



Trabajo Fin de Grado

Estudio y evaluación de métodos de predicción de
scanpaths en vídeos 360º

Study and evaluation of scanpath prediction
methods for 360º video

Autor

Juan Lorente Guarnieri

Directores

Edurne Bernal Berdún
Daniel Martín Serrano

Ponente

Ana Belén Serrano Pacheu

AGRADECIMIENTOS

Quiero expresar mi más sincero agradecimiento a mis tutores, **Ana, Daniel y Edurne**, por su orientación y apoyo durante la realización de este trabajo. Su conocimiento y consejos han sido esenciales para llevar a cabo este proyecto.

Agradezco también a la **Universidad de Zaragoza** y al **Graphics and Imaging Lab** por los recursos y el entorno académico proporcionado, y a mis compañeros de clase y amigos, por su apoyo y colaboración.

Finalmente, a mi familia, especialmente a mis padres y mi hermano, por su amor y apoyo incondicional a lo largo de estos años.

Gracias a todos.

Juan Lorente

Estudio y evaluación de métodos de predicción de scanpaths en vídeos 360º

RESUMEN

En los últimos años la realidad virtual ha ganado una gran relevancia en diversas áreas como los videojuegos, la educación y la medicina, entre otras. Para crear experiencias inmersivas y realistas, es importante comprender la atención humana: ¿Qué llama más la atención de los observadores? ¿Qué regiones son más interesantes? Encontrar respuesta a estas preguntas es desafiante, ya que hay una gran variabilidad en el comportamiento visual humano. En este contexto, la capacidad de predecir los patrones de atención visual de los usuarios puede beneficiar la creación de contenido, permitiendo prever qué zonas son de mayor potencial interés, ayudando así a los creadores a optimizar sus diseños.

Este Trabajo de Fin de Grado se centra en analizar y comparar diferentes enfoques de predicción de *scanpaths* en vídeos 360º, i.e., representaciones de la trayectoria que sigue la mirada de un observador a lo largo del tiempo. Estos scanpaths proporcionan información sobre qué elementos visuales capturan más la atención del usuario y cómo se procesan visualmente a lo largo del tiempo.

Primero, se han implementado siete métodos heurísticos basados en el muestreo de mapas de saliencia, una representación espacial común de la atención visual. Después, se ha diseñado e implementado un modelo de predicción con redes neuronales profundas, basándose en dos modelos del estado del arte, estudiando su potencial en la predicción de trayectorias visuales. Para la evaluación de estos métodos se han utilizado distintas métricas bien establecidas para evaluar la recursividad y las distancias mínimas entre trayectorias.

Los resultados obtenidos sugieren que los métodos avanzados de muestreo de saliencia, como el muestreo probabilístico y la inhibición de retorno, replican de manera más precisa los patrones de exploración visual humana en comparación con métodos más básicos. Por otro lado, el modelo de aprendizaje profundo ha obtenido resultados comparables con los métodos heurísticos, dejando abiertas numerosas líneas de investigación.

Índice

1. Introducción y objetivos	1
1.1. Contexto del trabajo	1
1.2. Objetivos del trabajo	3
1.3. Planificación y herramientas	4
2. Estudio previo	7
2.1. Trabajos Relacionados	7
2.1.1. ScanGAN360	7
2.1.2. SaltiNet	8
2.1.3. tSPM-Net	8
2.1.4. ScanpathNet	9
2.1.5. SST-Sal	10
2.1.6. VPT360	10
3. Estudio y evaluación de técnicas de muestreo de saliencia	13
3.1. Tipos de muestreos estudiados	14
3.1.1. Muestreo aleatorio	14
3.1.2. Valor máximo de la saliencia	14
3.1.3. Sampleo aleatorio con percentiles sobre la saliencia predicha . .	15
3.1.4. Sampleo probabilístico sobre la saliencia predicha	15
3.1.5. Inhibición de retorno sobre el sampleo probabilístico	15
3.1.6. Sesgo ecuatorial sobre el muestreo de inhibición de retorno . .	16
3.1.7. Máxima distancia entre cada punto sobre el muestreo de inhibición de retorno	16
3.2. Evaluación	17
3.3. Métricas	17
3.4. Dataset e implementación	18
3.5. Resultados	18

4. Diseño e implementación de un modelo de predicción de scanpaths en video 360º	23
4.1. Arquitectura	23
4.2. Función de pérdida	25
5. Evaluación y resultados	27
5.1. Sampleo con un modelo con sobreajuste	27
5.1.1. Implementación	27
5.1.2. Resultados	28
5.1.3. Conclusiones	28
5.2. Estudio de un modelo generalista	31
6. Conclusiones	33
6.1. Limitaciones y mejoras futuras	34
7. Bibliografía	37
Lista de Figuras	39
Lista de Tablas	41
Anexos	42
A. Métodos de Visualización de Scanpaths	45
A.1. Visualización de Puntos con Escala de Colores	45
A.2. Visualización de múltiples <i>scanpaths</i> con delimitaciones del viewport . .	46
A.3. Visualización de Thumbnails con el <i>scanpath</i>	47

Capítulo 1

Introducción y objetivos

1.1. Contexto del trabajo

La realidad virtual (RV) ha experimentado un crecimiento notable en los últimos años, revolucionando sectores como los videojuegos, el entretenimiento, o la educación. Esta tecnología permite a los usuarios sumergirse en entornos tridimensionales creados digitalmente, proporcionando experiencias interactivas que pueden simular situaciones tanto reales como fantásticas. La evolución constante del hardware, como los cascos de realidad virtual y los controladores hápticos, junto con los avances en software, han facilitado que la RV sea cada vez más accesible y atractiva para el público en general. Empresas líderes en tecnología y entretenimiento, como Meta, Adobe, o Valve han adoptado la RV para crear experiencias únicas que abarcan desde juegos y simulaciones de entrenamiento hasta aplicaciones en medicina y educación.

La creación de contenido atractivo e inmersivo en RV depende en gran medida de la comprensión de lo que capta la atención de las personas. Generar contenido interesante requiere un conocimiento profundo de la atención humana en entornos de realidad virtual: ¿Qué llama más la atención a las personas? ¿Cómo exploran las personas el contenido virtual? ¿Qué características influyen más a la hora de dirigir la atención? Encontrar respuestas a estas preguntas es esencial para diseñar experiencias que no solo sean visualmente impactantes, sino también intuitivas y naturales para los usuarios, mejorando la inmersión y la satisfacción general.

Una de las principales formas de representar y medir la atención es mediante los mapas de saliencia. Un mapa de saliencia es una representación visual que destaca las áreas de una imagen o video que son más probables de atraer la atención visual de los observadores. Estos mapas se computan agregando datos de la mirada de múltiples observadores viendo una escena, y asignan mayores valores a las zonas de interés, que pueden depender de características visuales (como el color, el contraste y la orientación) y contextuales (como la familiaridad y la relevancia de los objetos). Los mapas de

saliencia son útiles para identificar y predecir patrones de atención visual en diferentes tipos de contenidos visuales, o proporcionar a los creadores de contenido una idea preliminar sobre el comportamiento esperado de sus usuarios.

No obstante, la saliencia no tiene en cuenta el aspecto temporal de la atención, como el orden en que se explora una escena. Para paliar esta limitación, se recurre a otra representación de la atención visual: los *scanpaths*.

Un *scanpath* es la trayectoria que sigue la mirada de un observador al explorar una imagen o video a lo largo del tiempo. Se trata de una secuencia de fijaciones y movimientos sacádicos, que reflejan el interés y la atención del observador. Las fijaciones son los períodos en los que los ojos permanecen relativamente quietos y se centran en un punto específico, mientras que las sacadas son los movimientos rápidos entre fijaciones. El análisis de *scanpaths* proporciona información valiosa sobre cómo los usuarios procesan visualmente la información a lo largo del tiempo y qué elementos llaman más su atención.

Para entrenar modelos de predicción de *scanpaths*, se utilizan datos que incluyen registros de movimientos oculares y fijaciones de observadores humanos. Estos datos son esenciales para enseñar a los modelos a replicar patrones de atención visual humana. Los registros de movimientos oculares se obtienen mediante tecnologías de seguimiento ocular (eye-tracking), que permiten medir con precisión dónde y durante cuánto tiempo los observadores fijan su mirada. Capturar estos datos es costoso y laborioso, lo que ha motivado a muchos investigadores a desarrollar modelos capaces de predecir el comportamiento visual sin necesidad de tantos datos empíricos. Esta necesidad de eficiencia y precisión ha impulsado el diseño de técnicas de muestreo basadas en aprendizaje automático, buscando superar las limitaciones de los métodos heurísticos y mejorar la predicción de los *scanpaths* en entornos de realidad virtual.

Existen diferentes enfoques para predecir *scanpaths* en imágenes, pero en todos ellos, los modelos se centran en características estáticas del contenido visual [1].

Sin embargo, en videos 360º, es crucial considerar la dimensión temporal y la interactividad del entorno [2], [3].

Los videos 360º permiten a los usuarios explorar libremente el contenido visual, lo que añade una capa adicional de complejidad a la predicción de *scanpaths*. Los modelos deben ser capaces de anticipar cómo los usuarios variarán su atención a lo largo del tiempo y en respuesta a cambios en el entorno visual.

La integración de métodos de predicción de *scanpaths* en videos presenta desafíos adicionales debido a la naturaleza dinámica del contenido. A diferencia de las imágenes estáticas, los videos requieren que los modelos adapten sus predicciones en tiempo real a medida que el video se desarrolla. Esto implica no solo predecir las áreas de mayor

saliencia en cada frame del video, sino también cómo estas áreas cambian y cómo los observadores pueden cambiar su foco de atención en respuesta a los eventos en el video [4].

La capacidad de hacer predicciones precisas en tiempo real es crucial para aplicaciones prácticas, como la edición de contenido, la publicidad dirigida y la mejora de la experiencia del usuario en entornos de realidad virtual.

La importancia de esta investigación radica en su capacidad para optimizar la creación de contenidos en RV, mejorando la experiencia del usuario y potencialmente incrementando la retención y el impacto de las aplicaciones inmersivas. Al entender mejor cómo los usuarios interactúan visualmente con entornos 360º, los desarrolladores pueden diseñar contenidos más efectivos y atractivos, que maximicen la atención y la inmersión.

Dado que la predicción de scanpaths en videos 360º es un área aún poco explorada, este trabajo de fin de grado se propone estudiar y comparar diferentes enfoques para predecir scanpaths en videos 360º, analizando su eficacia y precisión en diferente contenido de realidad virtual.

