractal		
Modelo	Euclidean distance↓	ScanMatch↑	DTW↓
G-eymol	1186.60	0.2678	2489.47
CLE (Itti&Koch)	652.57	0.1653	874.34
CLE (Itti&Koch GBVS)	667.65	0.1602	909.09
CLE (DeepGazeIIE)	**638.24**	0.1733	**870.95**
IOR-ROI LSTM	1525.86	**0.2943**	3755.11

Tabla 4.4: Métricas obtenidas con los diferentes modelos de predicción *scanpath* para la categoría *Fractal* del banco de pruebas diseñado.

	Nature		
Modelo	Euclidean distance↓	ScanMatch↑	DTW↓
G-eymol	1430.61	0.2858	2885.85
CLE (Itti & Koch)	610.92	0.2707	839.02
CLE (Itti & Koch GBVS)	614.24	0.2435	828.15
CLE (DeepGaze IIE)	569.71	0.2053	783.39
IOR-ROI LSTM	1531.79	0.2256	3446.31

Tabla 4.5: Métricas obtenidas con los diferentes modelos de predicción *scanpath* para la categoría *Nature* del banco de pruebas diseñado.

	Urban		
Modelo	Euclidean distance↓	ScanMatch↑	DTW↓
G-eymol	1344.64	0.2512	2896.54
CLE (Itti & Koch)	787.38	0.2095	1053.88
CLE (Itti & Koch GBVS)	729.62	0.1680	974.98
CLE (DeepGaze IIE)	685.98	0.1784	920.00
IOR-ROI LSTM	1647.46	0.2657	4686.06

Tabla 4.6: Métricas obtenidas con los diferentes modelos de predicción *scanpath* para la categoría *Urban* del banco de pruebas diseñado.

	Pink noise		
Modelo	Euclidean distance↓	ScanMatch↑	DTW↓
G-eymol	1293.15	0.2570	2948.51
CLE (Itti&Koch)	601.11	0.2658	804.91
CLE (Itti&Koch GBVS)	593.30	0.1850	807.87
CLE (DeepGazeIIE)	598.68	0.2113	829.46
IOR-ROI LSTM	1394.55	0.2540	3310.82

Tabla 4.7: Métricas obtenidas con los diferentes modelos de predicción *scanpath* para la categoría *Pink noise* del banco de pruebas diseñado.

	Web		
Modelo	Euclidean distance↓	ScanMatch↑	DTW↓
G-eymol	1150.26	0.2250	2486.59
CLE (Itti&Koch)	849.57	0.1559	1111.55
CLE (Itti&Koch GBVS)			
CLE (DeepGazeIIE)	728.14	0.0986	978.98
IOR-ROI LSTM		0.1752	5282.31

Tabla 4.8: Métricas obtenidas con los diferentes modelos de predicción *scanpath* para la categoría *Web* del banco de pruebas diseñado.

de compatibilidad con el diseño interno de la implementación. Tampoco se han añadido los resultados de los *scanpaths* generados por el modelo de generación de *scanpaths* de *Itti & Koch* debido a que solo pudo analizar la mitad de los conjuntos de datos, nuevamente por problemas relacionados con la lectura de las imágenes. Sin embargo, es de esperar que el modelo de *Itti & Koch* no sea mejor que el resto de modelos analizados, de la misma manera que el modelo de *Itti & Koch* de generación de mapas de saliencia es superado por la mayoría de modelos actuales.

5. Diseño de un modelo propio

Como se puede comprobar en la sección de resultados, no podemos afirmar categóricamente que ninguno de los modelos implementados presente un rendimiento superior al resto a la hora de demostrar su capacidad de generalización ante distintos tipos de imágenes o tareas visuales. Lo que es más, tampoco podemos predecir cómo estos modelos se comportarían ante otros estímulos no probados en este trabajo. Debido a la cantidad reducida de datos disponibles para estas situaciones que van más allá de las imágenes naturales o la tarea visual de *free-viewing* (Sección 2.1) no es trivial entrenar un modelo que conserve el rendimiento de los modelos actuales ante imágenes naturales y sea a la vez capaz de generalizar para las situaciones menos comunes. A modo de ejemplo, durante este trabajo se ha entrenado el modelo *PathGan*[19] únicamente con los datos disponibles del banco de pruebas. Sin embargo, su rendimiento ha sido notablemente inferior al del resto de modelos existentes, por lo que se ha descartado su uso.

