



UNIVERSIDAD
DE GRANADA

Memoria de Trabajo Fin de Grado

Modelos de Atención Visual Multiespectral y RGB basados en Deep Learning

Grado en Óptica y Optometría. Universidad de Granada

Curso 2018-2019

Alumno/a:

Código del TFG

María José Rueda Montes

TFGOO-A-17-18_3

Tutor académico / Tutores Académicos:

Departamento

Eva M. Valero Benito

Óptica

Juan L. Nieves Gómez

Óptica

Granada, a 6 de Junio de 2019



UNIVERSIDAD
DE GRANADA

Declaración de originalidad de los Trabajos Fin de Grado

Grado en Óptica y Optometría. Universidad de Granada

Curso 2018-2019

Alumno/a: *Maria José Rueda Montes*

Título del TFG: *Modelos de Atención Visual Multiespectral y RGB basados en Deep Learning*

Tutor académico: *Eva M. Valero Benito*

Declaro explícitamente que el Trabajo Fin de Grado presentado es original, entendido en el sentido de que no he utilizado fuentes sin citarlas debidamente.

En Granada, a 6 de Junio de 2019

Firmado: María José Rueda Montes

ÍNDICE

| | |
|---|----|
| 1. INTRODUCCIÓN | 1 |
| 2. OBJETIVOS | 2 |
| 3. CONCEPTOS Y DESARROLLO | 2 |
| 3.1. Saliencia..... | 2 |
| 3.1.a. Medida de la saliencia..... | 3 |
| 3.1.b. Predicción y detección de saliencia | 4 |
| 3.1.c. Mapas de saliencia | 4 |
| 3.1.d. Bases de datos más usadas | 5 |
| 3.1.e. Métricas..... | 6 |
| 3.1.f. “Saliency Benchmark” | 10 |
| 3.2. Tipos y evolución de los modelos de saliencia | 10 |
| 3.2.a. Métodos “bottom-up” | 10 |
| 3.2.b. Métodos “top down” | 11 |
| 3.2.c. Arquitecturas con “deep learning” | 11 |
| 3.2.d. Saliencia multiespectral | 11 |
| 3.3. Modelos de saliencia basados en “Deep Learning” | 11 |
| 3.3.a. “DeepLearning”. Redes Neuronales Convolucionales “Convolutional Neural Networks” (CNN)..... | 12 |
| 3.4. Descripción de modelos de saliencia actuales basados en CNN | 16 |
| 3.4.a. Modelos de saliencia RGB..... | 17 |
| 3.4.b. Modelos de saliencia Multiespectral..... | 21 |
| 3.5. Evaluación cuantitativa de los modelos..... | 26 |
| 3.5.a. Modelos de predicción de saliencia | 26 |
| 3.5.b. Modelos de detección de saliencia..... | 28 |
| 4. CONCLUSIONES | 29 |
| 5. AGRADECIMIENTOS | 30 |
| 6. REFERENCIAS | 31 |
| 7. ANEXOS | 33 |

1. INTRODUCCIÓN

En el campo de la visión artificial, una de las tareas de investigación actuales es la predicción de las fijaciones oculares en una escena. Como consecuencia, han surgido una gran variedad de modelos cuya finalidad es localizar las regiones de una imagen en las que existe una mayor predisposición a orientar la mirada. A estos modelos se les denomina *modelos de saliencia* o *modelos de atención visual* y su empleo tiene múltiples aplicaciones; como el reconocimiento de objetos, segmentación de imágenes, evaluación de la calidad de imagen, sustracción de fondo, en anuncios, etc.

A lo largo de la historia, desde el primer modelo de saliencia en 1998, han surgido modelos que se fundamentan en diferentes características de la percepción humana. Sin embargo, en los últimos años han ido adquiriendo relevancia los modelos de saliencia basados en un sistema que imita el funcionamiento neural de la corteza visual V1. Este sistema son las *redes neuronales convolucionales* y está obteniendo resultados pioneros en la actualidad. Con la aparición de las redes neuronales convolucionales, han ido surgiendo investigaciones que proponen nuevos modelos de saliencia o modificaciones a modelos anteriores que incorporan esta estructura. Cada una de estas investigaciones aportan novedades que contribuyen al progreso en este campo. Una de las últimas aportaciones, el uso de imágenes *multiespectrales* o *hiperespectrales*, ha supuesto poder obtener información adicional de la escena, y los modelos que utilizan esta implementación han demostrado un rendimiento superior respecto del uso tradicional de imágenes RGB.

Debido al creciente auge que están teniendo los modelos de saliencia, se ha hecho necesario el uso de diferentes *métricas* de evaluación y la aportación de bases de datos que contengan imágenes de diferentes tipos con información para la saliencia. De esta forma se intenta lograr un resultado objetivo en el rendimiento de cada modelo de saliencia y la posibilidad de comparar los diferentes modelos entre sí. Como consecuencia de la continua aparición de nuevos modelos, se han desarrollado servicios encargados de clasificar e informar sobre los modelos de saliencia más recientes. Estos servicios, los “*saliency benchmark*”, mantienen actualizadas sus páginas webs donde se puede ver la calificación de los modelos con mejores puntuaciones.

En este Trabajo Fin de Grado (TFG), se van abordar los conocimientos básicos y útiles para poder entender la estructura y el funcionamiento de los modelos de saliencia, especialmente los que están basados en “deep learning”, como las redes neuronales convolucionales. También se van a exponer diferentes modelos de saliencia así como sus contribuciones, distinguiendo entre los modelos que utilizan imágenes RGB, y los modelos que usan imágenes multiespectrales o hiperespectrales.

El documento va a seguir la siguiente estructura: comienza con la sección 2, en la que se van a exponer los objetivos de este TFG. En la sección de conceptos y desarrollo 3 -que es la más extensa-, en el apartado sobre saliencia 3.1, se van abordar los conocimientos básicos y útiles para poder entender el objetivo de los modelos basados en este concepto. En el segundo apartado, 3. 2,

se va a desarrollar un pequeño resumen de la aparición en el tiempo de algunos de los modelos de saliencia, dividiéndolos en cuanto al fundamento de sus estructuras. El apartado 3. 3, incluye una descripción general del funcionamiento de las redes neuronales convolucionales. Finalmente, en el último apartado de conceptos y desarrollo, 3. 4, se van a evaluar los modelos expuestos. El documento continua con la sección 4, en la que se van a aportar las conclusiones. La sección 5 y 6, de agradecimientos y referencias respectivamente, y por último la sección 7, en la que aparece el glosario como anexo.

Key words—Saliency prediction, saliency object detection, deep learning-based models, convolutional neural networks, multiespectral, hyperspectral

2. OBJETIVOS

El objetivo de este TFG consiste en revisar las aportaciones de algunos de los modelos de atención visual más actuales basados en “deep learning”, analizando su estructura y funcionamiento. Se pretende dar una idea del fundamento de las redes neuronales convolucionales y su aplicación en los modelos de saliencia, abordando también los modelos que utilizan imágenes multiespectrales, que son de los más recientes y están obteniendo muy buenos resultados.

3. CONCEPTOS Y DESARROLLO

3. 1. Saliencia

El 80% de la información exterior que recibimos proviene del sistema visual. Cada segundo la retina recibe una gran cantidad de información que desde el principio del proceso visual se ha de seleccionar, puesto que la capacidad cerebral es limitada. El método de selección de la información relevante se lleva a cabo mediante la atención, y a los subconjuntos visuales de importancia seleccionados se les denomina objetos salientes (Li y Gao, 2014). Incluso en comparación con los ordenadores más potentes, el procesamiento visual del cerebro humano supera con creces al de un ordenador, tanto en velocidad como en eficiencia. Por lo tanto, el principal objetivo en el desarrollo de la visión artificial es poder imitar al sistema visual humano. De esta forma nace el problema del cálculo de saliencia, cuyo objetivo es localizar y extraer la información visual saliente en una escena, tal y como lo haría el sistema visual humano (Li y Gao, 2014).

La atención es el primer paso a la percepción, y aunque existen muchos tipos de atención, en este ámbito se intenta simular la habilidad de atención visual selectiva, en la que se atiende a un grupo de estímulos concretos a los que se está mirando, ignorando los demás. (Goldstein *et al.*, 2010). Este tipo de atención se encarga de discriminar los aspectos importantes de una escena, mejorando así la velocidad de procesamiento (Mancas *et al.*, 2016).

3. 1. a. Medida de la saliencia

La medida de saliencia puede relacionarse con la medida de la atención, ya que están íntimamente relacionadas (Goldstein *et al.*, 2010). En la medida cuantitativa de la saliencia, se presenta una imagen al sujeto, de la que se genera una representación espacial y temporal detallada de la atención (Mancas *et al.*, 2016).

Los conjuntos de datos obtenidos mediante los procedimientos que se van a nombrar a continuación, se utilizan posteriormente como “ground truth”, es decir, como datos obtenidos de la observación directa, datos fiables para utilizar como referencia en los modelos de saliencia obtenidos de forma artificial.

3. 1. a. 1. Seguimiento Ocular (“Eye Tracking”)

El seguimiento ocular es uno de los métodos más usados para medir la atención. Un mecanismo de atención visual selectiva son los movimientos oculares, orientamos nuestra fóvea a la parte de la escena que nos resulta interesante para así poder verla con más detalle. Además, dirigimos nuestra atención al objeto enfocado. De esta forma, siguiendo el recorrido de los movimientos oculares, se puede obtener un registro de las partes de la escena en las que el observador está prestando atención. Los “Eye Trackers” son los dispositivos encargados de registrar la dirección de la mirada (Mancas *et al.*, 2016).

Aunque a lo largo de la historia se han desarrollado diferentes técnicas de seguimiento ocular, como la electrooculografía o sistemas que incluyen dipolos magnéticos, en la actualidad la técnica más utilizada es la video-oculografía (VOG), que consta de una cámara que registra las reflexiones pupilar y corneal. En este método, una fuente de luz infrarroja ilumina ambos ojos. La luz reflejada por la pupila y la córnea son registradas por una cámara. Después, el software de procesamiento de imágenes incorporado en el dispositivo, utiliza la información recogida sobre la reflexión pupilar para detectar los bordes pupilares, y la información de la reflexión corneal, para descontar los movimientos de la cabeza. De esta forma se obtiene una estimación fiable sobre la orientación de los ojos en el espacio (Duchowski, 2017).

El “Eye Tracker” puede presentarse en diferentes formatos; incorporados directamente a la pantalla en la que se presentan las imágenes, con cámaras independientes, o en gafas (útiles para escenas en exteriores) (Mancas *et al.*, 2016).



*Figura 1. “Eye tracking” inalámbrico en gafa (Mancas *et al.*, 2016).*



*Figura 2. Sistema binocular de “Eye Tracking” independiente para una pantalla (Mancas *et al.*, 2016).*

Finalmente, de los registros de los movimientos oculares se obtiene un mapa de calor (“heat map”) o un mapa de densidad (“density map”). En función de las localizaciones y tiempo de la fijación, el software aplica un filtro gaussiano en las zonas fijadas, cuyo aumento en la amplitud de la curva, que supone una mayor intensidad, implica haber prestado un mayor tiempo de fijación. En este mapa cada pixel indica su capacidad para captar la atención.

3. 1. a. 2. “Mouse Tracking”

Una alternativa de bajo coste para el “Eye Tracker”, es la técnica “Mouse Tracking”. Para realizar este método, se utiliza el registro de la posición del cursor como referencia de la situación de la línea de mirada. Se le pide al participante que desplace el cursor (que imita la resolución foveal) sobre las zonas a las que está mirando. A pesar de que parezca un método poco fiable, puede conseguir resultados parecidos a los del “Eye Tracker”, aunque como es lógico presenta varias desventajas, como por ejemplo, la pérdida de los movimientos oculares inconscientes (Mancas *et al.*, 2016).