1.2. Objetivos del trabajo

El objetivo principal de este trabajo de fin de grado es diseñar, implementar y evaluar diferentes métodos de muestreo, al igual que un modelo de aprendizaje automático, para la generación de scanpaths en videos 360º. Los objetivos específicos incluyen:

- Estudiar la literatura existente sobre predicción de scanpaths y saliencia en imágenes y videos, con un enfoque particular en contenido 360º.
- Desarrollar y evaluar distintas técnicas de muestreo para generar scanpaths a partir de mapas de saliencia, analizando su efectividad utilizando diversas métricas de evaluación.
- Diseñar e implementar un modelo de generación de scanpaths en videos 360º, utilizando técnicas de aprendizaje profundo.
- Evaluar el modelo desarrollado comparándolo con las técnicas de muestreo previos, utilizando métricas estándar de rendimiento y análisis de resultados.

1.3. Planificación y herramientas

El trabajo se ha desarrollado usando el lenguaje de programación Python, haciendo uso de PyTorch, un framework de aprendizaje automático, y realizando un control de versiones mediante Github. El entrenamiento final del modelo se ha llevado a cabo en una GPU Nvidia Quadro RTX 6000 con 24 GB de memoria facilitada por el Graphics and Imaging Lab de la Universidad de Zaragoza.

El desarrollo del trabajo se ha dividido en varias tareas principales. Inicialmente, se llevó a cabo un estudio previo sobre el estado del arte en temas relacionados y modelos de aprendizaje automático. Posteriormente, se procedió a la implementación de métodos de muestreo, donde se diseñaron y evaluaron diferentes estrategias para seleccionar datos representativos. A continuación, se desarrolló el modelo principal, que incluyó el diseño, implementación y evaluación iterativa del mismo, realizando ajustes y mejoras continuas basadas en los resultados obtenidos. Finalmente, se redactó la memoria del proyecto, documentando detalladamente cada fase del trabajo y los hallazgos más significativos. En la Tabla 1.1 se recogen las horas dedicadas a cada tarea, y en la Figura 1.1 se puede observar el diagrama de Gantt.

Tareas	Horas Dedicadas
Trabajo previo	Estudio previo
	Familiarización con el código
	Implementación visualización
	Total
Desarrollo de los métodos de muestreo	Implementación muestreos
	Evaluación de muestreos
	Total
Desarrollo del modelo	Diseño del modelo
	Implementación del modelo
	Evaluación del modelo
	Total
Elaboración de la memoria	71
Total	322

Tabla 1.1: Horas totales dedicadas al desarrollo del trabajo.

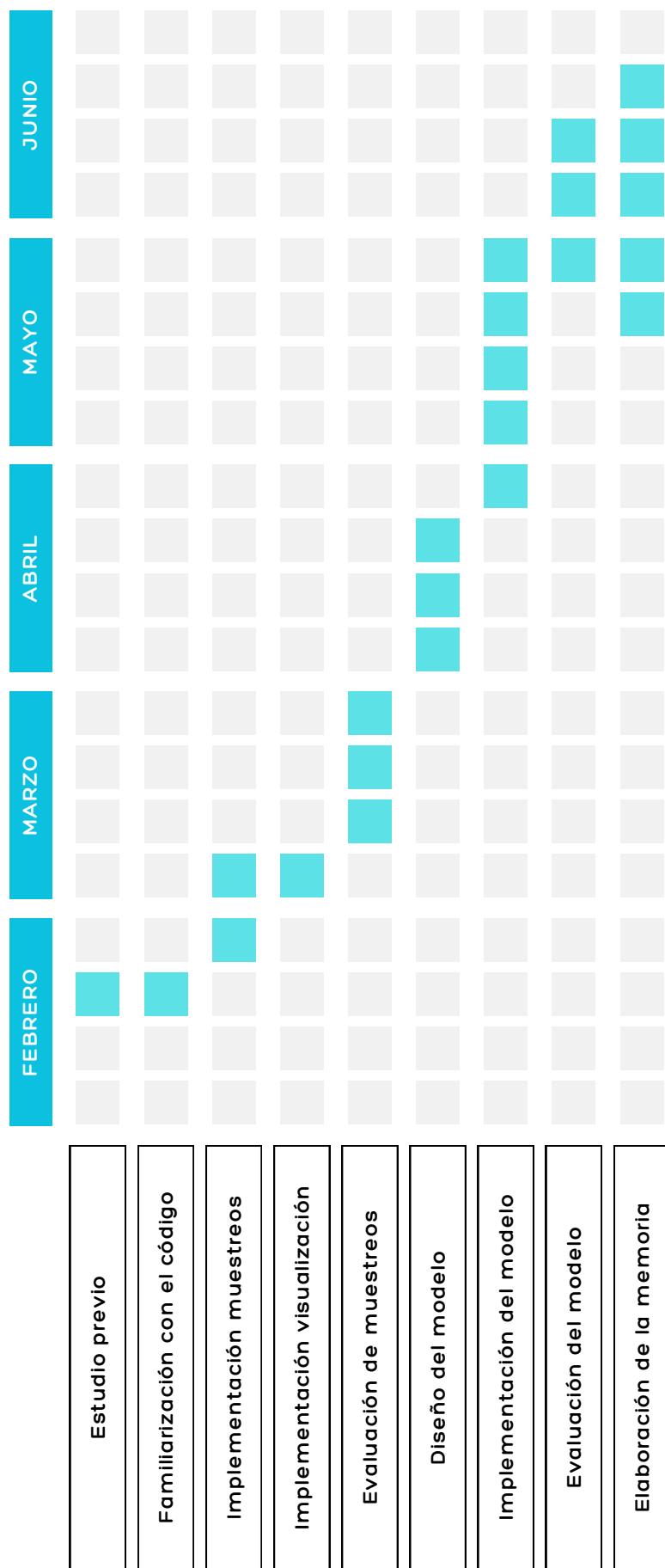


Figura 1.1: Diagrama de Gantt.

Capítulo 2

Estudio previo

El desarrollo de modelos para la predicción de *scanpaths* en videos 360º ha captado recientemente la atención de la comunidad investigadora debido a su potencial para mejorar la experiencia del usuario en aplicaciones de realidad virtual. La tarea de predecir cómo los usuarios moverán su atención en un entorno visual complejo presenta numerosos desafíos, especialmente cuando se consideran las características dinámicas y altamente interactivas de los videos 360º. Varios enfoques han sido propuestos para abordar estos desafíos, desde técnicas heurísticas basadas en características visuales tradicionales hasta el uso de métodos avanzados de aprendizaje automático.

2.1. Trabajos Relacionados

2.1.1. ScanGAN360

ScanGAN360 [5] es un modelo generativo adversarial diseñado para generar rutas de exploración realistas en imágenes 360°. Utiliza una arquitectura de cGAN, una variación de GAN que incorpora información condicional en el proceso de generación de datos. La función de pérdida se basa en Dynamic Time Warping, una métrica que mide la similitud entre dos series temporales, considerando tanto la forma como el orden de los elementos de una secuencia.

ScanGAN360 es un modelo entrenado para generar trayectorias cuyo patrón visual se asemeje al de observadores reales. El modelo es capaz de generar cientos de trayectorias por segundo. Sin embargo, aunque ScanGAN360 genera rutas de exploración de alta calidad, su desempeño podría mejorarse con conjuntos de datos más grandes y variados. El modelo está limitado a una frecuencia de 1Hz, es decir, sólo genera un punto por segundo. Además, actualmente no maneja contenido dinámico como videos 360°.

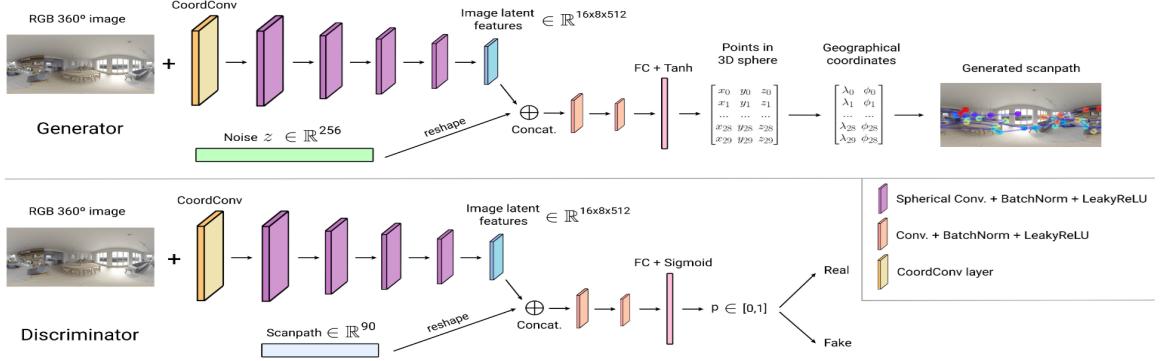


Figura 2.1: Arquitectura del modelo de ScanGAN360 [5].

2.1.2. SaltiNet

SaltiNet [6] es una red neuronal profunda para la predicción de rutas de exploración en imágenes 360°. El modelo se basa en una novedosa representación consciente del tiempo de la información de *saliencia* llamada volúmenes de *saliencia*. La primera parte de la red consiste en un modelo entrenado para generar volúmenes de *saliencia*. Las estrategias de muestreo sobre estos volúmenes se utilizan para generar rutas de exploración en las imágenes 360°. No obstante, aunque SaltiNet genera rutas de exploración efectivas, la probabilidad de una fijación no está condicionada a fijaciones previas, y la longitud de las rutas de exploración y la duración de cada fijación se tratan como variables aleatorias independientes, lo cual limita su precisión.

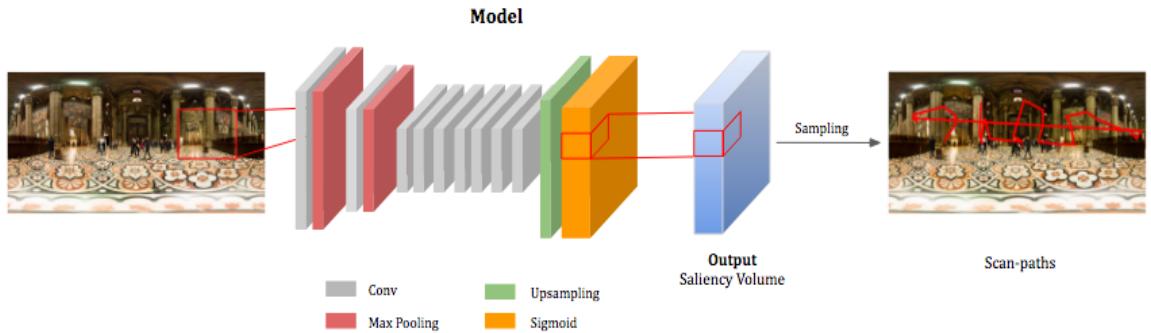


Figura 2.2: Arquitectura del modelo de SaltiNet [6].