Alternativamente, y tomando inspiración del modelo DeepGazeIIE (basado en un conjunto de redes neuronales en vez de en un único modelo para predecir mapas de saliencia), se ha decidido crear una especie de metamodelo de predicción de *scanpaths*, capaz de elegir de forma autónoma qué modelo subyacente utilizar en cada ocasión dependiendo de las características de la imagen de entrada. En concreto, utilizamos como capa superior un categorizador de imágenes basado en *Deep Learning*, y cotejamos los resultados obtenidos con el rendimiento de todos los modelos implementados y analizados con cada una de las categorías de imágenes de nuestro banco de pruebas. De esta forma, podemos igualar el máximo rendimiento posible del estado del arte en cada caso.

El categorizador que utilizamos es el que pertenece al servicio de *Computer Vision* de Azure [1]. Este servicio detecta un total de 86 categorías posibles de un nivel de abstracción alto (como se puede observar en la Figura 5.1), lo que en nuestro caso presenta una ventaja fundamental comparado con un segmentador de objetos en imágenes al no depender del contenido concreto de la imagen para elegir una categoría (lo que requeriría analizar miles de etiquetas para cada imagen). En cuanto a la implementación, Azure tiene un convenio gratuito para estudiantes. La comunicación con el servicio se realiza mediante llamadas a una API que permiten como entrada pasar una imagen en formato binario o incluso una URL de cualquier imagen online. Se ha implementado un *script* en Python que permite obtener las categorías de las imágenes para las que se quieren generar *scanpaths*, y que después permite elegir un modelo u otro basándose en las etiquetas que devuelve el servicio de categorización. Como limitación, el servicio gratuito sólo permite 20 llamadas por minuto, lo que podría suponer un cuello de botella en una aplicación completamente desarrollada. Sin embargo, ese ancho de banda ha sido suficiente para realizar

[1] https://docs.microsoft.com/es-es/azure/cognitive-services/computer-vision/concept-categorizing-images

una prueba de concepto de nuestro metamodelo.

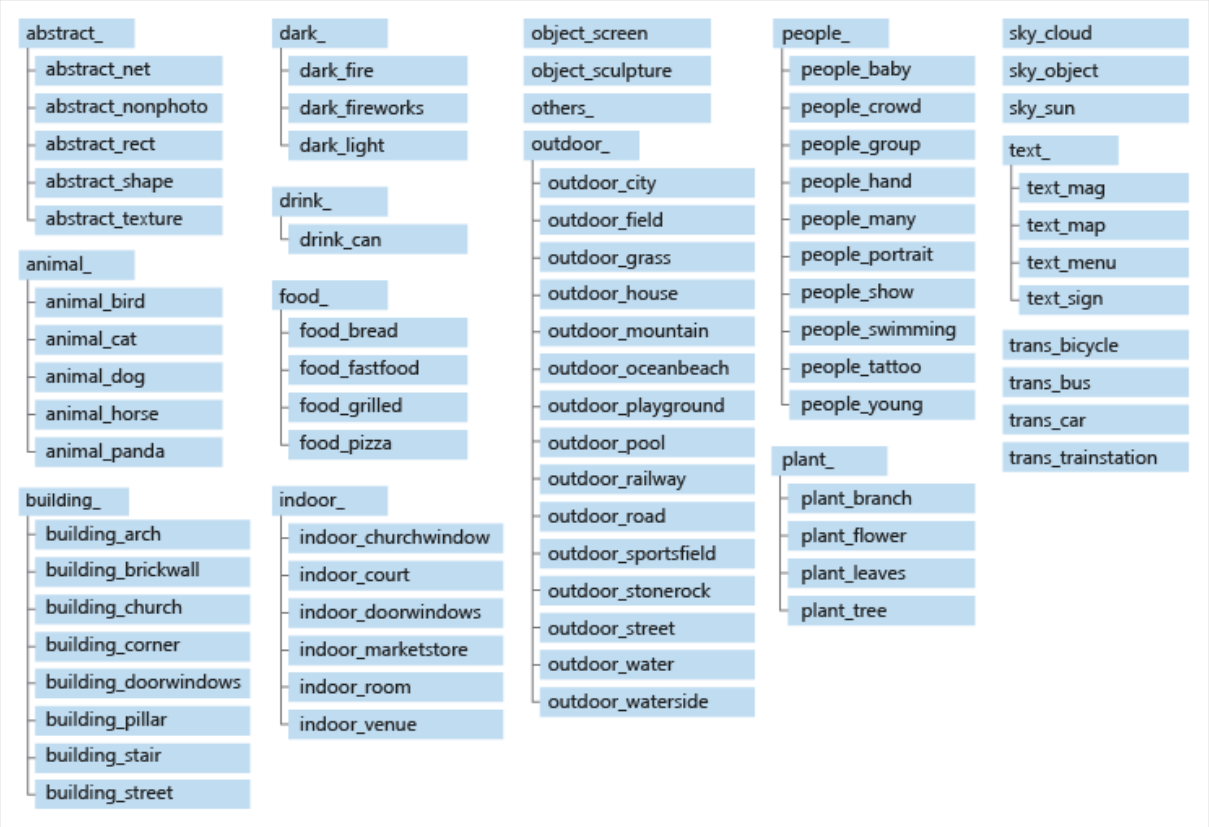

Figura 5.1: Taxonomía de categorías de imagen para el servicio *Computer Vision* de *Azure*.