3. 1. b. Predicción y detección de saliencia

En este documento se van a tratar dos tipos diferentes de modelos de saliencia: los modelos de *predicción de saliencia* y los de *detección de saliencia*, que aunque en principio aparenten tener la misma finalidad, presentan varias diferencias.

A rasgos generales, los modelos de predicción de saliencia, como su nombre indica, intentan predecir las fijaciones que va a realizar un sujeto tras la visualización libre (sin pedir ninguna tarea) de una imagen, durante 5 o 3 segundos normalmente; mientras que los modelos de detección de saliencia, buscan detectar y segmentar los objetos salientes de una imagen. Realmente, un modelo que obtiene buenos resultados en la predicción de saliencia no tiene porqué dar también buenos resultados en la detección de saliencia, y viceversa. Los modelos de predicción de saliencia deberán de resaltar las zonas de la imagen en las que se han realizado las fijaciones, pudiendo ser dentro o fuera del objeto saliente, mientras que los modelos de detección de saliencia resaltan únicamente los objetos que contienen mayor número de fijaciones. De esta forma, los modelos dedicados a la predicción generan falsos negativos en la detección, y los modelos de detección generan falsos positivos en la predicción. Sin embargo, en la práctica, varias investigaciones han aplicado un umbral a los mapas generados por modelos de predicción de saliencia, para detectar y segmentar los objetos salientes (Borji, 2014).

3. 1. c. Mapas de saliencia

Para la representación de saliencia, surge el concepto de *mapa de saliencia*. Los mapas de saliencia pueden representarse de dos formas diferentes, dependiendo de si el mapa va a estar relacionado con la predicción o con la detección de saliencia. Para la predicción de saliencia, el mapa de saliencia se muestra en una escala de grises en un rango de [0, 255]. En este formato,

cuando un píxel presenta mayor intensidad, se interpreta como que tiene un valor más alto de saliencia. A diferencia de estos mapas, los mapas para la detección de saliencia se representan con una máscara binaria, que asigna el 1 para el píxel que forma parte del objeto saliente, y el 0 para los píxeles de fondo (Mancas *et al.*, 2016). Es también usual encontrar mapas de detección de saliencia con “bounding box”, es decir, cuadros que delimitan el objeto saliente.

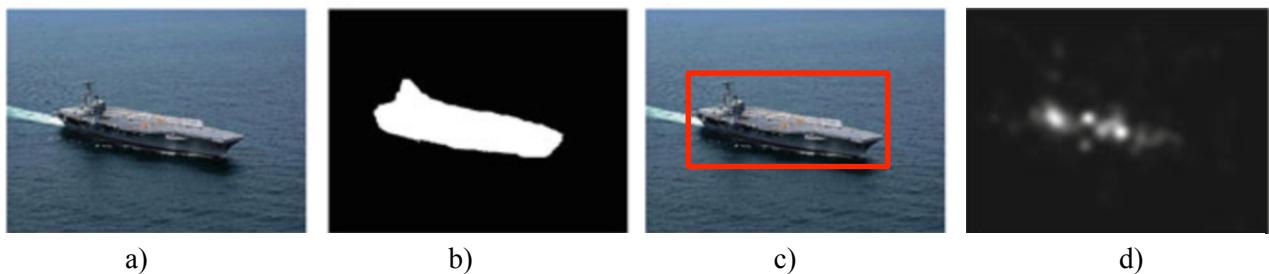


Figura 3. a) Imagen original b) Mapa de saliencia con máscara binaria. c) Mapa de saliencia con “bounding box”. d) Mapa de saliencia continuo en escala de grises (Mancas *et al.*, 2016).

3. 1. d. Bases de datos más utilizadas

Para comparar modelos de saliencia, en el entrenamiento de las redes neuronales convolucionales (aportando “ground truth”), o para evaluar el modelo y comprobar que obtiene resultados fiables (sin aportar “ground truth”), es necesario poseer un conjunto de imágenes con datos de saliencia que sirvan de referencia.

3. 1. d. 1. Conjuntos de datos para la predicción de saliencia

MIT1300 (2009)

Contiene 1.003 imágenes recopiladas de Flickr y LabelMe. “Ground Truth” se obtiene de los datos de “Eye Tracking” obtenidos de la observación libre de las imágenes de 15 observadores. Posee 779 imágenes de paisajes y 228 imágenes de retratos (Judd *et al.*, 2009).

MIT300 (2012)

Tiene 300 imágenes de escenarios interiores y exteriores. En este caso “Ground Truth” se genera de los datos de “Eye Tracking” obtenidos de la observación libre de 39 sujetos. Este conjunto de datos no es público (Judd *et al.*, 2012).

SALICON (2015)

Este conjunto de imágenes es el que recoge más datos para la predicción de saliencia. Ofrece un amplio conjunto de anotaciones destacadas en la base de datos de imágenes “Microsoft Common Objects in Context” (MS COCO), donde “ground truth” se etiquetó basándose en “Mouse

Tracking”. Esta base de datos contiene 10.000 imágenes para el entrenamiento, 5.000 imágenes para la validación y 5.000 imágenes para “testing”. Los mapas de “ground truth” del conjunto de “testing” no están disponibles públicamente y las predicciones deben enviarse al sitio web de SALICON para su evaluación (Jiang *et al.*, 2015).

CAT2000 (2015)

Esta base de datos contiene 4.000 imágenes para la predicción de saliencia, 2.000 para el entrenamiento y 2.000 para “testing”. Tiene 30 categorías de imágenes, incluyendo: acción, afecto, arte, blanco y negro, dibujos animados, fractales, interiores, exteriores, dibujos de líneas, baja resolución, imágenes con ruido, objetos, exterior artificial, exterior natural, estampados, aleatorio, satélite, dibujo y social. Los mapas de saliencia del conjunto de “testing” no están disponibles (Borji y Itti, 2015).

3. 1. d. 2. Conjuntos de datos para la detección de objetos salientes

Con el surgimiento de los modelos de detección de objetos salientes, desde 2007 existe un segundo tipo de “ground truth”. Esta segunda “ground truth” representa los objetos salientes de la imagen, que como ya se ha comentado, pueden mostrarse con dos tipos de máscaras: con “bounding box”, o con una máscara binaria a nivel de píxel que delimita los contornos de los objetos (Mancas *et al.*, 2016).

DUT-OMRON (2014)

Este conjunto de datos es el único que tiene “ground truth” con anotaciones sobre las fijaciones oculares, “bounding box”, y segmentación de los objetos salientes. Contiene 5.168 imágenes con fondos relativamente complejos, que han sido vistas por 5 observadores (Ruan *et al.*, 2014)

HS-SOD (2018)

Está compuesto por 60 imágenes hiperespectrales con su respectiva “ground truth” en binario. Para cada imagen hay 81 canales espectrales que abarcan las longitudes de onda de entre 380 nm y 780 nm con intervalos de 5 nm (Imamoglu *et al.*, 2018).

3. 1. e. Métricas para la validación de modelos de saliencia

Las métricas son medidas estándar utilizadas para evaluar los modelos de saliencia. Proporcionan datos cuantitativos de la comparación entre el mapa de saliencia obtenido por el modelo, y un mapa de saliencia de referencia, el mapa “ground truth”. Por lo tanto, es necesario el uso de “ground truth”, que dependiendo de la métrica, se utiliza el mapa directamente obtenido de “Eye Tracking”, o aplicándole una máscara binaria (Mancas *et al.*, 2016).

La estructura y los resultados de las métricas dependen de varios factores: el tipo de “ground truth”, el parámetro que se desea medir, etc. Como consecuencia, aparece la necesidad de

conseguir una gran variedad de métricas que permitan evaluar los modelos de saliencia de la forma más objetiva posible (Mancas *et al.*, 2016).

3. 1. e. 1. Métrica para la detección de objetos salientes

En este tipo de métricas se utilizan los verdaderos positivos (TP) (puntos que se obtienen en el mapa de saliencia como pertenecientes del objeto, y pertenecen al objeto en “ground truth”); verdaderos negativos (TN) (puntos que se obtienen en el mapa de saliencia como no pertenecientes del objeto, y no pertenecen al objeto en “ground truth”); falsos positivos (FP) (puntos que se obtienen en el mapa de saliencia como pertenecientes del objeto, y no pertenecen al objeto en “ground truth”), y falsos negativos (FN) (puntos que se obtienen en el mapa de saliencia como no pertenecientes del objeto, y sí pertenecen al objeto en “ground truth”). Estos resultados comparan los puntos del mapa de saliencia obtenido por el modelo, con los puntos del mapa binario “ground truth” de referencia. Para ello, a los mapas de saliencia se le aplica una máscara binaria (Mancas *et al.*, 2016).

3. 1. e. 1. 1. “F-Score” por “Precision-Recall” (2009)

Esta métrica calcula la puntuación mediante “Precision-Recall”. “Precision” corresponde al número de verdaderos positivos comparados con el número total de positivos. Y “Recall” es el número de verdaderos positivos comparados con el número total de positivos reales (Mancas *et al.*, 2016).

$$\text{Precision} = \frac{tp}{tp + fp} \quad \text{Recall} = \frac{tp}{tp + fn} \quad F\text{-score} = \frac{(1 + \beta^2) * \text{Precision} * \text{Recall}}{\beta^2 * \text{Precision} + \text{Recall}}$$

Ambas ecuaciones se combinan para calcular “F-Score”. El término β^2 se establece como una constante de 0.3, y se utiliza para dar importancia al valor “Precision”. Por tanto, un valor de “F-score” similar a 1 indicará mejores resultados (Mancas *et al.*, 2016).

3. 1. e. 1. 2. AUC: “Area Under the ROC Curve” (2011)

En este caso se calcula la tasa de verdaderos positivos (TPR), que mide la proporción de verdaderos positivos comparada con todos los positivos reales; y la tasa de falsos positivos (FPR), que calcula la proporción de falsos positivos comparados con todos los datos negativos (Mancas *et al.*, 2016).

$$TPR = \frac{tp}{tp + fn} \quad FPR = \frac{fp}{fp + tn}$$

Los resultados de ambas ecuaciones se combinan en una curva ROC. Después, para obtener el resultado, se mide el área debajo de la curva (Mancas *et al.*, 2016).

3. 1. e. 2. Métricas para la predicción de saliencia

Las métricas dirigidas a la predicción de saliencia pueden dividirse en tres tipos, dependiendo del factor que utilizan para comparar el mapa de saliencia y “ground truth”: métricas basadas en el valor, que se centran en los valores que se obtienen en los puntos de saliencia; métricas basadas en la distribución, que se fundamentan en la distribución estadística de los puntos de saliencia; y métricas basadas en la localización, que se basan en la ubicación de los puntos salientes (Mancas *et al.*, 2016).

Estas métricas cogen como entrada el mapa de predicción de saliencia obtenido por el modelo (SM) y el mapa “ground truth” (FM). En estos casos “ground truth” se representa en un mapa discreto para las métricas basadas en valor y localización, y continuo para las métricas basadas en distribución (Mancas *et al.*, 2016).

3. 1. e. 2. 1. Métricas basadas en el valor

En este tipo de métricas se comparan los valores o amplitudes de los mapas de saliencia con los mapas de “Eye Tracking” que corresponden a “ground truth”.

3. 1. e. 2. 1. 1. NSS: Normalized Scanpath Saliency (2005)

Con este método se pretende cuantificar los valores de los puntos salientes y normalizar el resultado con la varianza obtenida del mapa de saliencia (Mancas *et al.*, 2016).

$$NSS(p) = \frac{SM(p) - \mu_{SM}}{\sigma_{SM}}$$

(p) corresponde a la ubicación del punto saliente, μ_{SM} es el mapa de saliencia normalizado a una media de 0, y σ_{SM} es la desviación estándar del mapa. De esta forma, si la varianza es pequeña, o la diferencia entre los valores de los puntos salientes y la media es alta, se verifica que el mapa de saliencia obtenido es adecuado para la predicción. Un valor de NSS(p) alto será más favorable (Mancas *et al.*, 2016).