2.1.3. tSPM-Net

tSPM-Net [7] es un modelo basado en un enfoque probabilístico para predecir trayectorias de exploración visual completas y plausibles dada una imagen 2D de entrada. Utiliza redes recurrentes de memoria a corto y largo plazo convolucionales (ConvLSTM) construidas sobre aprendizaje profundo bayesiano y una novedosa función de pérdida para la optimización espaciotemporal conjunta. El modelo es capaz de

generar trayectorias realistas que simulan la explicación de observadores reales. Sin embargo, el rendimiento del modelo disminuye cuando las características visuales de una imagen son complejas o demasiado abstractas. Utilizar conjuntos de datos más grandes y variados probablemente haría el modelo más robusto en estos casos.

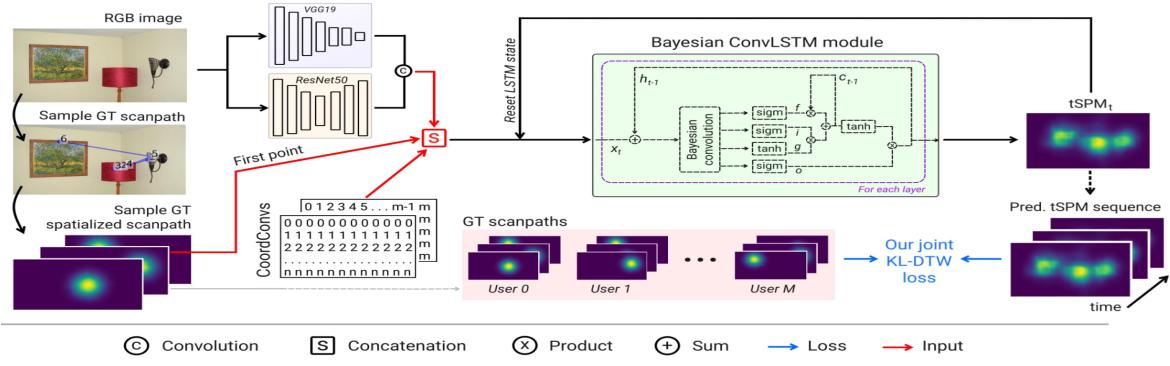


Figura 2.3: Arquitectura del modelo de tSPM-Net [7].

2.1.4. ScanpathNet

ScanpathNet [8] es un modelo de aprendizaje profundo inspirado en el modelo de Búsqueda Guiada 6 (GS6). Este modelo utiliza una red convolucional para extraer características semánticas, una red convLSTM para modelar las dependencias secuenciales de las fijaciones y una red de mezcla de densidades (MDN) para predecir la distribución de probabilidad de las fijaciones en cada píxel. ScanpathNet genera trayectorias de exploración visual simuladas mediante el muestreo secuencial de la salida del modelo. No obstante, ScanpathNet es sensible al número de componentes gaussianos en la red de mezcla de densidades. La complejidad de la escena en las imágenes puede afectar el rendimiento del modelo, y su desempeño podría mejorar con conjuntos de datos más grandes y variados.

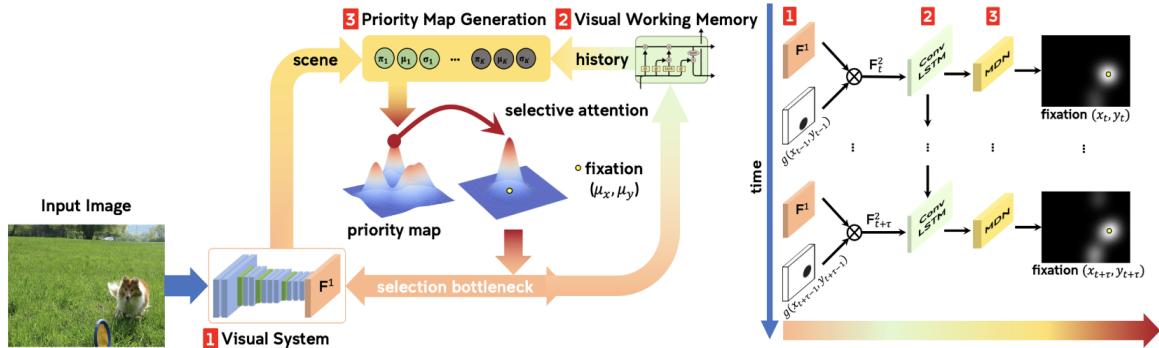


Figura 2.4: Arquitectura del modelo de ScanpathNet [8].

2.1.5. SST-Sal

SST-Sal [9] sigue una arquitectura de encoder-decoder basada en ConvLSTM que tiene en cuenta la información temporal. La red de predicción de *saliencia* está construida sobre convoluciones esféricas y propone una nueva función de pérdida KLDiv esférica, que pondera cada píxel según su ángulo sólido para compensar la distorsión de la proyección equirectangular, así como estimaciones de flujo óptico para aprender las relaciones entre movimiento y *saliencia*. Sin embargo, el modelo de predicción de *saliencia* en videos 360° se limita a videos estáticos y no considera cámaras en movimiento ni estímulos multimodales como pistas acústicas direccionales.

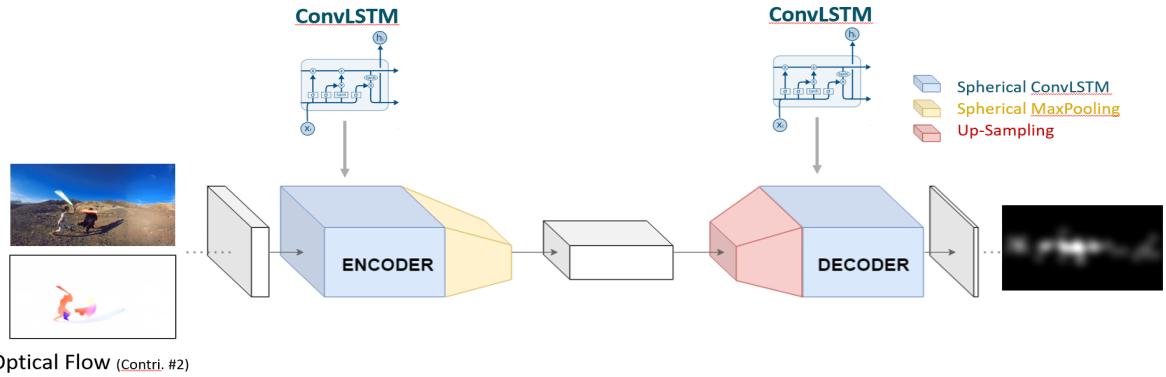


Figura 2.5: Arquitectura del modelo de SST-Sal [9].

2.1.6. VPT360

Un ejemplo notable de un enfoque innovador en la predicción de *scanpaths* en videos 360° es el propuesto por Chao et al [10]. Este enfoque utiliza una arquitectura basada en transformers, denominada VPT360, que únicamente se apoya en el *scanpath* histórico del usuario para predecir la trayectoria futura de su viewport. El objetivo principal de este modelo es mejorar la precisión de la predicción del viewport en tiempo real, lo que es esencial para optimizar las decisiones de transmisión adaptativa en sistemas de realidad virtual. El VPT360 se destaca por su uso del mecanismo de autoatención para capturar dependencias temporales en la secuencia de *scanpaths*, evitando la necesidad de características de contenido adicionales como mapas de *saliencia* o los frames de video, lo que reduce significativamente la complejidad computacional sin comprometer la precisión.

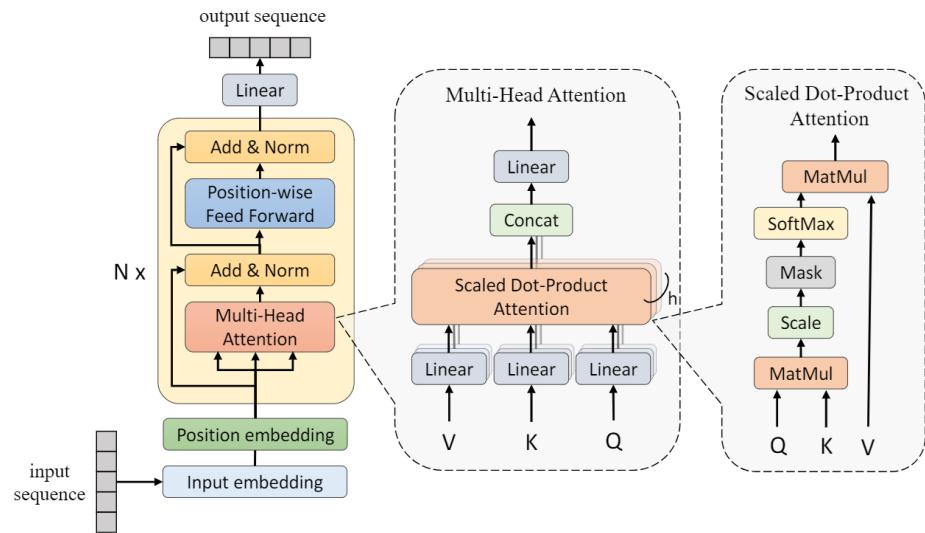


Figura 2.6: Arquitectura del modelo de VPT360 [10].

Capítulo 3

Estudio y evaluación de técnicas de muestreo de saliencia

El objetivo de este capítulo es estudiar y evaluar diferentes técnicas heurísticas de muestreo de saliencia, con el propósito de generar *scanpaths* a partir de mapas de saliencia. Como se ha comentado previamente, los mapas de saliencia son representaciones visuales que indican las regiones de una imagen o un vídeo que son más propensas a captar la atención visual de las personas. Estas predicciones se basan en diversos factores visuales y cognitivos que influyen en cómo los seres humanos perciben y procesan la información visual.

Los métodos de muestreo de saliencia permiten simular la forma en que los observadores exploran visualmente una escena, identificando los puntos de fijación más probables del mapa de saliencia en base a diferentes reglas.

Evaluar estas técnicas de muestreo es crucial para determinar cuáles se asemejan más al comportamiento real de los usuarios. Esto implica comparar los *scanpaths* generados por cada método de muestreo con los datos de fijaciones oculares obtenidos de estudios empíricos con participantes humanos. Al identificar los métodos que mejor replican el comportamiento visual humano, se pueden desarrollar modelos más efectivos y fiables para predecir la atención visual.

El proceso de evaluación incluye el uso de métricas específicas que cuantifican la similitud entre los *scanpaths* generados y los observados. Estas métricas pueden considerar aspectos como la distancia entre puntos de fijación, la coherencia temporal de las fijaciones, o la cobertura de las áreas de interés. A través de esta evaluación, se busca no solo validar la precisión de los métodos de muestreo, sino también comprender las ventajas y limitaciones de cada enfoque en diferentes contextos visuales.