El categorizador genera varias etiquetas para cada imagen, en orden de probabilidad. Debido a que hasta ahora la etiqueta principal es capaz de distinguir las categorías estudiadas en este trabajo, sólo se utiliza esta. Si en algún momento hiciera falta una separación más fina de categorías se podrían utilizar varias etiquetas de forma simultánea. A modo de ejemplo y para las categorías incluídas en nuestro banco de pruebas, estas son las etiquetas principales que devuelve el categorizador:

- Categoría *faces*: etiqueta *people_portrait*.
- Categoría *illusions*: etiqueta *abstract_nonphoto*.
- Categoría *landscape*: etiqueta *outdoor_mountain*.
- Categoría *nature*: etiqueta *plant_tree*.
- Categoría *pink noise*: etiqueta *abstract_texture*.
- Categoría *urban*: etiqueta *outdoor_road*.
- Categoría *web*: etiqueta *text_menu*.

Con este metamodelo somos capaces de obtener el mejor rendimiento posible de entre los modelos implementados, siempre que hayamos precomputado las métricas para esa categoría de imágenes con anterioridad. Como limitación, en este momento sólo somos capaces de seleccionar modelos dependiendo de la categoría visual de la imagen y no de la tarea visual que se esté llevando a cabo. Para poder incluir la tarea, necesitaríamos añadir esta explícitamente como parámetro de entrada y diseñar por lo menos una métrica adicional que nos permitiera estudiar el comportamiento de los modelos existentes dependiendo del tipo de tarea y no la categoría del estímulo visual. Para ello, necesitaríamos datos de usuarios mirando a la misma imagen realizando diferentes tareas visuales (por ejemplo, viendo una escena de un parque de forma libre, buscando un objeto y recordando cuánta gente aparece en ella, etc.). Es una vía de trabajo futuro muy interesante, pero que se encuentra fuera del alcance de este proyecto.

6. Conclusiones

En este trabajo se han implementado modelos de predicción de *scanpaths* de diferente inspiración, y se han extraído métricas para cada uno de ellos, tanto con bases de datos existentes y ampliamente reconocidas, como con un banco de pruebas diseñado explícitamente para este trabajo, y el cuál tenía el objetivo de someter a los modelos a imágenes y tareas distintas para las que no están entrenados. Como se ha podido observar, los resultados de las métricas son muy dispares, tal y como cabía esperar.

Pese a que la mayoría de los modelos existentes se centran en predicciones de imágenes naturales en tareas de *free-viewing*, las aplicaciones más directas y actuales de los modelos de generación de *scanpaths* se alejan de ese caso en concreto. Hasta ahora se ha dependido de entrenar modelos de nicho para cada paso específico: por ejemplo, a la hora de predecir comportamiento visual en páginas web [20] para poder realizar estudios de usabilidad o mejorar su diseño, estudiando el flujo narrativo y de atención en el arte gráfico [21] (un proceso íntimamente relacionado con la lectura), e incluso en aplicaciones relacionadas con el márketing, el diseño y el emplazamiento de anuncions para maximizar la atención que suscitan [22]. Aunque estamos lejos de obtener una solución definitiva y completa, esperamos que el trabajo propuesto sirva para saber qué modelos existentes se podrían aplicar en determinados casos y qué técnicas se podrían desarrollar para combinar las ventajas de cada uno de estos modelos.

Tras la realización de este trabajo nos ha quedado claro que las implementaciones o código de licencia pública no siempre implican su correcto funcionamiento. Inicialmente, se planteó la implementación de modelos adicionales que en el periódo de estudio del campo atrajo nuestro interés. Sin embargo, debido a un código fuente a medias (con fallos o que sólo funcionaba bajo condiciones muy específicas) y al tiempo disponible para este trabajo, hemos tenido que limitar el número de modelos implementados. Esta experiencia pone en evidencia la necesidad de concienciar a los investigadores y a los desarrolladores en general sobre la importancia de compartir un código que no sólo sea público, si no también fácilmente usable y entendible.

6.1. Trabajo futuro

Como el campo en el que se encuentra este trabajo es relativamente nuevo, existen incontables vías de exploración del mismo y en este trabajo tan solo se ha podido abarcar una pequeña parte de él. Existen a día de hoy varios modelos de predicción de *scanpaths* que emplean *Deep Learning* además del implmentado (Sección 2.4.7). Los resultados de estos modelos son generalmente mejores que para el resto, por lo que se podría seguir explotando todavía más las posibilidades

que ofrece el *Deep Learning*.