El valor final de NSS se calcula como el promedio de NSS(p) para todos los puntos de saliencia de “ground truth” (Mancas *et al.*, 2016).

3. 1. e. 2. 2. Métricas basadas en la distribución

Existen dos tipos de métricas que están basadas en la distribución: las métricas que calculan las similitudes entre dos distribuciones, y las que calculan la diferencia. Como ejemplo se va a comentar una de las métricas que utiliza la diferencia (Mancas *et al.*, 2016).

3. 1. e. 2. 2. 1. EMD: Earth Mover's Distance (2012)

Esta métrica se encarga de medir la distancia entre dos distribuciones de probabilidad sobre una región. Calcula el mínimo coste necesario para transformar la distribución del mapa de saliencia del modelo, en la distribución de saliencia de “ground truth” (Mancas *et al.*, 2016).

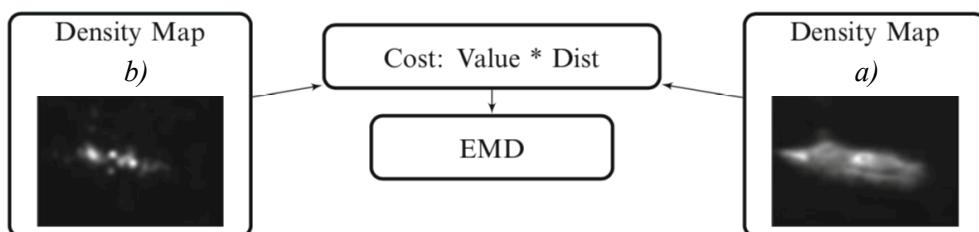


Figura 4. Operación que realiza la métrica EMD. El mapa a) corresponde al mapa de saliencia obtenido por el modelo y el mapa b) “ground truth” (Mancas *et al.*, 2016).

Un valor 0 o cercano a 0, indicará una diferencia nula o casi inapreciable entre las distribuciones del mapa de saliencia y el mapa de “ground truth”. Por tanto, cuanto más se acerque a 0 el valor de EMD, mejor será el resultado del mapa del modelo (Mancas *et al.*, 2016).

3. 1. e. 2. 3. Métricas basadas en la localización

Estas métricas son muy usadas para evaluar los mapas de predicción de saliencia. Se basan en la métrica “Area Under the ROC Curve” y existen cuatro tipos diferentes (Mancas *et al.*, 2016). A continuación se va a poner un ejemplo.

3. 1. e. 2. 3. 1. AUC-Judd: Hit Rate for Area Under the ROC Curve (2012)

Para poder validar los modelos de saliencia con la métrica AUC, Judd propuso una versión actualizada del método anterior. En esta métrica el primer paso es contar los píxeles salientes del mapa “ground truth”, y después extraer de forma aleatoria el mismo número de píxeles del mapa de saliencia. A los píxeles obtenidos del mapa de saliencia se le aplica un umbral, que asignará como puntos salientes a los valores encima del umbral, y como puntos de fondo, los valores por debajo del umbral. La operación de clasificación de ambos grupos de píxeles se repite unas cien veces (Mancas *et al.*, 2016).

De esta forma se obtienen dos grupos de píxeles, unos que coinciden con los píxeles de saliencia del mapa de “ground truth” (verdaderos positivos), y otro grupo de píxeles que se han clasificado como puntos salientes pero no lo son realmente (falsos positivos). Con estos dos valores se forma la curva ROC y se calcula el área debajo de la curva. Un valor aproximado a 1 indicará un buen resultado, mientras que 0,5 indica aleatoriedad (Mancas *et al.*, 2016).

3. 1. f. Saliency Benchmark

Los modelos de saliencia han ido cogiendo relevancia durante los últimos 10 años. En la actualidad existen una gran cantidad de modelos diferentes cuyos algoritmos están disponibles online. Con el fin de poder comparar los modelos entre sí de una forma objetiva, se crean los “saliency benchmark”. Los “saliency benchmark” son servicios encargados de clasificar e informar sobre los modelos de saliencia más recientes. Además, mantienen actualizadas las bases de datos disponibles para los modelos. Cada modelo se evalúa calculando el rendimiento para predecir la saliencia utilizando múltiples métricas y diferente bases de datos. Para evitar el ajuste de los modelos a los conjuntos de datos, las bases de datos que se utilizan para calificar los modelos no están publicadas (Li y Gao, 2014).

Por ejemplo, “MIT Saliency Benchmarck” publica los resultados de calificación en su página web, a la que cada persona puede mandar su modelo para ser evaluado. Los dos conjuntos de datos de referencia que proporcionan son los de MIT300 y CAT2000 (Mit saliency benchmark, 2019).

3. 2. Tipos y evolución de los modelos de saliencia

El sistema visual está compuesto por dos mecanismos de atención, “bottom-up” y “top-down”. Los factores “bottom-up” dependen únicamente de la imagen (como el color o la textura), y son los responsables de que la atención se dirija de forma automática a las zonas importantes de la imagen. Sin embargo, en el mecanismo “top-down” se fundamenta en el conocimiento previo, el contexto y la tarea a realizar, por lo que depende de la situación y es subjetivo. Utiliza el conocimiento para orientar la atención a las localizaciones relevantes de la imagen (Kruthiventi *et al.*, 2015).

Desde los primeros modelos de saliencia, basados en los mecanismos de atención “bottom-up”, han surgido una gran variedad de modelos hasta llegar a los más actuales que utilizan “deep learning” en su estructura. A continuación se van a citar algunos modelos de cada tipo, a modo de resumen.

3. 2. a. Métodos “Bottom-Up”

Los métodos “bottom-up”, se basan en la saliencia de características de la imagen de bajo nivel, tal como el color, contraste, orientación, textura, etc. En 1998 Itti *et al.*, proponen el primer método de saliencia. En este modelo se extraen las características de la imagen de intensidad, color y orientación, teniendo en cuenta las diferencias centro-periferia, y todas ellas se integran en etapas posteriores para generar el mapa final de saliencia (Itti *et al.*, 1998). En años posteriores surgieron varios modelos basados en el mismo método, sin embargo, todos estos modelos obtienen malos resultados con imágenes de fondos complejos, ya que están basados en el contraste local. En 2006, Zhai y Shah proponen un método basado en el contraste global, en el que calcularon la

saliencia a nivel de cada píxel en todos los píxeles, utilizando los histogramas de color, además establecieron un modelo de atención jerárquico (Zhai y Shah, 2006).

En general, las características de bajo nivel de la imagen pueden ser efectivas para escenas simples pero, al centrarse en las partes informativas de la imagen, generan falsos positivos para imágenes más complejas (Wang *et al.*, 2017).

3. 2. b. Métodos “Top-Down”

Los métodos “top-down” tienen en cuenta el contexto de las imágenes y consideran información de alto nivel (por ejemplo las caras) (Wang *et al.*, 2017). Oliva *et al.* en 2003 incorporan la información de contexto teniendo en cuenta estudios estadísticos de observaciones humanas para determinar la saliencia (Oliva *et al.* 2003). En 2009, Judd *et al.* entrenaron una máquina de vectores de soporte lineal a partir de los datos de “Eye tracking” para localizar las zonas salientes de la imagen, utilizando características de bajo, medio y alto nivel (Judd *et al.*, 2009).

3. 2. c. Arquitecturas con “Deep Learning”

El reciente progreso de los modelos de saliencia se debe a los avances en “deep learning”. Las redes neuronales convolucionales examinan las características de alto nivel en más categorías de objetos, mostrando un excelente rendimiento. El primer modelo para predicción de saliencia con “deep networks” fue el modelo eDN en 2014, en el que propusieron una combinación de mapas de características en tres capas convolucionales diferentes, que finalmente combinaron con un clasificador lineal entrenado con positivo (saliente) o negativo (no saliente) (Vig *et al.*, 2014).

3. 2. d. Saliencia Multiespectral

Todos los trabajos que se han nombrado en los apartados anteriores, y la mayoría de los modelos tanto anteriores como actuales, utilizan imágenes RGB como entrada. Existen pocos trabajos que investigan con la información de varios espectros. En 2013 Wang *et al.* incorporaron a su trabajo imágenes en infrarrojo cercano, junto con las imágenes RGB, para la detección de saliencia, de donde se obtuvieron unos mapas de saliencia precisos (Wang *et al.*, 2013). Todos los métodos que se han utilizado con imágenes de varios espectros han demostrado un rendimiento prometedor, lo que inspira a realizar investigaciones en esta dirección.

3. 3. Modelos de saliencia basados en “Deep Learning”

En la última década las redes neuronales convolucionales han conseguido obtener resultados pioneros en una gran variedad de campos, especialmente en aplicaciones que incluyen datos de imágenes: como en la clasificación de imágenes, visión artificial, procesamiento de imágenes, etc (Albawi *et al.*, 2017). Además, también existe una gran cantidad de compañías que utilizan “deep learning”: Facebook utiliza redes neuronales para sus algoritmos de etiquetado automático;

Google en la búsqueda de imágenes; Instagram para su formato de búsqueda; incluso Amazon para recomendar sus productos (Deshpande, 2017).

El término “Deep Learning” o “Deep Neuronal Network”, hace referencia a una red neuronal artificial (“Artificial Neural Networks” (ANN)) con multi-capas. Una de las redes más populares de “deep learning” son las redes neuronales convolucionales o “Convolutional Neural Network” (CNN). Como su nombre indica, el mecanismo de las redes neuronales convolucionales está basado en el funcionamiento neural de la corteza visual (V1). La multi-capa de la CNN incluye una capa convolucional (“convolutional layer”), capa de no linealidad (“non-linearity layer”), capa de agrupación (“pooling layer”), y la capa completamente conectada (“fully-connected layer”). Conforme la imagen se propaga por las capas más profundas de la red, se van extrayendo características más complejas, comenzando por las características de bajo nivel, como los bordes, y acabando por características complejas, como por ejemplo una cara (Albawi *et al.*, 2017).

3. 3. a. “DeepLearning”. Redes Neuronales Convolucionales, “Convolutional Neural Networks” (CNN)

En este apartado se va a tratar el funcionamiento de las diferentes capas que forman la red neuronal convolucional, así como el motivo de incorporación de cada capa. En general también se va a plantear el funcionamiento y mecanismo detrás de la eficiencia de este tipo de red neuronal artificial.

3. 3. a. 1. Capa de convolución

I. Convolución

En la capa de convolución, una matriz de números de un tamaño determinado, se desplaza a través de la imagen con el fin de obtener información. Esta matriz se llama filtro, máscara, ventana o núcleo, y puede tener diferentes valores dependiendo de la característica que se pretenda extraer de la imagen (Deshpande, 2017). Conforme el filtro se va desplazando sobre una región determinada de la imagen (campo receptivo), los valores del filtro (pesos o parámetros) y los valores de la región del campo receptivo (valor original del pixel, en la primera convolución) se multiplican, y se obtiene una matriz de números que se suman para dar el valor de salida que servirá como entrada a la neurona correspondiente de la siguiente capa. El conjunto de valores obtenidos en el paso del filtro se denomina mapa de activación o mapa de características (Deshpande, 2017). Por cada región de valores se obtiene un número, mejorando la eficiencia de la red (Albawi *et al.*, 2017).

| | | | | |
|---|---|-----------------|-----------------|-----------------|
| 1 | 1 | 1 | 0 | 0 |
| 0 | 1 | 1 | 1 | 0 |
| 0 | 0 | 1 _{x1} | 1 _{x0} | 1 _{x1} |
| 0 | 0 | 1 _{x0} | 1 _{x1} | 0 _{x0} |
| 0 | 1 | 1 _{x1} | 0 _{x0} | 0 _{x1} |

a)

| | | |
|---|---|---|
| 4 | 3 | 4 |
| 2 | 4 | 3 |
| 2 | 3 | 4 |

b)

Figura 5. Operación de convolución. a) corresponde con los datos de entrada, b) es el mapa de activación. La matriz de números que opera sobre la imagen es el filtro (UFLDL Tutorial, 2019).