3.1. Tipos de muestreos estudiados

3.1.1. Muestreo aleatorio

El muestreo aleatorio (Figura 3.1) se basa en seleccionar puntos de interés de manera completamente aleatoria dentro del mapa de saliencia. Esta técnica no considera los valores de saliencia y, por lo tanto, no se enfoca en las áreas más probables de captar la atención visual. Sin embargo, puede ser útil como línea base para comparar con otros métodos más sofisticados y evaluar la efectividad de técnicas más avanzadas de muestreo de saliencia.

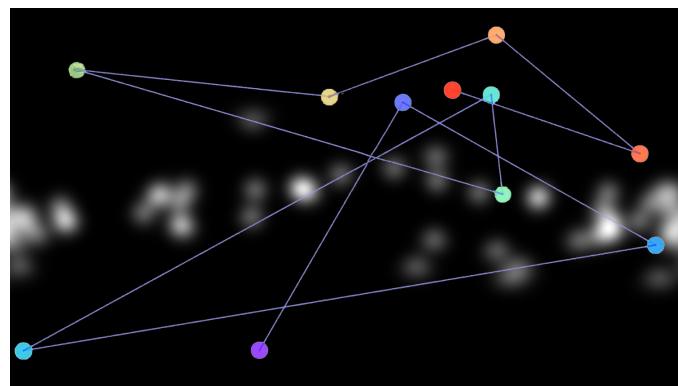


Figura 3.1: Ejemplo de muestreo aleatorio.

3.1.2. Valor máximo de la saliencia

Este método selecciona los puntos donde la saliencia predicha es máxima, de tal manera que siempre obtiene el punto más probable que el usuario vaya a ver en ese instante. Este enfoque es útil para identificar las áreas más llamativas y asegurar que los puntos de fijación se centren en las regiones de mayor interés (Figura 3.2). Al enfocarse en los picos de saliencia, se optimiza la precisión en la predicción de la atención visual.



Figura 3.2: Ejemplo del punto máximo obtenido en rojo sobre el mapa de saliencia.

3.1.3. Sampleo aleatorio con percentiles sobre la saliencia predicha

En este enfoque, los puntos de interés se seleccionan de forma aleatoria dentro de las zonas del mapa de saliencia que están por encima de un determinado percentil, descontando aquellas inferiores. Esto asegura un cierto grado de aleatoriedad mientras se introduce un sesgo hacia las áreas más relevantes.

Para obtener este resultado, se identifican los índices del mapa de saliencia donde los valores superan el umbral definido por el percentil. Los puntos con saliencia inferior a este umbral se descartan. Luego, se selecciona aleatoriamente uno de los puntos restantes para determinar el punto de fijación (Figura 3.3).

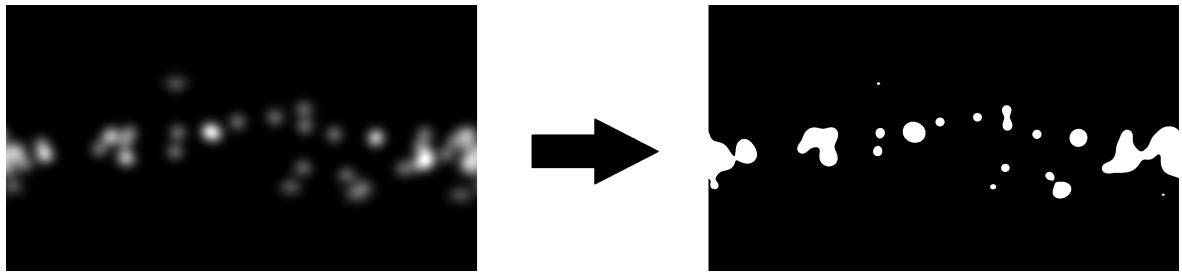


Figura 3.3: Ejemplo de obtención de las zonas superiores a un percentil en un mapa de saliencia. El muestreo se realiza sobre las áreas blancas de forma aleatoria.

3.1.4. Sampleo probabilístico sobre la saliencia predicha

Este método selecciona puntos de fijación basados en una probabilidad directamente proporcional a los valores de saliencia, sin usar percentiles. En lugar de centrarse únicamente en las áreas con saliencia máxima, este enfoque considera todo el mapa de saliencia, permitiendo que cada punto tenga una probabilidad de ser seleccionado proporcional a su valor de saliencia. Esto garantiza que las regiones más llamativas sean más propensas a ser elegidas, pero también permite cierta variabilidad, reflejando así una exploración visual más natural y menos determinística.

3.1.5. Inhibición de retorno sobre el sampleo probabilístico

La inhibición de retorno asegura que, después de seleccionar un punto de fijación, la probabilidad de volver a fijar en esa región disminuye, promoviendo la exploración de nuevas áreas. Este enfoque simula el comportamiento natural del ojo humano, que tiende a evitar fijar repetidamente en la misma área para explorar más eficazmente el entorno visual.

Para implementar esta técnica, se utiliza una máscara sobre el mapa de saliencia para ajustar las probabilidades de fijación (Figura 3.4). Específicamente, se crea una

máscara gaussiana centrada en el punto de fijación previamente seleccionado. Esta máscara se resta del mapa de saliencia original, reduciendo significativamente la probabilidad de seleccionar puntos cercanos a la fijación anterior. Como resultado, se incentiva la selección de nuevos puntos de fijación en áreas no exploradas.



Figura 3.4: Inhibición de retorno en el muestreo probabilístico.

3.1.6. Sesgo ecuatorial sobre el muestreo de inhibición de retorno

Este método aplica un sesgo para preferir puntos cercanos al ecuador del campo de visión, considerando que en entornos 360º, los usuarios tienden a explorar más estas áreas y no mirar hacia el suelo o el techo.

Para conseguir esto, se utiliza una máscara sobre el mapa de saliencia, similar a la utilizada en el método de inhibición de retorno. Esta máscara aumenta la probabilidad de seleccionar puntos que están más cerca del ecuador del mapa de saliencia, disminuyendo la probabilidad en las áreas más cercanas a los polos (suelo y techo, Figura 3.5).

El proceso de predicción del punto se realiza aplicando primero la máscara ecuatorial para ajustar el mapa de saliencia, seguido del muestreo de inhibición de retorno. De esta manera, se garantiza que los puntos de fijación sean preferentemente seleccionados en regiones más relevantes del campo de visión en entornos 360º.

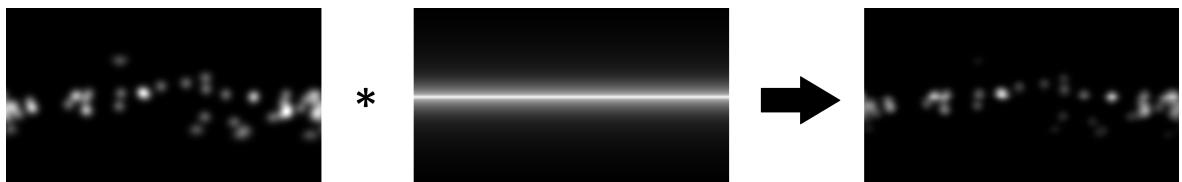


Figura 3.5: Aplicación del sesgo ecuatorial a un mapa de saliencia.

3.1.7. Máxima distancia entre cada punto sobre el muestreo de inhibición de retorno

Para este muestreo se impide que el punto predicho esté a una distancia límite del punto anterior para evitar saltos muy grandes en la predicción, asemejando el límite

de movimiento que tendría un usuario al ver el video 360° , ya que el usuario no tiende a hacer giros bruscos para darse la vuelta.

Para conseguir esto, se hace algo similar al paso de muestreo anterior, en el que se aplica una máscara en forma de mapa gaussiano centrada en el punto previamente seleccionado. Esta máscara se aplica al mapa de saliencia para evitar regiones alejadas del anterior punto predicho (Figura 3.6).

La predicción del punto es el resultado de aplicar la máscara descrita antes de realizar el muestreo de inhibición de retorno.

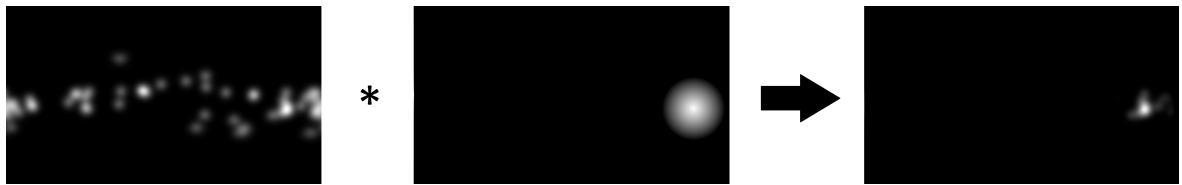


Figura 3.6: Aplicación de la máxima distancia a un mapa de saliencia.

3.2. Evaluación

La evaluación de las técnicas de muestreo de saliencia es un proceso crítico para determinar la efectividad y precisión de los métodos propuestos en la generación de *scanpaths* que se asemejen al comportamiento visual humano. En esta sección, se describen las métricas utilizadas para la evaluación y se presentan los resultados obtenidos.

3.3. Métricas

Para cuantificar la similitud entre los *scanpaths* generados y los observados en estudios empíricos, se han utilizado las siguientes métricas:

- **Dynamic Time Warping (DTW)**[7]: Mide la distancia mínima entre dos secuencias temporales permitiendo alineaciones no lineales. Es útil para comparar la forma de los scanpaths.
- **Distancia de Levenshtein (LEV)**: Calcula el número mínimo de operaciones necesarias para transformar un scanpath en otro, evaluando diferencias en la secuencia de puntos de fijación.
- **Recurrencia (REC)**[11]: Indica la proporción de fijaciones que ocurren en las mismas áreas en diferentes scanpaths, reflejando patrones recurrentes en la exploración visual.

- **Determinismo (DET)**[11]: Evalúa la proporción de fijaciones que forman líneas diagonales en la matriz de recurrencia, indicando la previsibilidad en el comportamiento de exploración.

3.4. Dataset e implementación

Para evaluar los distintos muestreos, se utiliza el dataset D-SAV360 [12]. Este conjunto de datos contiene 4,609 trayectorias de cabeza y mirada (*scanpaths*) en videos 360° con sonido ambisónico de primer orden, recopilados de 87 participantes que visualizaron 85 videos diferentes. A diferencia de otros datasets que carecen de estímulos multimodales, D-SAV360 permite un estudio más integral de la interacción multimodal en el comportamiento visual dentro de entornos de realidad virtual.