Otra vía posible de estudio sería profundizar más en las cuestiones referentes al mal funcionamiento de los modelos, sobre todo en cuanto al tipo de imágenes y tareas. Aunque harían falta más bases de datos para poder realizar más y mejores estudios.

6.2. Valoración personal

La realización de este Trabajo de Fin de Grado ha supuesto una perfecta forma de afianzar muchos de los conocimientos adquiridos durante el mismo y durante los años del grado. Más concretamente en lo referente al área de la inteligencia artifical, el cual siempre me ha generado un gran interés.

Una de las cosas que más me ha gustado, y que a su vez, más difícil me ha resultado de este trabajo, ha sido investigar y aprender por cuenta propia. Aunque contaba con la indiscutible e inmejorable ayuda de la directora, he tenido que afrontar grandes retos, los cuales no se habían presentado durante el grado o no al menos en la misma medida. La tarea de investigar al igual que es fascinante, es frustrante en algunas ocasiones, por motivos como los mencionados a lo largo de esta memoria. Aunque sacándole el lado positivo, al final he aprendido en innumerables aspectos durante el transcurso de este trabajo, no solo académicos, sino también personales.

Bibliografía

[1] Claudio M Privitera. The scanpath theory: its definition and later developments. In Bernice E. Rogowitz, Thrasyvoulos N. Pappas, and Scott J. Daly, editors, *Human Vision and Electronic Imaging XI*, volume 6057, pages 87 – 91. International Society for Optics and Photonics, SPIE, 2006.

[2] Torsten Betz, Tim C Kietzmann, Niklas Wilming, and Peter Koenig. Investigating task-dependent top-down effects on overt visual attention. *Journal of vision*, 10(3):15–15, 2010.

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

[4] Akis Linardos, Matthias Kümmerer, Ori Press, and Matthias Bethge. Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. *CoRR*, abs/2105.12441, 2021.

[5] Matthias Kümmerer and Matthias Bethge. State-of-the-art in human scanpath prediction. *CoRR*, abs/2102.12239, 2021.

[6] Matthias Kümmerer, Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit/tübingen saliency benchmark. https://saliency.tuebingen.ai/.

[7] Dirk Walther and Christof Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395–1407, 2006. Brain and Attention.

[8] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, page 545–552, Cambridge, MA, USA, 2006. MIT Press.

[9] Matthias Kummerer, Thomas S. A. Wallis, Leon A. Gatys, and Matthias Bethge. Understanding low- and high-level contributions to fixation prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[10] Dario Zanca, Stefano Melacci, and Marco Gori. Gravitational laws of focus of attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(12):2983–2995, 2020.

[11] Giuseppe Boccignone and Mario Ferraro. Modelling gaze shift as a constrained random walk. *Physica A: Statistical Mechanics and its Applications*, 331(1):207–218, 2004.

[12] Wanjie Sun, Zhenzhong Chen, and Feng Wu. Visual scanpath prediction using ior-roi recurrent mixture density network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):2101–2118, 2021.

[13] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.

[14] Daniel Martin, Ana Serrano, Alexander W. Bergman, Gordon Wetzstein, and Belen Masia. Scangan360: A generative model of realistic scanpaths for 360° images. *arXiv preprint arXiv:2103.13922*, 2021.

[15] Robert J. Peters, Asha Iyer, Laurent Itti, and Christof Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18):2397–2416, 2005.

[16] Filipe Cristino, Sebastiaan Mathôt, Jan Theeuwes, and Iain Gilchrist. Scanmatch: A novel method for comparing fixation sequences. *Behavior research methods*, 42:692–700, 08 2010.

[17] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2106–2113, 2009.

[18] Niklas Wilming, Selim Onat, José P. Ossandón, Alper Açik, Tim Christian Kietzmann, Kai Kaspar, Ricardo Ramos Gameiro, Alexandra Vormberg, and Peter König. An extensive dataset of eye movements during viewing of complex images. *Scientific Data*, 4, 2017.

[19] Kevin McGuinness Noel E. O'Connor Marc Assens, Xavier Giro-i-Nieto. Pathgan: Visual scanpath prediction with generative adversarial networks. 2018.

[20] Chen Xia and Rong Quan. Predicting saccadic eye movements in free viewing of webpages. *IEEE Access*, 8:15598–15610, 2020.

[21] Khimya Khetarpal and Eakta Jain. A preliminary benchmark of four saliency algorithms on comic art. In *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2016.

[22] Serhat Peker, Gonca Gokce Menekse Dalveren, and Yavuz İnal. The effects of the content elements of online banner ads on visual attention: Evidence from an-eye-tracking study. *Future Internet*, 13(1):18, 2021.