Para aprovechar este método, y aumentar la información obtenida de los datos de entrada, es posible agregar más capas tras la capa de entrada. Cada capa puede estar asociada a diferentes filtros, para así poder extraer diferentes características de la imagen (Albawi *et al.*, 2017).

En niveles más profundos de las capas, podemos poner como ejemplo un filtro de detección de curvas. Cuando se convoluciona el filtro a través de la imagen, se obtendrán valores altos justo cuando el filtro se superponga en una región que contenga valores similares, curvas en este caso (Deshpande, 2017). Sin embargo, los valores serán muy pequeños o incluso nulos cuando el filtro se superpone en una zona de la imagen diferente.

Cuando se pasa por la primera capa de convolución, el mapa de características generado sirve como entrada para la segunda capa de convolución, y así sucesivamente. Por lo tanto, y ya que las neuronas de la siguiente capa solo cogen la información de su campo receptivo, cada capa describe las ubicaciones de la imagen en la que se encuentran las características (Deshpande, 2017).

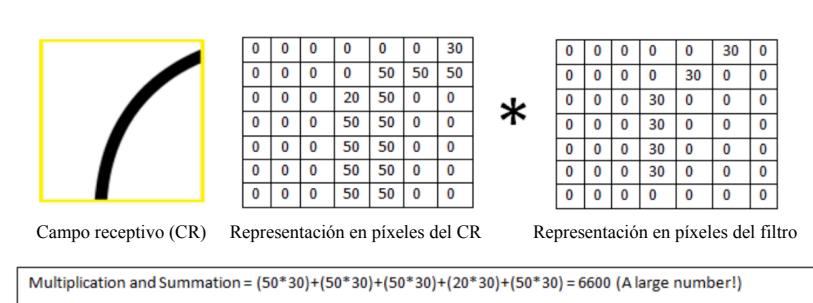


Figura 6. Resultado de la operación de convolución tras aplicar un filtro de detección de curvas que coincide sobre una curva en la imagen (Deshpande, 2017).

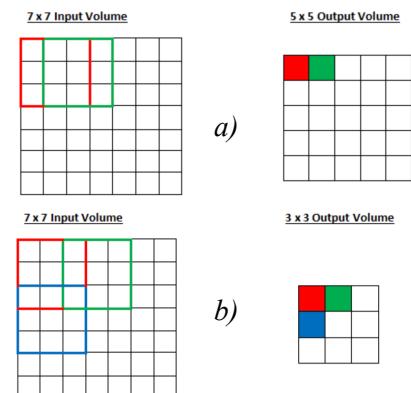


Figura 7. a) ocurre cuando “stride” es de valor 1, b) cuando “stride” es de valor 2 (Deshpande, 2017).

II. Paso (“Stride”)

El paso es el número de unidades que se desplaza el filtro sobre la matriz de entrada. Por ejemplo, cuando el paso es de 1 unidad, el filtro se desplaza de un pixel en un pixel. Ajustando el paso se pueden cambiar las dimensiones de la matriz de salida, es decir, del mapa de características. Además, con el paso también se puede modificar la superposición del campo receptivo. Conforme se aumenta el paso, se disminuye el tamaño mapa generado y la superposición de los campos receptivos (Albawi *et al.*, 2017).

III. Relleno 0 (“Padding”)

Consiste en rodear la matriz de entrada de valores 0. Se utiliza para evitar la pérdida de información en los bordes de la imagen de entrada, y poder aplicar el filtro a los mismos. También se obtiene como ventaja el control de las dimensiones de los mapas de características generados (Deshpande, 2017).

Hay que tener en cuenta que la elección de los parámetros es arbitraria, se pueden modificar dependiendo del tipo de red que se desea obtener. Los parámetros se modifican hasta encontrar la combinación que genere la extracción de información de la imagen en una escala adecuada (Albawi *et al.*, 2017).

3. 3. a. 2. Capa de no linealidad (“Nonlinearity layer”)

Después de cada capa de convolución se utiliza la operación ReLU (“Rectified Linear Units”), también llamada capa de activación (Albawi *et al.*, 2017). Debido a que las operaciones que se realizan en la red de convolución, multiplicación y suma, son operaciones lineales, la función ReLU permite obtener tras cada capa de convolución un mapa de características de salida no lineal. Esta capa aplica la función $f(x) = \max(0,x)$ a todos los valores del volumen de entrada, cambiando así todos los valores negativos al valor 0 (Deshpande, 2017).

3. 3. a. 3. Capa de agrupación (“Pooling layer”)

El objetivo de esta capa es simplificar los mapas de características que se generan, para así reducir la complejidad de tener capas adicionales. Para ello, uno los métodos más comunes de agrupación ,“MaxPooling”, aplica un filtro al volumen de entrada, del que se obtiene como salida el mayor valor de cada subregión (Deshpande, 2017). Normalmente el filtro es de tamaño 2x2 y se desplaza con un paso de tamaño 2 (Albawi *et al.*, 2017).

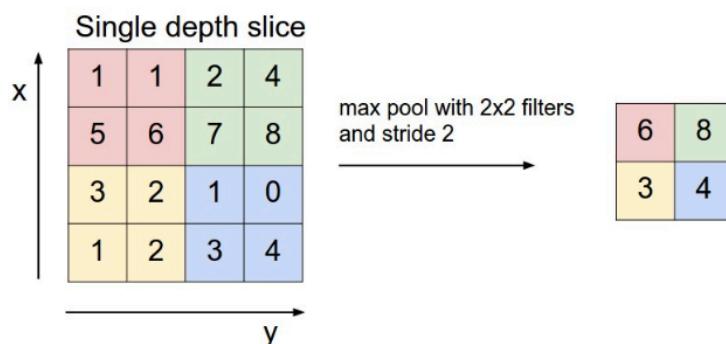


Figura 8. “Pooling” con un filtro de tamaño 2 y “stride” 2. De una entrada de tamaño 4x4 se obtiene una salida de 2x2, conservándose el máximo valor local (Albawi *et al.*, 2017).

3. 3. a. 4. Capa completamente conectada (“Fully connected layer”)

Esta capa se encarga de comprobar la relación entre las características de alto nivel generadas por la capa anterior, y los valores de una clase concreta, para que al calcular los productos entre los pesos y la capa anterior se obtenga la probabilidad de que los valores de la capa anterior pertenezcan a esa clase (Deshpande, 2017).

La capa completamente conectada, en algunos modelos de predicción de saliencia, se sustituye por capas convolucionales (capa de convolución), consideradas como un decodificador. La capa de deconvolución sirve para producir un mapa de saliencia con las mismas dimensiones que la

imagen de entrada. De esta forma al superponer el mapa de saliencia y la imagen de entrada, se podrán ver las zonas de escena con mayor preferencia de fijación.

3. 3. a. 5. Entrenamiento (“Training”)

Antes de que la red del modelo pueda obtener algún resultado fiable, es necesario entrenarla. Para el entrenamiento de los modelos de saliencia se utiliza una base de datos de imágenes con “ground truth” obtenida mediante los mapas de “Eye Tracking”. La forma mediante la que se ajustan los valores tras el entrenamiento se denomina “backpropagation”. Este proceso consta de 4 etapas, el paso hacia delante (“forward pass”), la función de pérdida (“loss function”), paso hacia atrás (“backward pass”) y la actualización de pesos (“weight update”).

I. “Forward Pass”: en esta etapa, la imagen de entrenamiento pasa a través de la red. Se obtiene como salida un valor sin preferencia, aleatorio, ya que los pesos de los filtros se inician de forma aleatoria.

II. “Loss Function”: la más común es el error cuadrático medio (MSE), que es $1/2$ (valor real - valor predicho) al cuadrado (aunque existen diferentes formas como “cross-entropy loss”).

$$E_{total} = \sum \frac{1}{2} (target - output)^2$$

Como se ha mostrado en el paso anterior, el resultado de “loss function” será muy grande para las primeras imágenes de entrenamiento. Para conseguir que el mapa de saliencia de la red coincida con el mapa de “ground truth” se deberá de minimizar el valor obtenido en “loss function” (Deshpande, 2017).

III. “Backward Pass”: se realiza un paso hacia atrás en la red, para determinar qué pesos de cada filtro han contribuido más en el aumento de “loss function” y encontrar la forma de ajustarlos para disminuir el resultado.

IV. “Weight Update”: en esta etapa se actualizan todos los pesos de los filtros para que cambien en la dirección adecuada (Deshpande, 2017).

El proceso de entrenamiento se desarrolla un número fijo de repeticiones para cada subconjunto de imágenes de entrenamiento (lote o batch), hasta que los pesos de las capas se sintonizan correctamente. Conforme mayores sean los datos de entrenamiento, y por tanto más veces se entrene la red, se obtendrá un número mayor de actualizaciones de pesos, que como consecuencia generan unos resultados más semejantes a los resultados obtenidos en observaciones humanas (Deshpande, 2017).

3. 3. a. 6. Transferencia de aprendizaje (“Transfer Learning”)

Es necesario tener una gran cantidad de datos para poder obtener unos resultados adecuados en una red. La transferencia de aprendizaje permite disminuir la demanda de datos necesarios. Este proceso se lleva a cabo escogiendo una red previamente entrenada con una gran cantidad de imágenes, y ajustando el modelo con el conjunto propio de datos. Para ajustar la red se elimina la última capa para sustituirla por capas propias del nuevo modelo, y después se entrena de forma normal habiendo congelado los pesos de las otras capas. En general, el modelo preentrenado se utiliza para extraer las características de las imágenes (Deshpande, 2017).

3. 3. a. 7. Prueba (“Testing”)

Tras el entrenamiento de la red, se debe de comprobar si esta es capaz de generar buenos resultados. Para ello, se escoge un conjunto diferente de imágenes y se pasan a través de la red neuronal convolucional, después se compara la salida obtenida con “ground truth” (Deshpande, 2017).

3. 4. Descripción de modelos de saliencia actuales basados en CNN

Con el desarrollo de las redes neuronales convolucionales, se han podido obtener representaciones más fieles de las características visuales. Desde la primera aplicación de las CNN en 2014 con el trabajo propuesto por Vig *et al.*, la predicción de saliencia ha mejorado notablemente (Jia y Bruce, 2019).

Aunque todos los modelos de CNN utilizados en la predicción de saliencia son poco profundos (menor precisión de predicción o detección de saliencia) en comparación con los modelos más modernos de reconocimiento de objetos, la posibilidad de usar las características aprendidas en los modelos de reconocimiento de objetos, para la predicción de mapas de saliencia, facilita la obtención de modelos de predicción de saliencia efectivos. Además, la reciente aparición de conjuntos de datos a gran escala, específicos para saliencia, ha supuesto una gran mejora de los resultados (Jia y Bruce, 2019). Otras implementaciones novedosas, como el uso de imágenes multiespectrales, que sustituyen a las imágenes RGB de entrada, también están consiguiendo modelos de saliencia que producen mapas más parecidos a “ground truth”.

En la actualidad existen una gran cantidad de modelos de predicción de saliencia basados en CNN, cuya posición en los “saliency benchmark” está en continuo cambio debido a la aparición de nuevos modelos. A continuación se van a describir algunos de los modelos más actuales que utilizan “deep learning”, ya que este tipo de modelos están en auge, y la mayoría de investigaciones dedicadas a saliencia están dirigidas en torno a este tipo de arquitectura.