Para esta evaluación, se han utilizado 10 vídeos. Para cada uno se han computado 10 *scanpaths* con cada uno de los métodos de muestreo. Después, se han comparado con los datos reales del dataset utilizando las métricas explicadas anteriormente, obteniendo la media y la desviación típica.

3.5. Resultados

Los resultados obtenidos en la evaluación de las técnicas de muestreo de saliencia se resumen en la Tabla 3.1. Estos resultados comparan los distintos métodos introducidos, con un baseline humano (Human BL) para evaluar su similitud con el comportamiento visual humano.

Metric	DTW ↓		DET ↓		REC ↑		LEV ↓	
Human BL	5831.540	1140.614	5.559	1.111	10.594	5.858	337.147	51.754
Random	18019.572	1443.864	7.215	1.401	2.940	0.568	740.369	47.512
Max saliency	6054.192	1398.365	5.884	2.573	11.744	8.453	342.056	56.360
Percentile	6049.975	1400.671	6.333	2.709	11.729	8.189	341.738	53.206
Probabilistic	6042.207	1390.408	6.368	2.732	11.751	8.225	341.755	53.856
Inhibition	6019.938	1379.746	6.325	2.705	11.772	8.169	341.562	53.992
SE	5976.197	1370.896	6.399	2.752	12.048	8.199	339.910	54.417
MD	6148,027	1428.771	5.550	1.380	9.014	5.166	361.431	25.504
SE + MD	5887.859	1154.755	7.029	1.474	9.780	4.669	356.153	20.648

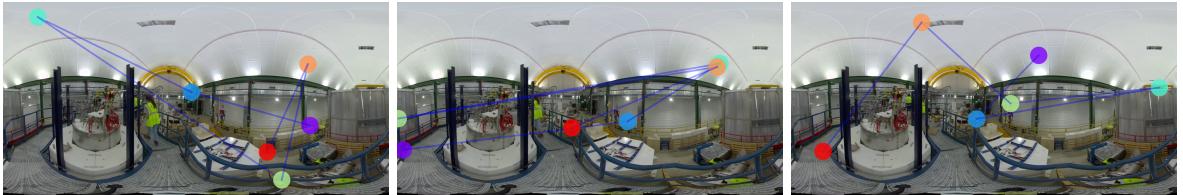
Tabla 3.1: Resultados comparativos de diferentes métodos de muestreo, SE corresponde al sesgo ecuatorial y MD al muestreo de máxima distancia

En la Figura 3.7 y en la Figura 3.8 se muestran ejemplos de *scanpaths* computados con los distintos tipos de muestreo utilizando uno de los métodos de visualización de *scanpaths* explicados en el Anexo A. Todos los videos generados se pueden ver en el <https://drive.google.com/drive/folders/1gkSex0MygpvBAmNsLWKtCKvXrsftBzvL>.

Ground Truth



Random



Max Saliency



Percentile



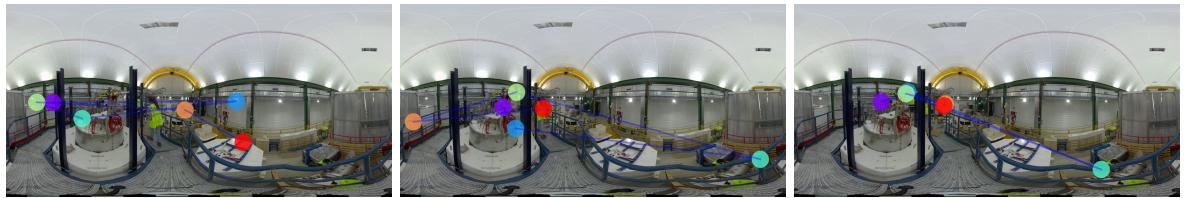
Probabilistic



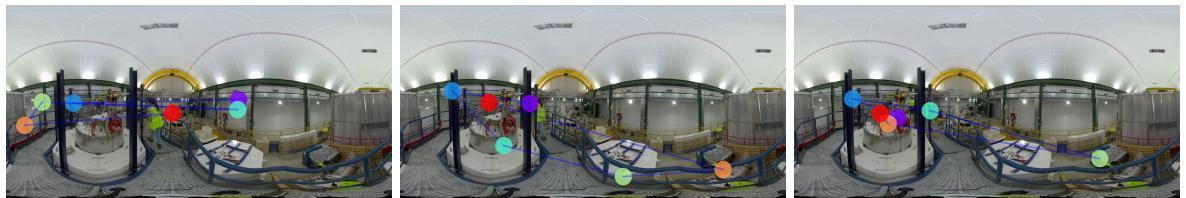
Time

Figura 3.7: Visualización de *scanpaths* para los distintos métodos de muestreo teniendo en cada fila un método y en columna un instante del video con id 1016 del dataset D-SAV360 [12]. El resto de métodos de muestreo está en la Figura 3.8. Cabe destacar que, después del punto rojo en una imagen, el *scanpath* continúa con el punto violeta en la imagen siguiente.

Inhibition of return



Equatorial Bias



Max Distance



Equatorial Bias + Max Distance

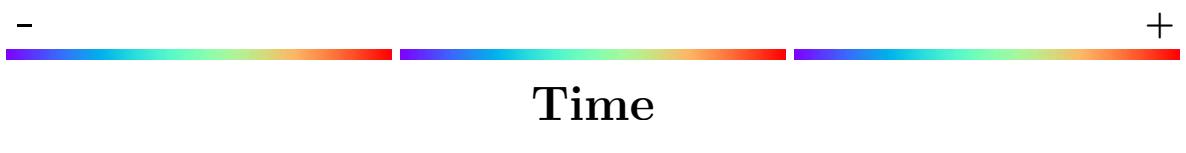


Figura 3.8: Visualización de *scanpaths* para los distintos métodos de muestreo teniendo en cada fila un método y en columna un instante del video con id 1016 del dataset D-SAV360 [12]. El resto de métodos de muestreo está en la Figura 3.7. Cabe destacar que, después del punto rojo en una imagen, el *scanpath* continúa con el punto violeta en la imagen siguiente.

Conclusiones:

- Muestreo aleatorio: El método de muestreo aleatorio muestra un rendimiento significativamente inferior en todas las métricas comparado con otros métodos. Esto se debe a su falta de consideración por los valores de saliencia, lo que resulta en puntos de fijación no representativos.

- Valor máximo de la saliencia: Seleccionar los puntos de máxima saliencia mejora notablemente la precisión del determinismo (DET) en comparación con el resto de muestreos, pero conlleva una menor variabilidad en los scanpaths.
- Sampleo aleatorio con percentiles: Introducir aleatoriedad dentro de zonas de alta saliencia mejora la variabilidad sin comprometer significativamente la precisión, aunque de los muestreos implementados es el que peor resultados obtiene.
- Sampleo probabilístico: Este método logra un equilibrio entre precisión y variabilidad, mostrando resultados consistentes en múltiples métricas.
- Inhibición de retorno: La incorporación de la inhibición de retorno mejora la exploración visual al evitar la fijación repetida en las mismas áreas, resultando en una mayor similitud con los patrones de exploración humanos, obteniendo mejores resultados tanto en la alineación temporal dinámica (DTW), como en la recurrencia (REC) en comparación con los métodos de muestreo anteriores.
- Sesgo ecuatorial y máxima distancia: Ambos enfoques adicionales proporcionan ajustes finos que pueden mejorar la relevancia en contextos específicos, mejorando resultados en la métrica de la distancia de Levenshtein (LEV), en la alineación temporal dinámica (DTW) y en la recurrencia (REC).

Los resultados de este estudio resaltan la importancia de considerar las características específicas del comportamiento visual humano al diseñar métodos de muestreo de saliencia. Los datos muestran que las técnicas que integran elementos como la inhibición de retorno y la maximización de la saliencia logran una mayor similitud con los patrones de exploración visual humanos, mejorando la precisión y la variabilidad de los scanpaths generados. En particular, la combinación de métodos probabilísticos y de sesgo ecuatorial ofrece un equilibrio óptimo entre la representatividad y la eficiencia del muestreo.

Capítulo 4

Diseño e implementación de un modelo de predicción de *scanpaths* en video 360º

La motivación principal para el diseño de este modelo radica en evaluar la posibilidad de desarrollar una técnica de muestreo basada en aprendizaje automático que supere en precisión a los métodos heurísticos descritos en el Capítulo 3. La idea es explorar si los enfoques de aprendizaje profundo pueden ofrecer predicciones de *scanpaths* más precisas y representativas del comportamiento visual humano en entornos de realidad virtual, superando las limitaciones de los métodos heurísticos.

4.1. Arquitectura

Uno de los objetivos de este trabajo es desarrollar un modelo de predicción probabilística de *scanpaths* en videos 360º. Para ello, se ha diseñado una arquitectura inspirada en las arquitecturas de SST-Sal [9] y tSPM-Net [7].

SST-Sal, con su arquitectura de encoder-decoder basada en ConvLSTMs, permite capturar relaciones temporales y espaciales en videos 360º mediante convoluciones esféricas y estimaciones de flujo óptico, mejorando así las predicciones de saliencia. Por su parte, tSPM-Net emplea un enfoque probabilístico con una función de pérdida espaciotemporal, utilizando KLDiv esférica y un módulo de atención, logrando modelar la incertidumbre y las dinámicas de la escena.

La arquitectura propuesta (Figura 4.1) integra estas fortalezas para crear un sistema robusto y eficiente en la predicción de trayectorias de la mirada en entornos inmersivos. A continuación, se detallan los pasos y consideraciones clave para la implementación de esta arquitectura.

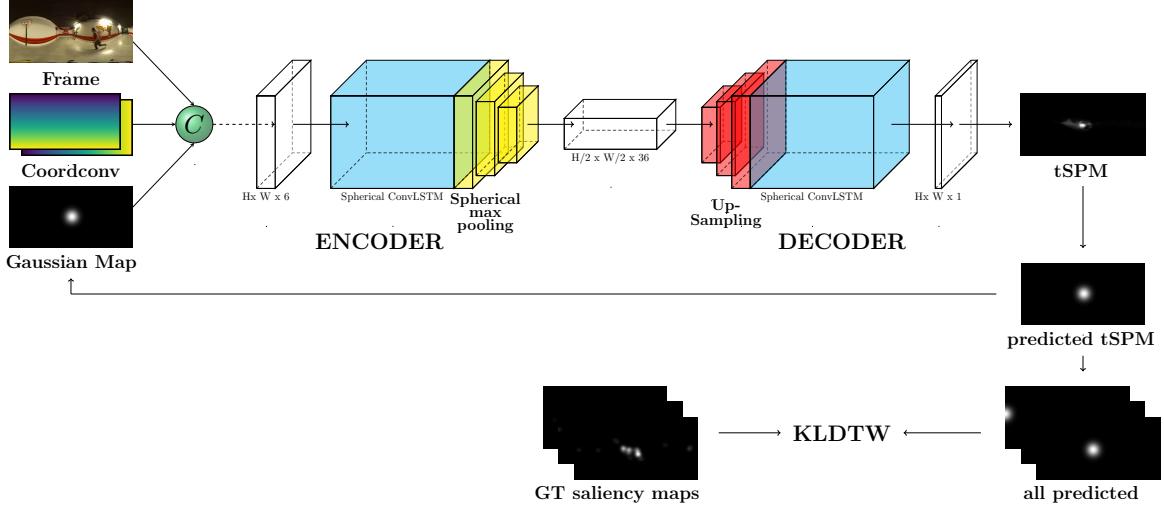


Figura 4.1: Diseño de la arquitectura del modelo de predicción de scanpaths en videos 360º, inspirada en SST-Sal [9] y tSPM-Net [7].