3. 4. a. Modelos de saliencia RGB

Como ya se ha comentado, este tipo de modelos cogen como entrada una imagen a color con tres canales, RGB. Los tres canales codifican el color de la imagen; el canal rojo (R) comprende una longitud de onda de 700 nm, el canal azul (B) con una longitud de onda de 546.1 nm, y el canal verde (G), 435.8 nm (Schanda, 2007).

3. 4. a. 1. Modelo de detección de saliencia “Stagewise Refinement” (Wang et al., 2017)

I. Introducción

Las múltiples capas de agrupación y convolución de las CNN reducen la resolución de la imagen de entrada, perdiendo parte de la estructura de la imagen. Aunque esta técnica presenta buenos resultados para la clasificación, para la segmentación de objetos salientes, en los modelos de detección de saliencia, resulta problemática.

En este modelo se muestra un nuevo algoritmo que proporciona un mecanismo de refinamiento por etapas, en el que las estructuras con mayor resolución son renovadas gradualmente mediante varias redes de refinamiento.

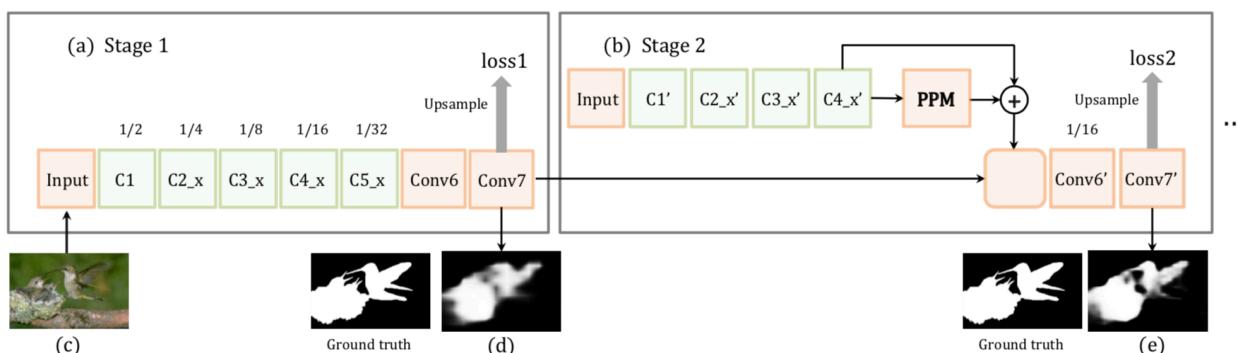


Figura 9. Estructura del modelo “Stagewise Refinement”. c) imagen de entrada. d) es el mapa generado en la etapa 1, S1, e) mapa generado en la etapa 2, S2 (Wang et al., 2017).

II. Modelo propuesto

Este modelo consta de dos etapas: una etapa inicial en la que se genera un mapa burdo de saliencia, y una segunda etapa de refinamiento en la que se incluye un módulo de agrupación piramidal (PPM). En la primera etapa se utilizan las 5 primeros bloques de convolución de la red de clasificación ResNet-50, añadiendo dos capas convolucionales. Esta red coge una imagen completa como entrada, de la que se obtiene un mapa de saliencia con la misma resolución, aumentando la resolución directamente del último mapa de características. Se genera así el primer mapa de saliencia S1.

La etapa de refinamiento F2 (Figura 9. “Stage” 2), cuya finalidad es recuperar información local perdida durante el proceso anterior, coge como entrada la imagen inicial, y utiliza los 4 primeros bloques de ResNet-50 con parámetros diferentes a los de la primera etapa. Finalmente se obtiene

un mapa de saliencia S2 más fino, al fusionar el mapa obtenido en esta etapa de refinamiento y el mapa de la etapa anterior, S1. De la misma forma, en una etapa de refinamiento posterior se utiliza el mapa S2 como entrada para fusionarlo con el nuevo mapa F3 obtenido, y generar el mapa S3, así sucesivamente.

$$\mathbf{S}^t = \mathbf{R}^{t-1}(\mathbf{S}^{t-1}, \mathbf{F}^t),$$

El módulo de agrupación piramidal se añade en cada etapa de refinamiento. Este módulo está formado por varias capas de agrupación con tamaños de filtro diferentes, 1x1, 2x2, 3x3 y 6x6. A los mapas de características generados en cada escala se les reduce la dimensión y se fusionan entre sí para obtener el mapa de salida. Al final de esta etapa, el mapa de salida del PPM se *concatena** con el mapa de características de salida del último bloque convolucional de refinamiento.

III. Bases de datos

El entrenamiento se inicia con los pesos de ResNet-50 y se ajusta con el conjunto de datos DUT-S. Cada etapa del algoritmo es entrenada para producir repetidamente un mapa de saliencia basado en el anterior, pero con detalles más finos recuperados y agregados.

Para evaluar el modelo se utilizan 5 bases de datos de “benchmarks” populares: ECSSD, THUR15K, DUT-OMRON, SED, HKU-IS, y DUT- S.

IV. Resultados

Como resultado se obtiene el incremento del rendimiento al añadir etapas de refinamiento t, esto se debe a que en la predicción de las siguientes etapas se utilizan la información de contexto proporcionada por los mapas anteriores de la red auxiliar. Además, el método de refinamiento de múltiples etapas consigue combinar efectivamente *semánticas del objeto** de alto nivel con características de imagen de bajo nivel para finalmente obtener un mapa de saliencia de alta resolución.

El módulo de agrupación piramidal también produce un aumento del rendimiento ya que añade información de contexto global, importante para distinguir los objetos con saliencia del fondo.

3. 4. a. 2. SAM-VGG, SAM-ResNet: “Saliency Attentive Model” basado en LSTM para la predicción de saliencia (Cornia et al., 2018)

I. Introducción

“Machine attention” es un paradigma de programación que se encarga de atender secuencialmente a diferentes partes de la entrada. Uno de los ejemplos en los que este paradigma obtiene buenos

resultados es aplicado en subtítulos y traducción automática, para enfocarse selectivamente en diferentes fragmentos de la oración.

Basándose en el concepto de “Machine attention”, en este trabajo se propone un nuevo modelo, “Saliency Attentive Model”, que incorpora una red “Attentive Convolutional Long Short-Term Memory” (Attentive ConvLSTM) que se centra en las zonas relevantes de la imagen para refinar las características de saliencia.

II. Modelo

El modelo consta de una CNN conocida, que es modificada, y de dos módulos posteriores, “Attentive Convolutional LTSM” y “Learned Priors”. La principal novedad en este modelo es la red “Attentive Convolutional LTSM”. Esta red procesa de forma iterativa las características de saliencia en diferentes localizaciones, al atender selectivamente a las diferentes regiones de la imagen. El modulo “Attentive ConvLTSM” coge como entrada las características extraídas de la imagen, y produce un conjunto refinado de mapas de características que pasan al módulo de “Learned Priors”.

En el módulo “Learned Priors”, la red aprende la predisposición que existe a orientar las fijaciones oculares hacia el centro de la imagen, sin necesidad de tener que incorporar manualmente esta información. Para ello se concatenan diferentes filtros gaussianos con los mapas de características obtenidos en el módulo “Attentive ConvLTSM”. Después el mapa generado de la fusión de los dos mapas anteriores se pasa a través de una capa convolucional que quita linealidad. Todo este módulo se repite dos veces.

Como estrategia para aumentar la resolución de la CNN inicial de extracción de características, se utilizan las redes ResNet-50 y VGG-16, que se modifican reduciendo el tamaño de “stride” y añadiendo “padding”.

III. Bases de datos

Para entrenar el modelo utilizan los conjuntos de datos de SALICON y MIT1003, y para el “testing” MIT300 y CAT2000.

IV. Resultados

Se obtiene una mejora constante de los modelos de predicción de saliencia, ya que al agregar los módulos de “Attentive Convolutional LTSM” y “Learned Priors” se genera un mapa más similar a “ground truth”. Los resultados muestran que los nuevos modelos, SAM-ResNet y SAM-VGG, obtienen mejor resultado en los mapas de saliencia, y cada componente contribuye al aumento del rendimiento.

Cualitativamente se obtienen buenos resultados en imágenes que no presentan regiones claramente salientes, como en paisajes. Además, el modelo es capaz de deducir la importancia relativa que poseen diferentes personas en una misma escena. Esta última función es un comportamiento humano que resulta difícil de imitar para los modelos de saliencia.

3. 4. a. 3. EML-NET: “Expandable Multi-Layer NETwork” para la predicción de saliencia (Jia y Bruce, 2019)

I. Introducción

Los modelos de predicción de saliencia necesitan utilizar bastante memoria para la representación de los datos, ya que se utilizan imágenes de un mayor tamaño que para los modelos de clasificación. Además, la incorporación de características de alto nivel como los rostros, y la combinación de varios modelos de CNN entrenados, amplían aún más la necesidad de memoria en el sistema.

Con fin de solucionar los problemas comentados, se propone un *sistema escalable** para la predicción de saliencia. En este modelo utilizan como método la fracción del sistema completo en módulos de entrenamiento, en el que cada módulo se entrena por separado.

II. Modelo

Para estudiar la saliencia se extraen las características utilizando diferentes modelos de CNN, e intentan utilizar el conocimiento previo aprendido durante la visualización de imágenes. Los modelos utilizados son de mucha profundidad, como DenseNet-166, con 160 capas convolucionales, y NasNet-Large de 568 capas convolucionales.

La división para el entrenamiento se realiza en torno a la red que genera el mapa de características (codificador), y la red de la que se obtiene la predicción de saliencia (decodificador). En la primera etapa, la etapa de codificado, los modelos usados de CNN son modelos preentrenados en clasificación de imágenes, en los que la capa “fully connected” (normalmente capas con muchos filtros en las que se requiere de bastante espacio) se sustituye por capas convolucionales que actúan de decodificador. Con la misma finalidad, la de ahorrar espacio, se comprime la predicción de salida con una capa de convolución 1@1 (tamaño de kernel @ número de mapas de filtro), y no se extraen las características de alto nivel.

En la etapa de decodificado, se entrena un decodificador para combinar las características aprendidas en los dos modelos de CNN. Se seleccionan las capas útiles y se comprimen en un mapa de características (mediante Conv1@1), individualmente, para posteriormente alinearlos y concatenarlos. Cuando se entrena el decodificador los pesos de los modelos CNN se congelan para así reducir el tamaño de los mismos. Finalmente, el resultado se redimensiona al tamaño del mapa de características más grande.

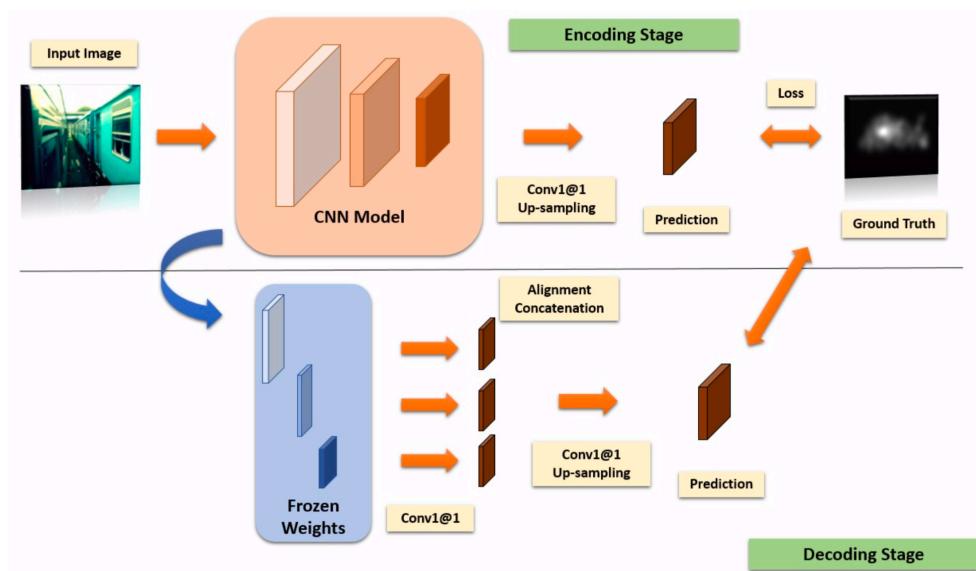


Figura 10. Diagrama de flujo de “Saliency Attentive Model” (Cornia et al., 2018).

III. Bases de datos

Durante los experimentos, el modelo DenseNet-161 se entrena previamente en el conjunto de datos PLACE365, mientras que el modelo NasNet-Large se entrena en ImageNet. Para afinar el modelo se utilizan las bases de datos de MIT1003 y CAT2000, y para evaluarlo las bases de datos de SALICON y MIT300.

IV. Resultados

Como cada modelo CNN se entrena por separado en la etapa de codificación, el espacio computacional que necesita EML-NET depende únicamente del modelo CNN de mayor tamaño. Por lo tanto a EML-NET se pueden incorporar otros modelos CNN con conocimiento previo de otras tareas. Se obtiene un modelo escalable que puede obtener mejores características de saliencia con un conjunto de entrenamiento grande.

3. 4. b. Modelos de saliencia Multiespectral

A diferencia de los modelos de saliencia RGB, que usan únicamente 3 canales de entrada, los modelos que utilizan imágenes multiespectrales cogen como entrada más de 3 canales, que pueden abarcar un amplio rango de longitudes de onda, incluyendo las localizadas fuera del espectro visible. Estas imágenes se consiguen con unos dispositivos de captura específicos que poseen una secuencia de filtros para las diferentes bandas del espectro.

3. 4. b. 1. “Semantic feature based multi-spectral” para la detección de saliencia (Wang et al., 2018)

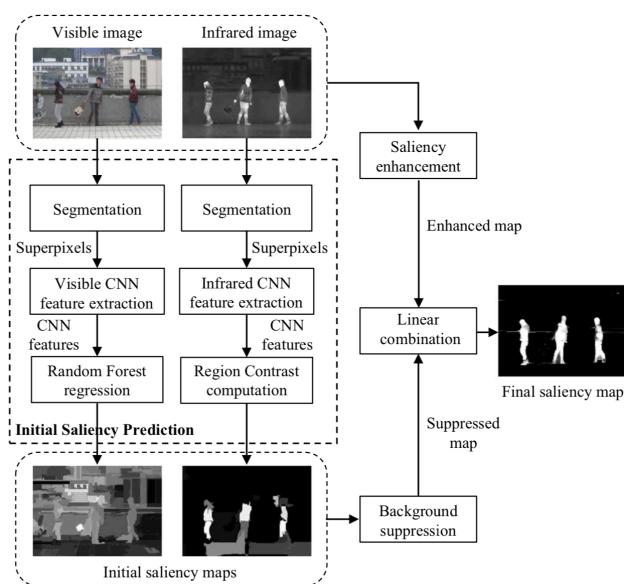
I. Introducción

Haciendo uso de imágenes tomadas en el *infrarrojo térmico**, que obtienen información de la radiación térmica del objeto, se puede mejorar el problema de detección de saliencia para fondos complejos que presentan las imágenes RGB. Esto se debe a que con las imágenes infrarrojas térmicas se obtiene un fondo de imagen que es uniforme y claro, incluso para escenas complejas. El inconveniente de este tipo de imágenes es que no aportan información sobre la textura y el color de los objetos, que con RGB sí obtenemos.

En este estudio se combinan tanto las imágenes RGB como las infrarrojas, para formar el mapa de saliencia.

II. Modelo propuesto

En el modelo, el método de detección de saliencia se puede dividir en 3 bloques. En el primer bloque (el módulo de obtención del mapa de saliencia inicial), con el fin de obtener la información semántica de cada región, se comienzan extrayendo las características de las imágenes visibles e infrarrojas mediante una CNN de 16 capas de profundidad, basándose en una previa *segmentación de superpixeles** en la imagen. Para las características del espectro visible se entrena el *algoritmo de regresión “Random Forest”**, obteniendo así los valores de saliencia. Sin embargo, para las imágenes infrarrojas, la saliencia se calcula con el valor de contraste de cada región.



Tras ambos procesos, en el segundo bloque (“Background Supression”), se fusionan los dos mapas obtenidos en la etapa anterior, para así eliminar el fondo complejo de la imagen. En el módulo de mejora de saliencia, el tercer bloque, se fusionan las imágenes iniciales con un método que captura la información más relevante de los espectros de las imágenes originales, eliminando el ruido de fondo. Finalmente, se obtiene el mapa de saliencia final combinando linealmente los mapas “Enhanced map” y “Supressed map” de las etapas previas.

Figura 11. Estructura del modelo *Semantic feature based multi-spectral*” (Wang et al., 2018).

III. Bases de datos

Las imágenes multiespectrales utilizadas en la investigación provienen de una base de datos de imágenes infrarrojas infAR, usadas para tareas de reconocimiento de acciones. Esta base de datos tiene datos multiespectrales de imágenes en las que 40 personas realizan 12 acciones habituales en ambientes diferentes. Todos los objetos salientes de la imagen son etiquetados manualmente ya que no existen datos para la saliencia en el conjunto de imágenes. Para las imágenes RGB se entrena el algoritmo de regresión “Random Forest” con la base de datos de saliencia MSRA-B.

Para el conjunto de “testing” escogen 300 imágenes de la base de datos infAR. Cada escena con su imagen infrarroja y de espectro visible correspondientes.

VI. Resultados

Con este método se consigue obtener un contorno definido en los objetos, además de un mapa de saliencia constante. También se logra una representación de cuerpos humanos completa, siendo también capaz de localizar los objetos salientes de los bordes de la imagen.

3. 4. b. 2. “Multispectral Saliency Validation” en el visible e infrarrojo cercano (Valero et al., 2019)

I. Introducción

El avance en los últimos años de los dispositivos capaces de capturar imágenes con un amplio rango de longitudes de onda, como el ultravioleta, en infrarrojo o térmico, ha permitido extraer información adicional contenida en cada escena. Estos avances han hecho posible el progreso en una gran cantidad de campos; como el de la robótica, imágenes por satélite, detección de objetos, etc.

En este trabajo se adaptan algunos de los modelos de saliencia más conocidos (ITTI, GBVS, RARE, BMS, LDS y SAM-ResNet) para que puedan utilizar la información multiespectral en la detección de objetos salientes. De esta forma se pretende analizar las ventajas del uso de las imágenes multiespectrales.

II. Modelo

Para el modelo, en primer lugar se extraen las características de las imágenes multiespectrales, obteniendo sus respectivos mapas de características, que en el siguiente paso servirán como entrada para los modelos de saliencia. Mediante una capa de convolución se extrae la información de color del espectro visible, en el espacio de color *CIELab**. Para evitar la correlación entre bandas espectrales adyacentes y mantener las características relevantes, se utiliza la técnica de reducción de dimensionalidad *PCA**, que se usa como características. Por último, para analizar las

diferencias entre los datos espectrales y hallar las zonas distintivas, se emplea una métrica para la discriminación numérica de señales espectrales, en este caso *SAM-SID**.

El siguiente paso es adaptar los modelos a las características obtenidas. Para los modelos de Itti y GBVS, puesto que utilizan intensidad, color y orientación como características, se reemplaza la intensidad y el color por las características de CIELab, la orientación se calcula con el canal de luminosidad, y como características adicionales se añaden PCA y SAM-SID. Para RARE y LDS, ya que utilizan PCA, como entrada se emplea una imagen con 7 canales, sus 3 componentes de PCA, 3 canales de CIELab; y la imagen SAM-SID. En el caso de BMS, los 3 canales a los que el modelo aplica el umbral aleatorio, se sustituyen por los 7 canales obtenidos del conjunto de CIELab, PCA y SAM-SID. Por último, en el modelo SAM-ResNet, se utiliza como entrada cada característica espectral de forma independiente, para después fusionar los mapas de saliencia obtenidos, sin modificar la estructura de la red.

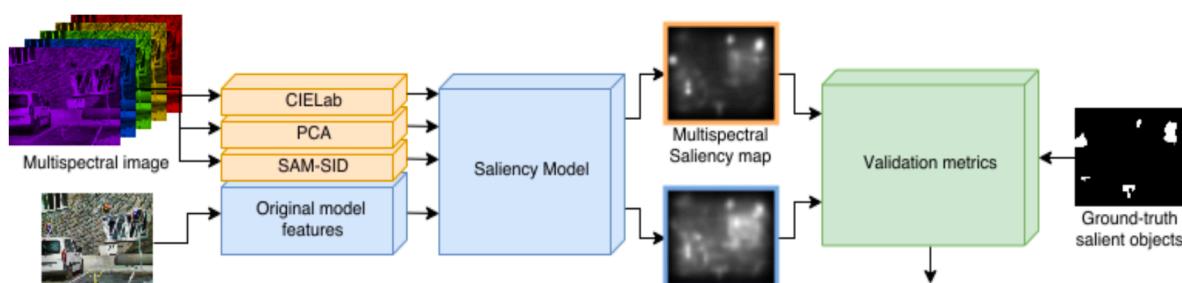


Figura 12. Diagrama del flujo del trabajo “Multispectral Saliency Validation” (Valero et al., 2018).

III. Bases de datos

Para realizar los experimentos con los modelos se utiliza un conjunto propio de 9 imágenes multiespectrales en escenas urbanas, de 370 nm a 1100 nm en 8 canales, y su versión en RGB.

En la obtención de “ground truth” se usan las imágenes RGB, que fueron vistas de forma libre por un total de 6 sujetos, mientras que los movimientos oculares se registraban con un “Eye Tracker”. Los objetos que comprendían un mayor número de fijaciones se segmentaron de forma manual para formar los objetos salientes de la imagen.

IV. Resultados

Primero se utilizan las imágenes RGB en el modelo original para obtener el mapa de saliencia. Después se obtienen un segundo mapa de saliencia utilizando el modelo adaptado en el que se utilizan las características de las imágenes multiespectrales como entrada. Ambos mapas de saliencia se comparan con “ground truth” para comprobar el rendimiento de los modelos en las dos situaciones. Finalmente, de la comparación de cada modelo, utilizando las imágenes multiespectrales respecto de RGB, se obtienen mejores puntuaciones para las características

espectrales, con una media de ganancia del 31.3% para la métrica AUC, y del 56% para los datos de la métrica NSS.

3. 4. b. 3. SUDF: “Saliency from Unsupervised Deep Features” para la detección de saliencia (Imamoglu et al., 2019)

I. Introducción

En algunos trabajos anteriores ya se han propuesto varios estudios que utilizan las características de bajo nivel en imágenes hiperespectrales para la detección de objetos salientes. Sin embargo, las características de alto nivel se pueden utilizar de forma “self-supervised” para datos de imágenes hiperespectrales. Es decir, la contribución de cada banda espectral para la representación del mapa de saliencia, se puede aprender con una red neuronal sin necesidad de utilizar “ground-truth” (“unsupervised”).

Se propone un modelo de detección de saliencia en imágenes hiperespectrales aplicando “*Manifold Ranking*”* a las características de la red neuronal convolucional “self-supervised”, aprendidas durante la segmentación “unsupervised” de la imagen.