En primer lugar, para representar los *scanpaths* en un formato adecuado, se sigue el enfoque de tSPM-Net, donde cada punto del *scanpath* se representa como un mapa gaussiano centrado en dicho punto. Este método permite convertir las coordenadas de la mirada en una representación espacial continua y diferenciable, facilitando su uso como entrada en redes neuronales convolucionales.

Se ha modificado la arquitectura de SST-Sal para que acepte estos mapas gaussianos como entrada. Se han mantenido los 6 canales en la entrada, 3 canales para la información RGB del frame actual del vídeo, 1 canal para el mapa gaussiano del punto anterior predicho, y 2 canales para introducir una matriz con coordenadas convolucionales (CordConv [13]) para que el modelo sea capaz de entender la arquitectura espacial de las imágenes, tal y como sucede en tSPM-Net [7]. Nótese que se podría configurar el modelo para añadir más canales y módulos adicionales para que tenga mayor contexto o para experimentar con diferentes configuraciones y elegir la que mejor se adapte al objetivo deseado.

La salida del modelo es un “tSPM” (Mapa Probabilístico de Scanpath en evolución Temporal), que indica las probabilidades de los posibles puntos de mirada futuros. Debido a que la arquitectura de SST-Sal ya está diseñada para predecir mapas de saliencia, no es necesario editar el modulo de salida, ya que el modelo será el que aprenderá a generar el tSPM durante el entrenamiento.

tSPM-Net necesita un mapa gaussiano como entrada. Para obtenerlo, se toma el tSPM predicho en el paso anterior y se utiliza un método de muestreo (Capítulo 3) sobre él. Para este caso se ha decidido usar el sampleo probabilístico ya que en la evaluación de estos sampleos no se notó una diferencia muy notoria en todas las métricas, además

de que el sesgo ecuatorial y la inhibición de retorno debería aprenderla el modelo y no depender de la variabilidad inconsistente que aportaría el usar estos sampleos. También teniendo en consideración que el predecir un punto dependiendo de las probabilidades de la saliencia es uno de los sampleos más rápidos computacionalmente y con mejor ratio precisión-variabilidad.

4.2. Función de pérdida

Para optimizar el modelo de predicción de *scanpaths*, se utiliza la función de pérdida KLSoftDTW (*Dynamic Time Warping* Suavizado con Divergencia de *Kullback-Leibler*) modelada en el trabajo de tSPM-Net [7]. Esta función combina las ventajas del *Dynamic Time Warping* (DTW) con la divergencia de *Kullback-Leibler* (KL) para evaluar la similitud entre secuencias temporales de distinta longitud.

El DTW mide la similitud entre dos secuencias temporales considerando tanto la forma como el orden de los elementos, sin forzar una correspondencia uno a uno. Esto se logra alineando las dos series temporales de manera que la distancia entre ellas sea mínima. La fórmula para DTW es:

$$\text{DTW}(r, s) = \min_A \langle A, \Delta(r, s) \rangle,$$

donde A es una matriz binaria de alineación entre dos series temporales r y s , $\Delta(r, s) = [\delta(r_i, s_j)]_{i,j}$ es una matriz que contiene las distancias $\delta(\cdot, \cdot)$ entre cada par de puntos en r y s , y $\langle \cdot, \cdot \rangle$ denota el producto interior entre ambas matrices.

Dado que la función mínima no es diferenciable, se ha propuesto [7] una versión suavizada:

$$\text{DTW}_\gamma(r, s) = \min_A \gamma \langle A, \Delta(r, s) \rangle, \quad \gamma > 0.$$

La función \min_γ está definida como:

$$\min_\gamma(a_1, \dots, a_N) = -\gamma \log \sum_{i=1}^N \exp\left(-\frac{a_i}{\gamma}\right).$$

El parámetro γ ajusta la similitud entre la versión suavizada y el algoritmo DTW original, siendo ambos iguales cuando $\gamma = 0$.

Por otro lado, el KLDiv mide la diferencia entre dos distribuciones de probabilidad. La divergencia de Kullback-Leibler (DKL) se define como:

$$\text{DKL}(P \parallel Q) = \sum_j P(j) \log \frac{P(j)}{Q(j)},$$

donde P y Q son las distribuciones de probabilidad a comparar, y j se refiere a cada punto de la distribución [7]. En nuestro caso particular, cada punto de mirada se representa de manera espacial, por lo que KL-Div puede proporcionar una medida cuantitativa sobre cuán diferentes son dos puntos de mirada basados en sus mapas de probabilidad.

Al combinar estas dos métricas en la función de pérdida KLSoftDTW, se puede evaluar y optimizar la red de manera más eficaz, asegurando que las predicciones del modelo sean lo más precisas y realistas posible. Esta combinación permite capturar tanto las variaciones temporales como las probabilísticas en las trayectorias de mirada, mejorando significativamente el rendimiento del modelo en entornos dinámicos y multimodales [7].

Capítulo 5

Evaluación y resultados

Para entrenar y evaluar el modelo, se ha utilizado D-SAV360 [12], el mismo que en la evaluación de los muestradores (Capítulo 3). En esta sección, se discutirá un enfoque principal en el que la evaluación del modelo de predicción de *scanpaths* se realiza mediante sobreajuste (overfitting) sobre el mismo video a evaluar, y luego se realizarán pruebas adicionales para examinar la capacidad de generalización del modelo.

Estos análisis permitirán determinar la precisión y eficacia del modelo en diferentes contextos y su potencial para superar a los métodos heurísticos tradicionales. Nótese que este modelo pretende ser una primera aproximación basada en aprendizaje automático para el muestreo de *scanpaths*. Las potenciales líneas de mejora y limitaciones se describen en el Capítulo 6.

5.1. Sampleo con un modelo con sobreajuste

5.1.1. Implementación

Para evaluar la eficacia del modelo de predicción de *scanpaths* en videos 360º, se ha adoptado una estrategia de sobreajuste, entrenando el modelo únicamente con un video específico y evaluándolo con el mismo video durante la inferencia. Esta metodología tiene como objetivo explorar la capacidad del modelo para aprender patrones detallados de *scanpaths* y evaluar su precisión en un entorno controlado. Posteriormente, se evaluará si el modelo, actuando como técnica de muestreo basada en aprendizaje automático, puede ofrecer resultados más precisos que los métodos heurísticos.

Los entrenamientos se han realizado con un ratio de aprendizaje (lr) de 0.001, durante 1000 épocas, tamaño de batch de 1 (debido a limitaciones de la arquitectura original de tSPM-Net [7]), en una GPU Nvidia Quadro RTX 6000 con 24 GB de memoria con una duración media de entrenamiento de 2 horas.

5.1.2. Resultados

Para comparar el resultado con las otras técnicas de muestreo, se entrenó 10 modelos (uno para cada vídeo) y se generaron 10 *scanpaths* para cada uno. Después, se compararon (Tabla 5.1) usando la misma metodología y métricas que en la evaluación de los métodos de muestreo (Capítulo 3).

Metric	DTW ↓		DET ↓		REC ↑		LEV ↓	
Human BL	5831.540	1140.614	5.559	1.111	10.594	5.858	337.147	51.754
Random	18019.572	1443.864	7.215	1.401	2.940	0.568	740.369	47.512
Max saliency	6054.192	1398.365	5.884	2.573	11.744	8.453	342.056	56.360
Percentile	6049.975	1400.671	6.333	2.709	11.729	8.189	341.738	53.206
Probabilistic	6042.207	1390.408	6.368	2.732	11.751	8.225	341.755	53.856
Inhibition	6019.938	1379.746	6.325	2.705	11.772	8.169	341.562	53.992
SE	5976.197	1370.896	6.399	2.752	12.048	8.199	339.910	54.417
MD	6148.027	1428.771	5.550	1.380	9.014	5.166	361.431	25.504
SE + MD	5887.859	1154.755	7.029	1.474	9.780	4.669	356.153	20.648
DL model	6138.950	1057.466	10.416	2.129	10.756	5.672	335.997	32.225

Tabla 5.1: Resultados comparativos de diferentes métodos de sampleo con el modelo con sobreajuste.

En la Figura 5.1 se muestran ejemplos de *scanpaths* computados con los distintos tipos de muestreo utilizando uno de los métodos de visualización de *scanpaths* explicados en el Anexo A. Todos los videos generados se pueden ver en <https://drive.google.com/drive/folders/1gkSex0MygpvBAmNsLWKtCKvXrsftBzvL>.

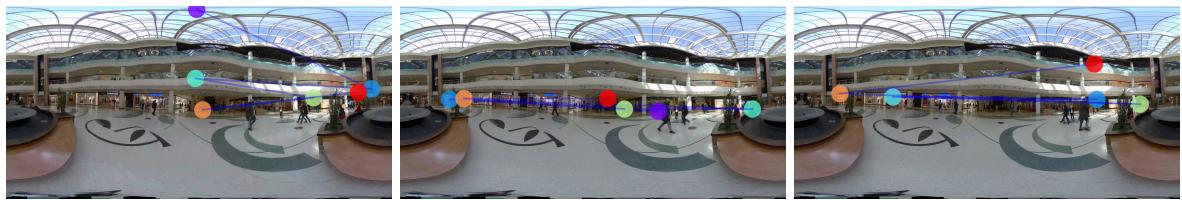
5.1.3. Conclusiones

En la evaluación comparativa de las técnicas de muestreo utilizando el modelo con sobreajuste, se observaron resultados variados según las diferentes métricas. A continuación, se presentan las conclusiones basadas en los resultados cuantitativos y un análisis cualitativo de las características de la atención visual aprendidas por el modelo:

Resultados Cuantitativos:

- Dynamic Time Warping (DTW): Los resultados para DTW fueron peores en comparación con cualquier otro método de muestreo. Esto sugiere que los *scanpaths* generados por el modelo presentaban mayor disimilitud en su forma respecto a los *scanpaths* reales. Aunque el modelo puede aprender patrones específicos de *scanpaths*, no consigue replicar la forma general de los *scanpaths* con precisión.

id 0002



id 0011



id 1004



id 1005



id 1016



Figura 5.1: Visualización de *scanpaths* para distintos vídeos usando respectivo modelo sobreajustado, en el que para cada fila hay un vídeo y en una columna es el instante del video con el id del dataset D-SAV360 [12]. Cabe destacar que, después del punto rojo en una imagen, el *scanpath* continúa con el punto violeta en la imagen siguiente.