II. Modelo

Para obtener el mapa de detección de objetos salientes, en primer lugar se propone un método de “backpropagation”, utilizando la segmentación de las imágenes de forma “unsupervised” (sin aportar datos de saliencia a las imágenes), para que el modelo aprenda las características visuales de alto nivel de forma “self-supervised”, que serán usadas por el algoritmo “*Manifold Ranking*” para el cálculo del mapa de saliencia.

Como entrada se cogen las imágenes hiperespectrales, donde los píxeles se normalizan. Luego se incorpora una CNN para extraer los mapas de características del lote de imágenes de entrada. Después de normalizar el mapa características obtenido en la red, se le aplica la función “Argmax Classification” que clasifica cada pixel eligiendo la dimensión de máximo valor, obteniendo así la etiqueta “Cluster”. Seguidamente, la etiqueta “Cluster” se somete a un proceso de refinado, basado en la etiqueta de “Superpixel”, que se ha obtenido anteriormente mediante la extracción de características de alto nivel en la CNN. Se utilizan las etiquetas “Cluster” y “Superpixel” para obtener “Refined Label”, la etiqueta final. El proceso de refinado se consigue asignando la misma etiqueta a cada área de superpíxeles, en función de la frecuencia más alta. Tras el proceso anterior, se utiliza la función de pérdida “cross-entropy loss” entre la respuesta de la CNN y de “Refined Label” para actualizar los filtros de la CNN.

El proceso anterior se repite un número determinado de veces (hasta que se consigue un error definido), conformando un método “self-supervised” para la red de segmentación. En cada iteración se utiliza la etiqueta “Superpixel Label” (que incluye las características de alto nivel

aprendidas, generadas por la CNN) para obtener un mapa de saliencia empleando “Manifold Ranking”.

III. Bases de datos

Para evaluar el modelo se utiliza la base de datos HS-SOD, que contiene 60 imágenes hiperespectrales con sus respectivas “ground truth” binarias. Cada imagen se compone por 81 canales que comprenden un rango de longitudes de onda de 380 nm hasta 780 nm, en intervalos de 5 nm.

VI. Resultados

Los mapas de saliencia para las diferentes iteraciones del aprendizaje “self-supervised”, en la tarea de segmentación “unsupervised”, muestran que conforme avanza el número de iteraciones aumenta la semejanza entre el mapa obtenido y “ground truth”. Se obtiene un buen rendimiento de la propuesta incorporada al modelo.

De los resultados experimentales se obtiene que este modelo supera a los modelos de saliencia hiperespectral de vanguardia, lo que corrobora un modelo útil. Así, la obtención de un modelo “self-supervised” aplicable, resulta muy beneficioso teniendo en cuenta la falta de datos que existen para los modelos de saliencia, en especial para modelos que utilizan imágenes multiespectrales o hiperespectrales.

3. 5. Evaluación cuantitativa de los modelos

En este apartado se van a evaluar y comparar algunos de los modelos que se han tratado durante el documento. Para ello se utilizan diferentes métricas que han sido explicadas en el apartado 3.1.e, y algunas de las bases de datos del apartado 3.1.d.

3. 5. a. Modelos de predicción de saliencia

| | AUC-Judd | EMD | NSS |
|------------|----------|------|------|
| EML-NET | 0.87 | 1.05 | 2.38 |
| SAM-ResNet | 0.88 | 1.04 | 2.38 |
| SAM-VGG | 0.88 | 1.07 | 2.38 |
| eDN | 0.85 | 2.64 | 1.30 |

Tabla 1. Evaluación en las diferentes métricas con la base de datos CAT2000. La mejor marca dentro de cada métrica se muestra de color rojo, y la peor marca con el color azul (Mit saliency benchmark, 2019).

| Mit300 | AUC-Judd | EMD | NSS |
|------------|----------|------|------|
| EML-NET | 0.88 | 1.84 | 2.47 |
| SAM-ResNet | 0.87 | 2.15 | 2.34 |
| SAM-VGG | 0.87 | 2.14 | 2.30 |
| eDN | 0.82 | 4.56 | 1.14 |

Tabla 2. Evaluación en las diferentes métricas con la base de datos MIT300. La mejor marca dentro de cada métrica se muestra de color rojo, y la peor marca con el color azul (Mit saliency benchmark, 2019).

Tanto en la base de datos de MIT300 como en la de CAT2000, para el orden de clasificación de los modelos se toma como referencia la métrica de AUC-Judd. Con esta métrica, en la base de datos MIT300, el modelo EML-NET consigue el mejor resultado, logrando el segundo puesto para “MIT Saliency Benchmark”. Mientras que para la base de datos CAT2000, en la métrica AUC-Judd, EML-NET se coloca por debajo de SAM-ResNet y SAM-VGG, en la sexta posición. El motivo, tras la bajada de rendimiento para la base de datos CAT2000, parece ser debido a la diferencia que existe entre las imágenes de CAT2000 y las imágenes de entrenamiento del modelo EML-NET (ImageNet y PLACE365) (Jia y Bruce, 2019).

Teniendo en cuenta los modelos de SAM-ResNet y SAM-VGG, para la métrica AUC-Judd en la base de datos de MIT300, consiguen el sexto y séptimo puesto para “MIT Saliency Benchmark”, respectivamente. Sin embargo, para la base de datos CAT2000, SAM-ResNet obtiene el primer puesto y SAM-VGG el segundo.

Para las dos bases de datos, en todas las métricas, se puede observar que los modelos de “deep learning” actuales consiguen resultados muy por encima del primer modelo con “deep learning”, eDN. En “MIT Saliency Benchmark”, para la base de datos de MIT300, el modelo eDN se posiciona en el puesto 25, y para CAT2000, en octava posición, aunque en puestos superiores respecto de otros modelos que no utilizan “deep learning”

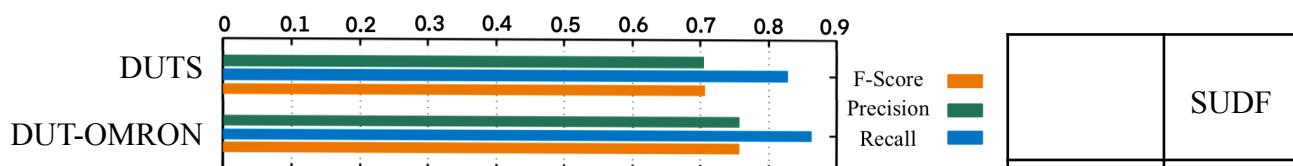
Para las métricas EMD y NSS, el modelo eDN sigue obteniendo los peores resultados, mientras que los modelos de EML-NET, SAM-ResNet y SAM-VGG, intercambian la mejor marca dependiendo de la métrica y la base de datos utilizada. Tanto la diferencia de resultados en cuanto al uso de diferentes bases de datos para una misma métrica, como la diferencia que se muestra al utilizar diferentes métricas para una misma base de datos, resaltan la importancia de aplicar distintas métricas y poder tener acceso a diversas bases de datos para validar los modelos de saliencia de una forma objetiva.

| | | AUC | | NSS | |
|------------|----------------|------|--------|------|--------|
| SAM-ResNet | RGB | 0.42 | 63.7 % | 2.34 | 0.04 % |
| | Hiperespectral | 0.69 | | 2.44 | |

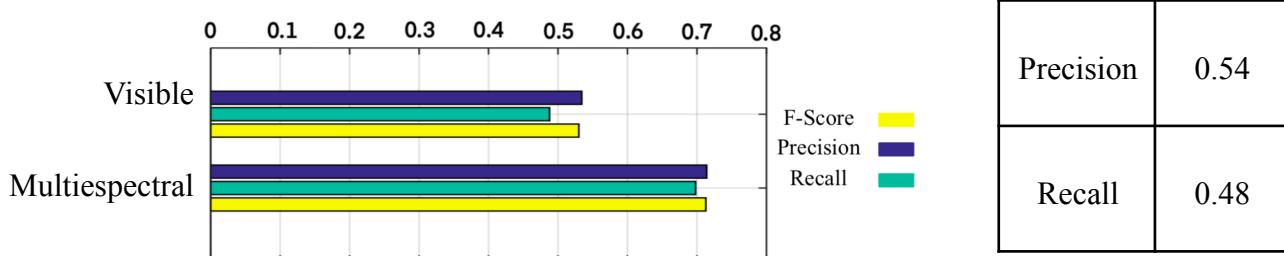
Tabla 3. Evaluación en las métricas de AUC y NSS en la base de datos del trabajo “Multispectral Saliency Validation”. La mejor marca dentro de cada métrica se muestra de color rojo, y la peor marca con el color azul (Valero, 2019).

En el estudio “Multispectral Saliency Validation”, para su propia base de datos (en RGB e hiperespectral), tomando como referencia el modelo SAM-ResNet, se obtiene una mejora del rendimiento cuando se utilizan las características de las imágenes hiperespectrales, tanto en la métrica AUC como para la métrica NSS. Con un aumento considerable para la métrica AUC, del 63.7%. Aunque para la métrica NSS la mejora haya sido solamente del 0.04%, es un buen resultado, teniendo en cuenta que para el modelo SAM-ResNet no se modificó la estructura interna de la red y simplemente se utilizaron las características espectrales como entrada para el modelo (Eva Valero *et al.*, 2018).

3. 4. b. Modelo de detección de saliencia



Gráfica 1. Modelo “Stagewise Refinement”. Métricas “F-Score”, “Precision” y “Recall” para las bases de datos DUTS y DUT-OMRON (Wang *et al.*, 2017).



Gráfica 2. Modelo “Semantic feature based multi-spectral”. Métricas “F-Score”, “Precision” y “Recall” para las bases de datos infAR (Wang *et al.*, 2018).

| | |
|-----------|------|
| | SUDF |
| F-Score | 0.47 |
| Precision | 0.54 |
| Recall | 0.48 |

Tabla 4. Datos de las métricas para el modelo SUDF en la base de datos HS-SOD (Imamoglu *et al.*, 2019).

En las dos gráficas y la tabla anteriores se muestran los resultados que obtienen cada modelo para las métricas de “Precision”, “Recall” y “F-Score”. Para “Precisión” y “Recall”, puede observarse que tanto el modelo “Stagewise Refinement” como el de “Semantic feature based multi-spectral” consiguen valores cercanos a 1, que sería el valor ideal para ambas métricas. También puede verse que en “Semantic feature based multi-spectral” el uso de las imágenes multiespectrales mejora

considerablemente los resultados. Siendo también importante que para “Precision”, “Recall” y “F-Score”, la diferencia entre los resultados sea la mínima posible. En este sentido, el modelo SUDF consigue buenos resultados, en “F-Score” logra un 7% y 1% de diferencia hasta “Precision” y “Recall”, respectivamente, y una diferencia de 6% entre “Precision” y “Recall”. En cambio, en la gráfica 1 se muestra que el modelo “Stagewise Refinement” obtiene peores resultados.

Indicar, que la diferencia entre los resultados de las métricas para cada modelo, igual que para los modelos de predicción de saliencia, dependen de la base de datos utilizada. Por este motivo, una marca superior no implica un mejor rendimiento estrictamente, debiendo de compararse con las mismas bases de datos y de diferentes tipos.

4. CONCLUSIONES

Todos los modelos de saliencia tienen en común pretender obtener un resultado lo más parecido posible a “ground truth”, es decir, intentar simular al máximo la respuesta visual humana durante la observación de una escena. Las contribuciones de los diferentes modelos se utilizan con este fin. En este sentido, “Saliency Attentive Model” logra imitar de forma eficiente la selectividad a las diferentes regiones de la imagen que existe en el comportamiento humano. Otras investigaciones han permitido el uso de redes preentrenadas, y han aumentado la disponibilidad de bases de datos con información sobre saliencia, ayudando así en el progreso de los resultados. En este aspecto, el modelo EML-NET permite incorporar redes profundas preentrenadas, al conseguir una estructura que disminuye la necesidad de memoria en el sistema. Mientras que el modelo de “Saliency from Unsupervised Deep Features” obtiene buenos resultados con un sistema “self-supervised” en el que no se necesita “ground truth”. Como otra aportación, “Stagewise Refinement” ha solventado el problema de la falta de resolución en las imágenes de salida, necesaria para poder obtener resultados precisos. Por otra parte, teniendo en cuenta los modelos que investigan con el uso de imágenes multiespectrales o hiperespectrales, se están obteniendo resultados prometedores que, como se demuestra en la investigación “Multispectral Saliency Validation”, suponen una mejora bastante significativa respecto de los modelos que utilizan imágenes RGB.

Todas las propuestas sobre los modelos de saliencia, ya sea para mejorar su estructura o investigar nuevos métodos, consiguen aportar novedades que día tras día contribuyen a la obtención de modelos con un mayor rendimiento. De esta manera, los “saliency benchmarks” juegan un papel bastante importante al permitir acceder a una calificación actualizada sobre los modelos de saliencia que consiguen los mejores resultados. Con esta finalidad, utilizan varias métricas y diferentes bases de datos, que como se ha comprobado en el apartado anterior, es fundamental teniendo en cuenta la diferencia de resultados que existen al evaluar un mismo modelo con métricas o bases de datos distintas.

Según la información de los artículos más novedosos, el futuro parece estar orientado al uso de estructuras con “deep learning” e imágenes multiespectrales, aunque en la actualidad, existen muy pocas investigaciones que traten este tipo de imágenes para los modelos de saliencia. Conforme se

expanda el uso de los modelos multiespectrales, aumentando las bases de datos específicas y enfocando la metodología a estos modelos, se podrían conseguir resultados más favorables. Además, el conocimiento del sistema visual y su funcionamiento es determinante para poder obtener modelos de saliencia cada vez más fiables, ya que se precisan sistemas semejantes a la visión humana. El avance en el conocimiento de la respuesta y mecanismo del sistema visual, desde el procesado retiniano hasta el cerebral, ha supuesto grandes avances a lo largo de la historia, y de la misma forma lo hará en el futuro, ya que aún existen muchos aspectos visuales y cerebrales que se desconocen. En definitiva, los modelos de atención o modelos de saliencia siguen en desarrollo y les queda un largo recorrido, siendo un campo en auge para la inteligencia artificial al que poder aportar los conocimientos de muchas disciplinas, incluyendo, por supuesto, el de la óptica.

5. AGRADECIMIENTOS

Gracias a la Universidad de Granada por haber aportado y facilitado los permisos y las bases de datos necesarias para adquirir los documentos científicos utilizados. Mi gratitud también al departamento de Óptica de la Universidad de Granada, en especial a la Profesora Eva M. Valero por tutorizar, aportar material y resolver las dudas de este trabajo fin de grado. Y finalmente, al Co-tutor Juan L. Nieves.

6. REFERENCIAS

*Todas las referencias que se aportan son de fuentes en inglés. Cada una de ellas aparece mencionada durante el texto.

6. 1. Libros

- Duchowski, AT. *Eye Tracking Methodology. Theory and Practice*. Londres: Springer; 2017.
- Goldstein, EB. *Sensation and Perception*. Internacional; Wadsworth; 2009.
- Li, J; Gao, W. *Visual Saliency Computation. A machine Learning Perspective*. Suiza: Springer; 2014.
- Mancas, M; Ferrera, VP; Riche, N; Taylor, JG. *From Human Attention to Computational Attention. A Multidisciplinary Approach*. Nueva York: Springer; 2016.
- Schanda, J. (Ed.). *Colorimetry: understanding the CIE system*. Panonia: John Wiley & Sons; 2007.

6. 1. Artículos científicos y webs científicas

- Albawi, S; Mohammed, T; Al-Zawi, S. Understanding of a convolutional neural network. IEEE. *2017 International Conference on Engineering and Technology (ICET)*; Agosto, 2017; Antalya, Turquía: p. 1-6.
- Borji A, Bylinskii Z, Judd T, Itti L, Durand F, Oliva A; Torralba A, “Mit saliency benchmark,” *Mit saliency benchmark*, consulta 1 de Julio de 2019 en: http://saliency.mit.edu/results_mit300.html
- Borji A; Itti L, “Cat2000: A large scale fixation dataset for boosting saliency research,” *arXiv preprint*, 14 de Mayo de 2015, en: <https://arxiv.org/abs/1505.03581>
- Borji, A. What is a salient object? A dataset and a baseline model for salient object detection. *IEEE Transactions on Image Processing* 2014; 24(2): 742-756.
- Cornia, M; Baraldi, L; Serra, G; Cucchiara, R. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing* 2018; 27(10): 5142-5154.
- Deshpande, A, “A Beginner’s Guide To Understanding Convolutional Neural Networks”, *CS Undergrad at UCLA ('19)*, 2017, en: <https://adeshpande3.github.io/A-Beginner%27s-Guide-To-Understanding-Convolutional-Neural-Networks/>
- Imamoglu N, Ding G, Fang Y, Kanezaki A, Kouyama T; Nakamura R, Salient object detection on hyperspectral images using features learned from unsupervised segmentation task. *arXiv preprint*, 28 de Febrero de 2019, en: <https://arxiv.org/abs/1902.10993>
- Imamoglu, N; Oishi, Y; Zhang, X; Ding, G; Fang, Y; Kouyama, T; Nakamura, R. Hyperspectral Image Dataset for Benchmarking on Salient Object Detection. IEEE. *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*; Mayo 2018; Cagliari, Italia: p. 1-3.
- Itti, L; Koch, C; Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 1998; (11): 1254-1259.
- Jia S; Bruce N, “Eml-net: An expandable multi-layer network for saliency prediction. *arXiv preprint*, 2 de Mayo de 2018, en: <https://arxiv.org/abs/1805.01047>
- Jiang, M; Huang, S; Duan, J; Zhao, Q. Salicon: Saliency in context. IEEE. Proceedings of the IEEE conference on computer vision and pattern recognition; Junio, 2015; Boston, Massachusetts: p. 1072-1080.
- Judd T, Durand F; Torralba A, “A benchmark of computational models of saliency to predict human fixations.,” *MIT Libraries*, 13 de Enero de 2012, en: <https://dspace.mit.edu/handle/1721.1/68590>
- Judd, T; Ehinger, K; Durand, F; Torralba, A. Learning to predict where humans look. IEEE. *2009 IEEE 12th international conference on computer vision*; Septiembre, 2009; Kyoto, Japón: p. 2106-2113.
- Kruthiventi, S; Ayush, K; Babu, R. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing* 2017; 26(9): 4446-4456.

- Ng A; Ngiam J; Foo C; Mai Y; Suen C; Coates A; Maas A; Hannun A; Huval B; Wang T; Tandon S, “Feature Extraction Using Convolution.,” *UFLDL Tutorial*, 2 de Junio 2019, en: <http://ufldl.stanford.edu/tutorial/supervised/FeatureExtractionUsingConvolution/>
- Oliva, A; Torralba, A; Castelhano, M; Henderson, J. M. Top-down control of visual attention in object detection. IEEE. *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*; Septiembre, 2003; Barcelona, España: p. 1-253.
- Ruan X, Tong N; Lu H, “How far we away from a perfect visual saliency detection - DUT-OMRON: a new benchmark dataset,” *The DUT-OMRON Image Dataset*, 7 de Enero de 2019, en: <http://saliencydetection.net/dut-omron/>
- Valero, E; Nieves, JL; Etchebehere, S. Multispectral saliency validation in the visible and near-infrared. *Sometido a Color Research and Application*, 2019.
- Vig, E; Dorr, M; Cox, D. Large-scale optimization of hierarchical features for saliency prediction in natural images. IEEE. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2014; Columbus, Ohio: p. 2798-2805.
- Wang, T; Borji, A; Zhang, L; Zhang, P; Lu, H. A stagewise refinement model for detecting salient objects in images. IEEE. *Proceedings of the IEEE International Conference on Computer Vision*; 2017; Venecia, Italia: p. 4019-4028.
- Wang, L; Gao, C; Jian, J; Tang, L; Liu, J. Semantic feature based multi-spectral saliency detection. *Multimedia Tools and Applications* 2018; 77(3): 3387-3403.
- Zhai, Y; Shah, M. Visual attention detection in video sequences using spatiotemporal cues. ACM. *Proceedings of the 14th ACM international conference on Multimedia*; Octubre, 2006; California, EEUU: p. 815-824.

7. ANEXOS

7. 1. GLOSARIO

Las palabras que aparecen en el glosario se muestran durante el texto en cursiva y con un asterisco ().*

Algoritmo de regresión “Random Forest”: es un método estadístico y de análisis de datos que se utiliza en “machine-learning”. Es útil en el aprendizaje para la clasificación y regresión (Wang *et al.*, 2018).

CIELab: es una escala colorimétrica que describe a los colores con 3 ejes, blanco y negro (L^*), rojo y verde (a^*), y amarillo y azul (b^*). Esta representación simula bien la percepción humana del color (Schanda, 2007).

Concatenar: se refiere a la operación de solapar mapas de características con la misma dimensión, aumentando la profundidad y la información del mapa resultante (Wang *et al.*, 2017).

Imágenes infrarrojas térmicas: la longitud de onda del infrarrojo térmico va desde los 15 μm hasta los 8 μm (NASA, 2010).

Manifold Ranking: es un algoritmo de aprendizaje “semisupervised” que explora la relación entre todos los puntos de los datos. Tiene dos versiones dependiendo de la tarea que realice: clasificar los puntos de los datos, o predecir las etiquetas para los datos (Tong *et al.*, 2006).

PCA: “Principal Component Analyses” es una técnica de transformación lineal que se utiliza para reducir la dimensionalidad. El objetivo de PCA es encontrar el mejor conjunto de componentes ortogonales para representar los datos, mientras que retienen la mayor variabilidad en los datos (Valero *et al.*, 2019).

SAM-SID: es una técnica utilizada en la discriminación de señales multiespectrales de forma numérica. Para ello se compara cada píxel con la señal media para todos los canales. SAM es “Spectral Angle Mapper” y SID “Spectral Information Divergence”, ambas métricas son combinadas (Valero *et al.*, 2019).

Segmentación de superpíxeles: es un procesamiento que se encarga de dividir la imagen en regiones uniformes, denominadas superpíxeles. Cada píxel contenido en un superpíxel comparte características de intensidad, color o textura con los demás píxeles. Esta representación reduce la complejidad de la imagen y facilita la extracción de características (Liu *et al.* 2014).

Semánticas del objeto: se refiere a la información que existe al relacionar el contexto de la imagen con las propiedades del objeto (Wang *et al.*, 2017).

Sistema escalable: es un sistema que no muestra efectos negativos cuando su tamaño o complejidad aumentan. Estos sistemas pueden manejar una cantidad de trabajo creciente con facilidad, o pueden ampliarse para que el trabajo requerido se desarrolle sin aumentar el costo de recursos, como la memoria (Jia y Bruce, 2019).

7. 2. REFERENCIAS ADICIONALES USADAS EN EL GLOSARIO

- “Infrared Waves,” National Aeronautics and Space Administration, *Science Mission Directorate*, 2010 en: https://science.nasa.gov/ems/07_infraredwaves
- Liu, M; Tuzel, O; Ramalingam, S; Chellappa, R. Entropy-rate clustering: Cluster analysis via maximizing a submodular function subject to a matroid constraint. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2013; 36(1): 99-112.
- Tong, H; He, J; Li, M; Ma, W; Zhang, H; Zhang, C. Manifold-ranking-based keyword propagation for image retrieval. *EURASIP Journal on Advances in Signal Processing* 2006; 2006(1): 079412.