- Determinismo (DET): Similar al DTW, los resultados del DET fueron notablemente peores para el modelo. Los resultados bajos indican que los *scanpaths* generados por el modelo mostraban un comportamiento menos predecible y más aleatorio en comparación con los métodos heurísticos.
- Recurrencia (REC): Aunque los resultados de REC fueron un poco peores que los de los otros métodos de muestreo, la diferencia no fue tan marcada. Una ligera desventaja en esta métrica sugiere que el modelo tiene cierta capacidad para identificar áreas de interés recurrente, aunque no lo haga tan efectivamente como los métodos heurísticos.
- Distancia de Levenshtein (LEV): En contraste, los resultados de la distancia de Levenshtein (LEV) fueron mejores para el modelo en comparación con los métodos heurísticos. Un mejor desempeño aquí indica que el modelo es capaz de replicar de manera más fiel la secuencia de fijaciones, incluso si no logra alinear perfectamente las trayectorias temporales o la previsibilidad.

Análisis Cualitativo:

El análisis cualitativo revela que el modelo ha aprendido características relevantes de la atención visual, lo que se refleja en varios aspectos importantes:

- Inhibición de retorno: El modelo muestra una tendencia a evitar volver a fijarse en áreas ya exploradas, una característica consistente con la inhibición de retorno observada en el comportamiento visual humano. Esta propiedad ayuda a aumentar la eficiencia exploratoria al prevenir fijaciones redundantes.
- Saltos bruscos: Se observa que el modelo evita realizar saltos de fijación a grandes distancias, prefiriendo movimientos más suaves y controlados. Esta característica es importante para evitar movimientos bruscos que podrían resultar incómodos o no naturales en una visualización 360º.
- Sesgo ecuatorial: Al igual que en el comportamiento humano, el modelo parece exhibir un sesgo hacia áreas centrales del campo visual, donde los objetos tienden a ser más prominentes. Este sesgo ecuatorial indica que el modelo está aprendiendo a priorizar regiones visuales más relevantes.
- Muestreo probabilístico: Utilizando un enfoque de muestreo probabilístico, el modelo es capaz de generar *scanpaths* que no sólo replican la secuencia de fijaciones sino que también reflejan la distribución probabilística de la atención visual. Esto mejora la representatividad de los patrones de exploración generados.

En conclusión, aunque el modelo de muestreo basado en aprendizaje automático muestra potencial en términos de replicar la secuencia de fijaciones (LEV), su desempeño en términos de forma general de *scanpaths* (DTW) y previsibilidad (DET) es inferior al de los métodos heurísticos. La ligera desventaja en la métrica de recurrencia (REC) sugiere que, aunque el modelo puede identificar áreas de interés, no lo hace con la misma precisión que los métodos heurísticos de muestreo (Capítulo 3). Estos hallazgos indican que, si bien el modelo tiene áreas donde sobresale, especialmente en la secuencia de fijaciones, necesita mejoras significativas para competir con los métodos heurísticos en términos de forma y previsibilidad de los *scanpaths*. Además, el análisis cualitativo sugiere que el modelo ha capturado algunas características inherentes de la atención visual humana, como la inhibición de retorno, la evitación de saltos bruscos, y el sesgo ecuatorial, lo que abre la puerta a futuras investigaciones para refinar estas capacidades y mejorar el rendimiento global del modelo.

5.2. Estudio de un modelo generalista

Para evaluar la capacidad de generalización del modelo, se realizaron pruebas en videos distintos a los de entrenamiento. Se entrenó el modelo con un conjunto de videos y se evaluó su rendimiento en otro diferente, dividiendo el dataset en 80 % para entrenamiento y 20 % para validación, asegurando la representación de la diversidad de escenarios y comportamientos.

Se utilizaron 15 videos: 12 para entrenamiento y 3 para validación, con los mismos hiperparámetros que el modelo con sobreajuste. El entrenamiento duró 11 horas.

Los resultados mostraron que el modelo tenía tendencia a mezclar características de distintos videos, produciendo tSPMs con zonas que combinaban elementos de varios videos. Esto sugiere que, aunque efectivo en un entorno controlado, el modelo enfrenta desafíos en la generalización a nuevos datos.

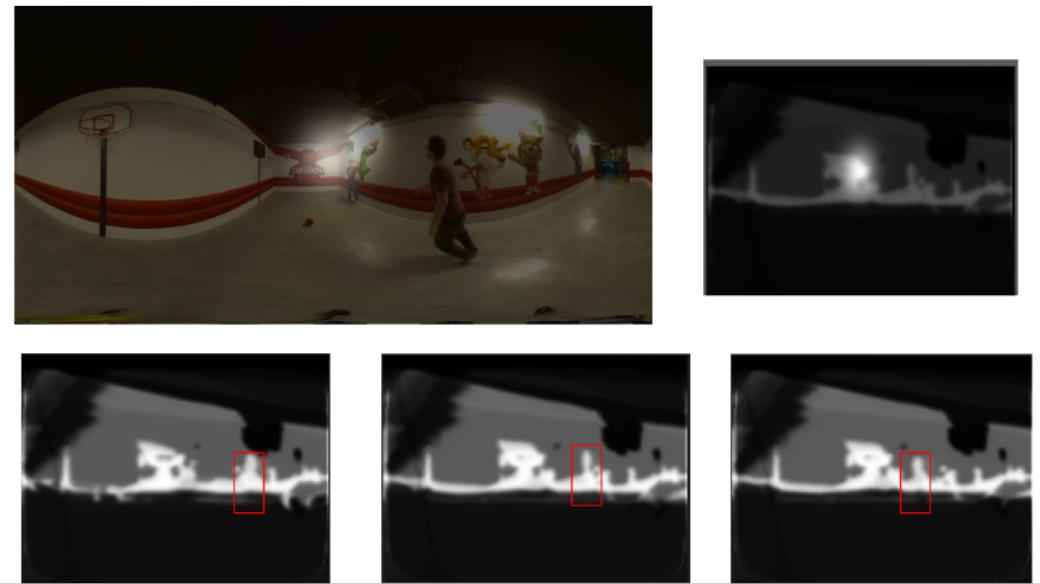


Figura 5.2: Ejemplo de tSPMs devuelto por el modelo generalista.

En la Figura 5.2 se puede ver cómo el modelo genera tSPMs que, aunque capturan la trayectoria de una persona avanzando en el video marcado con rectángulos rojos, también incluyen muchas otras partes importantes como si se mezclaran de otros videos. Por ejemplo, se observan áreas de alta probabilidad de fijación en partes del video que no deberían ser de interés en el contexto actual, lo que indica que el modelo está incorporando características aprendidas de otros videos en los tSPMs generados. Este fenómeno sugiere que el modelo necesita refinarse para mejorar su capacidad de generalización y evitar la contaminación cruzada de características entre diferentes videos, pero deja abierta una línea de investigación con gran potencial.

Capítulo 6

Conclusiones

En este Trabajo de Fin de Grado (TFG), se ha llevado a cabo un estudio exhaustivo sobre diferentes métodos de predicción de *scanpaths* en videos 360º. El objetivo principal fue analizar y comparar la efectividad de métodos heurísticos y modelos de aprendizaje profundo en la generación de *scanpaths* que se asemejen al comportamiento visual humano.

En primer lugar, se evaluaron varios métodos heurísticos de muestreo de saliencia, incluyendo muestreo aleatorio, valor máximo de la *saliencia*, sampleo probabilístico, inhibición de retorno, y combinaciones de sesgo ecuatorial y máxima distancia. Cada técnica fue analizada en términos de su capacidad para replicar patrones de exploración visual humana utilizando métricas clave como Dynamic Time Warping (DTW), determinismo (DET), recurrencia (REC) y distancia de Levenshtein (LEV).

Posteriormente, se diseñó e implementó un modelo de predicción de *scanpaths* basado en redes neuronales profundas. Este modelo, inspirado en arquitecturas como SST-Sal [9] y tSPM-Net [7], fue entrenado y evaluado utilizando el conjunto de datos D-SAV360 [12]. Se adoptó una estrategia de sobreajuste para evaluar la precisión del modelo en un entorno controlado y se realizaron pruebas adicionales para examinar su capacidad de generalización a nuevos datos.

El análisis comparativo los métodos de muestreo de saliencia estudiados, los resultados mostraron que las técnicas avanzadas, como el sampleo probabilístico y la inhibición de retorno, lograron una mayor similitud con los patrones de exploración visual humana. Estos métodos superaron significativamente al muestreo aleatorio y al valor máximo de la saliencia en términos de métricas clave como la recurrencia (REC) y la distancia de Levenshtein (LEV). La combinación de sesgo ecuatorial y máxima distancia también demostró ser eficaz en contextos específicos, mejorando la relevancia y la precisión de los *scanpaths* generados. En cuanto al método de muestreo utilizando los modelos con sobreajuste, si bien el modelo de aprendizaje automático propuesto presenta algunas ventajas, especialmente en la secuencia de fijaciones (LEV),

aún enfrenta desafíos en términos de forma general de *scanpaths* (DTW) y previsibilidad (DET).

La principal conclusión de este estudio es que, aunque los métodos avanzados de muestreo de saliencia y los modelos de aprendizaje automático muestran un potencial significativo para replicar los patrones de atención visual humana en entornos 360º, todavía existen áreas importantes de mejora. La capacidad del modelo de aprendizaje profundo para replicar la secuencia de fijaciones es prometedora, pero su rendimiento en la forma general y la previsibilidad de los *scanpaths* necesita refinamiento. Esto sugiere que la integración de técnicas adicionales y el uso de datos más diversos y ricos en características podrían fortalecer considerablemente el modelo propuesto.

6.1. Limitaciones y mejoras futuras

A pesar de los resultados positivos obtenidos, el modelo desarrollado presenta algunas limitaciones que se deben abordar en futuros trabajos. Una de las principales limitaciones es la capacidad de generalización del modelo. Durante las pruebas se observó que el modelo tenía tendencia a mezclar características de diferentes videos, sugiriendo que, aunque efectivo en un entorno controlado, enfrenta desafíos al generalizar a nuevos datos.

Para mejorar la generalización del modelo, se podrían explorar varias estrategias. Una opción es la integración de módulos adicionales que permitan la introducción de nuevos tipos de información, como el flujo óptico (optical flow) o estimaciones de profundidad para escenas en 360º, similar a lo propuesto en 360MonoDepth [14]. Estos módulos podrían ayudar a capturar mejor las características dinámicas y tridimensionales de las escenas, mejorando la precisión de los tSPMs generados.

Otra posible mejora es el uso de fijaciones en lugar de puntos en bruto. Las fijaciones proporcionan una representación más precisa de la atención visual, permitiendo que el modelo aprenda patrones de atención más representativos. Incorporar estas mejoras podría resultar en una mayor similitud con los patrones de exploración visual humana, mejorando tanto la precisión como la variabilidad de los *scanpaths* generados.

Finalmente, el uso de conjuntos de datos mayores y más variados sería crucial para entrenar el modelo de manera más robusta. Datos adicionales permitirían al modelo aprender una gama más amplia de patrones de atención visual, mejorando su capacidad para generalizar a nuevas escenas y condiciones. Además, la inclusión de datos multimodales, que integren tanto estímulos visuales como auditivos, podría ofrecer una visión más completa de cómo los usuarios interactúan con entornos 360º. En este sentido, se podrían aprovechar los datos del conjunto D-SAV360 [12], que ya incluye tanto información visual como auditiva, permitiendo al modelo capturar de manera más efectiva las interacciones multimodales que ocurren en entornos de realidad virtual.

Este trabajo es un primer paso prometedor hacia la predicción de *scanpaths* en video 360º, un campo subexplorado y con gran potencial.

Capítulo 7

Bibliografía

- [1] 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.
- [2] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics*, 2017.
- [3] Daniel Martin, Ana Serrano, and Belen Masia. Panoramic convolutions for 360° single-image saliency prediction. In *CVPR Workshop on Computer Vision for Augmented and Virtual Reality*, 2020.
- [4] Mai Xu, Yuhang Song, Jianyi Wang, Minglang Qiao, Liangyu Huo, and Zulin Wang. Predicting head movement in panoramic video: A deep reinforcement learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–1, 07 2018.
- [5] Daniel Martin, Ana Serrano, Alexander W Bergman, Gordon Wetzstein, and Belen Masia. Scangan360: A generative model of realistic scanpaths for 360° images. *IEEE Transactions on Visualization and Computer Graphics*, 28(5):2003–2013, 2022.
- [6] Marc Assens Reina, Xavier Giró-i Nieto, Kevin McGuinness, and Noel E. O’Connor. Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In *ICCV Workshop on Egocentric Perception, Interaction and Computing*, Oct 2017.
- [7] Daniel Martin, Diego Gutierrez, and Belen Masia. tspm-net: A probabilistic spatio-temporal approach for scanpath prediction. *To appear in Computers Graphics*, 1:1–11, 2042.

- [8] Ryan Anthony Jalova de Belen, Tomasz Bednarz, and Arcot Sowmya. Scanpathnet: A recurrent mixture density network for scanpath prediction. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 5006–5016, 2022.
- [9] Edurne Bernal-Berdun, Daniel Martin, Diego Gutierrez, and Belen Masia. Sst-sal: A spherical spatio-temporal approach for saliency prediction in 360° videos. Computers Graphics, 106:200–209, 2022.
- [10] Fang-Yi Chao, Cagri Ozcinar, and Aljosa Smolic. Transformer-based long-term viewport prediction in 360° video: Scanpath is all you need. In 2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP), pages 1–6, 2021.
- [11] Lilla M. Gurtner, Walter F. Bischof, and Fred W. Mast. Recurrence quantification analysis of eye movements during mental imagery. Journal of Vision, 19(1):17–17, 01 2019.
- [12] Edurne Bernal-Berdun, Daniel Martin, Sandra Malpica, Pedro J. Perez, Diego Gutierrez, Belen Masia, and Ana Serrano. D-sav360: A dataset of gaze scanpaths on 360° ambisonic videos. IEEE Transactions on Visualization and Computer Graphics, pages 1–11, 2023.
- [13] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution, 07 2018.
- [14] Manuel Rey-Area, Mingze Yuan, and Christian Richardt. 360monodepth: High-resolution 360° monocular depth estimation. CoRR, abs/2111.15669, 2021.

Listado de Figuras

1.1.	Diagrama de Gantt	5
2.1.	Arquitectura del modelo de ScanGAN360 [5].	8
2.2.	Arquitectura del modelo de SaltiNet [6].	8
2.3.	Arquitectura del modelo de tSPM-Net [7].	9
2.4.	Arquitectura del modelo de ScanpathNet [8].	9
2.5.	Arquitectura del modelo de SST-Sal [9].	10
2.6.	Arquitectura del modelo de VPT360 [10].	11
3.1.	Ejemplo de muestreo aleatorio.	14
3.2.	Ejemplo del punto máximo obtenido en rojo sobre el mapa de saliencia.	14
3.3.	Ejemplo de obtención de las zonas superiores a un percentil en un mapa de saliencia. El muestreo se realiza sobre las áreas blancas de forma aleatoria.	15
3.4.	Inhibición de retorno en el muestreo probabilístico.	16
3.5.	Aplicación del sesgo ecuatorial a un mapa de saliencia.	16
3.6.	Aplicación de la máxima distancia a un mapa de saliencia.	17
3.7.	Visualización de <i>scanpaths</i> para los distintos métodos de muestreo teniendo en cada fila un método y en columna un instante del video con id 1016 del dataset D-SAV360 [12]. El resto de métodos de muestreo está en la Figura 3.8. Cabe destacar que, después del punto rojo en una imagen, el <i>scanpath</i> continúa con el punto violeta en la imagen siguiente.	19
3.8.	Visualización de <i>scanpaths</i> para los distintos métodos de muestreo teniendo en cada fila un método y en columna un instante del video con id 1016 del dataset D-SAV360 [12]. El resto de métodos de muestreo está en la Figura 3.7. Cabe destacar que, después del punto rojo en una imagen, el <i>scanpath</i> continúa con el punto violeta en la imagen siguiente.	20
4.1.	Diseño de la arquitectura del modelo de predicción de scanpaths en videos 360º, inspirada en SST-Sal [9] y tSPM-Net [7].	24

5.1. Visualización de <i>scanpaths</i> para distintos vídeos usando el respectivo modelo sobreajustado, en el que para cada fila hay un vídeo y en una columna es el instante del video con el id del dataset D-SAV360 [12]. Cabe destacar que, después del punto rojo en una imagen, el <i>scanpath</i> continúa con el punto violeta en la imagen siguiente.	29
5.2. Ejemplo de tSPMs devuelto por el modelo generalista.	32
A.1. Visualización de un <i>scanpath</i> en un frame del vídeo.	46
A.2. Visualización de múltiples <i>scanpaths</i> en un frame de un vídeo.	46
A.3. Visualización de un <i>scanpath</i> de un vídeo mostrando el viewport que vería el usuario.	47

Lista de Tablas

1.1.	Horas totales dedicadas al desarrollo del trabajo.	4
3.1.	Resultados comparativos de diferentes métodos de muestreo, SE corresponde al sesgo ecuatorial y MD al muestreo de máxima distancia	18
5.1.	Resultados comparativos de diferentes métodos de sampleo con el modelo con sobreajuste.	28

Anexos

Anexos A

Métodos de Visualización de Scanpaths

En este anexo, se describen tres métodos de visualización de *scanpaths* utilizados en el análisis de datos de atención visual en videos 360º. Estos métodos de visualización adaptados de ScanGAN360 ofrecen una variedad de formas para analizar y comprender los *scanpaths* en videos 360º. Al utilizar escalas de colores, líneas de *viewport* y *thumbnails*, se facilita la interpretación de los datos de atención visual y se proporciona una herramienta valiosa para el análisis y optimización de contenido en entornos de realidad virtual.

A.1. Visualización de Puntos con Escala de Colores

Este método representa cada *scanpath* mostrando los últimos cinco puntos de fijación en una escala de colores. En esta escala, el rojo indica los puntos más antiguos y el azul los más recientes (Figura A.1). Además puede superponer los mapas de saliencia en la propia visualización. Para cada fotograma del video, se actualizan los puntos de fijación, permitiendo observar la trayectoria de la mirada del usuario a lo largo del tiempo.

Se seleccionan los cinco puntos de fijación más recientes en cada fotograma. Se asignan colores a estos puntos en una escala que va del rojo al azul, siendo el morado el punto más antiguo y el rojo el más reciente. Se superponen estos puntos sobre el fotograma correspondiente del video. Esta técnica permite identificar fácilmente la secuencia y la duración de las fijaciones, proporcionando una visualización clara de la dinámica de la atención visual del usuario.



Figura A.1: Visualización de un *scanpath* en un frame del vídeo.

A.2. Visualización de múltiples *scanpaths* con delimitaciones del *viewport*

Este método visualiza varios *scanpaths* simultáneamente, mostrando los límites del *viewport* con una línea que tiene en cuenta la distorsión esférica de la imagen (Figura A.2). Cada *scanpath* se representa con un color diferente, permitiendo diferenciar fácilmente entre las trayectorias de diferentes usuarios.



Figura A.2: Visualización de múltiples *scanpaths* en un frame de un vídeo.

Para cada fotograma, se dibujan los límites del *viewport* actual como una línea, considerando la distorsión esférica de la proyección equirectangular. Se interpolan los puntos de fijación para evitar saltos bruscos entre los puntos, actualizando el *viewport* cada cuatro fotogramas. Se asigna un color único a cada *scanpath* y se superpone sobre el fotograma correspondiente del video. Esta visualización facilita la comparación entre diferentes trayectorias de atención visual, proporcionando una perspectiva integral de cómo distintos usuarios exploran el entorno visual.

A.3. Visualización de Thumbnails con el *scanpath*

El tercer método genera una miniatura (*thumbnail*) que representa lo que el usuario vería en un dispositivo de realidad virtual (VR). Esta imagen procesada muestra solo la parte del *viewport* y distorsiona la imagen para que se pueda visualizar en un video normal (Figura A.3).



Figura A.3: Visualización de un *scanpath* de un vídeo mostrando el *viewport* que vería el usuario.

Para implementar esta visualización, se sigue el siguiente procedimiento. Se extrae la región del *viewport* correspondiente a cada punto de fijación. Se aplica una distorsión a esta región para ajustarla al formato de visualización de un video convencional. Se genera una miniatura para cada punto de fijación, mostrando únicamente la parte del *viewport* visible. Esta técnica proporciona una representación fiel de la experiencia visual del usuario, permitiendo visualizar directamente lo que el usuario está viendo en el entorno VR.